Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : **Martin Smal**

Titre : **Modélisation du risque sécheresse en MRH**

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

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

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : Exiom Partners

Signature :

Directrice du mémoire en entreprise :

Nom : Coralie Charbonnel,
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

Résumé

La fréquence et l'intensité croissante des catastrophes naturelles, en particulier de la sécheresse, en France, suscitent une préoccupation grandissante. Depuis 1982, de nombreuses évolutions réglementaires ont été mises en place pour la déclaration des catastrophes naturelles liées à la sécheresse. Ainsi, il devient essentiel pour les assureurs de modéliser efficacement ce risque. Cette modélisation permettrait de calculer des provisions appropriées, de mettre en place une stratégie de souscription affinée mais également d'optimiser leurs schémas de réassurance conformément aux exigences du SCR (Solvency Capital Requirement) établi par la réglementation Solvabilité 2.

Dans le cadre de ce mémoire, la création d'un modèle de prédiction est proposée pour les communes de la région Auvergne-Rhône-Alpes en utilisant les données fournies par Météo France ainsi que les données du portefeuille d'un assureur présent dans cette région. L'objectif principal est de créer un modèle qui prend en compte les réglementations mises en vigueur depuis 2019 pour le processus de déclaration de catastrophe naturelle, mais aussi de déterminer le coût du risque sous-jacent réellement observé par les assureurs.

La première partie de ce mémoire expose le contexte et présente les éléments essentiels à la compréhension approfondie du problème. Les parties suivantes abordent les notions théoriques et le développement du modèle proposé : la deuxième partie traite de la modélisation des indicateurs d'humidité des sols et de la déclaration du critère météorologique (cadre réglementaire) tandis que la troisième partie traite de la modélisation du coût du risque observé chez les assureurs. Enfin, la quatrième partie propose une application du modèle en assurance, ainsi que des limites du régime Cat Nat actuel.

En combinant les approches basées sur le risque de déclaration de catastrophe naturelle et sur le risque subsidence réellement observé chez les assureurs, ce mémoire offre une analyse approfondie de la modélisation du risque sécheresse pour les contrats Multirisques Habitation MRH.

Les résultats et les recommandations formulées dans ce travail de recherche peuvent servir de base pour aider les assureurs à mieux appréhender ce risque croissant et à prendre des décisions éclairées en matière de gestion du portefeuille, de politique de souscription et de pilotage stratégique.

Mots-clés : *Sécheresse, Multirisque habitation, Catastrophe naturelle, Régime Cat Nat, Argile, Retrait-gonflement, Risque subsidence, Critère météorologique, Critère géologique, Soil Wetness Index, Loi Rousseau, Réchauffement climatique, Auvergne-Rhône-Alpes, France.*

Abstract

The increasing frequency and intensity of natural disasters, particularly droughts, in France are raising increasing concerns. Since 1982, numerous regulatory developments have been implemented for the declaration of natural disasters related to drought. Thus, it is essential for insurers to effectively model this risk. This modeling would enable them to calculate appropriate provisions, refine underwriting strategies, optimize reinsurance schemes, and comply with the requirements of the Solvency Capital Requirement (SCR) established by Solvency 2 regulations.

In the context of this thesis, the creation of a prediction model is proposed for the municipalities of the Auvergne-Rhône-Alpes region, using data provided by Météo France as well as data from an insurer's portfolio present in this region. The main objective is to create a model that takes into account the recent regulations implemented in 2019, 2021, and 2023 for the natural disaster declaration process, as well as the actual underlying risk cost observed by insurers.

The first part of this thesis outlines the context and presents essential elements for a comprehensive understanding of the problem. The following sections address theoretical concepts and the development of the proposed model, the second part deals with the modeling of soil moisture indicators and the declaration of meteorological criteria (regulatory framework), while the third part addresses the modeling of the actual underlying risk cost. Finally, the fourth part proposes an application of the model in insurance, as well as the limitations of the current Cat Nat regime.

By combining the approaches of natural disaster declaration risk and the risk of subsidence risk observed by insurers, this thesis offers an in-depth analysis of drought risk modeling for Multirisks Habitation (MRH) contracts in the Auvergne-Rhône-Alpes region. The results and recommendations presented in this research work can serve as a basis to assist insurers in better understanding this growing risk. Additionally, they can aid in making informed decisions regarding portfolio management, underwriting policy, and strategic planning.

Keywords : *Drought, Multirisk Home Insurance, Natural disaster, Natural disaster insurance scheme, Clay, Shrink-swell, Subsidence risk, Meteorological criterion, Geological criterion, Soil Wetness Index, Rousseau Law, Climate change, Auvergne-Rhône-Alpes, France.*

Remerciements

Je tiens à exprimer mes sincères remerciements à toutes les personnes ayant contribué de près ou de loin à la réalisation de ce mémoire.

Tout d'abord, je souhaite remercier chaleureusement ma tutrice d'entreprise, Coralie Charbonnel, pour son précieux accompagnement tout au long de cette année. Sa disponibilité, ses conseils éclairés et son expertise dans le domaine m'ont été d'une aide précieuse pour avancer dans mon projet et réussir à élaborer les modèles nécessaires à cette étude.

Je tiens également à exprimer ma gratitude à Adrien Cortes et à Arthur Bourdon, mes amis, pour leurs conseils dans l'élaboration du modèle théorique. Leurs remarques pertinentes ont grandement contribué à améliorer la qualité de ce travail.

Un grand merci à notre client, qui m'a permis d'avoir une base de données sur laquelle travailler et dont les suggestions et les éclairages sur des aspects auxquels je n'avais pas pensé ont enrichi ma réflexion et apporté une dimension concrète à mon étude.

Enfin, je tiens à exprimer ma reconnaissance envers mon tuteur académique, Aurélien Coulomy, pour ses précieux conseils et son suivi attentif tout au long de ce projet. Sa disponibilité et son expertise m'ont permis de progresser et de réaliser ce travail dans les meilleures conditions.

En somme, je suis profondément reconnaissant envers toutes ces personnes qui ont contribué à la réalisation de ce mémoire. Leur soutien et leurs encouragements ont été essentiels pour mener à bien ce travail de recherche.

Sommaire

Introduction

Depuis la fin des années 1990, les catastrophes naturelles sont de plus en plus fréquentes et intenses en France. La sinistralité augmente donc fortement ce qui préoccupe de plus en plus les assureurs et les réassureurs.
Historiquement, ce sont les inondations qui ont coûté le plus cher aux assureurs (53 % de la sinistralité Cat Nat non-auto cumulée sur la période 1982-2021 selon le Bilan Cat Nat de la Caisse Centrale de Réassurance (CCR) [2]). A la deuxième place de ce classement se situent les sinistres indemnisés au titre de la sécheresse (avec 37 %).
Cependant, plusieurs évolutions notables de la sinistralité sécheresse sont constatées depuis 1982, notamment à cause de plusieurs changements dans les règles de reconnaissance de catastrophes naturelles concernant le péril sécheresse.

Par ailleurs, il est constaté que depuis 2016 environ (2018 en Auvergne-Rhône-Alpes), la sinistralité liée à la sécheresse augmente fortement. Selon la CCR, la part de la sinistralité sécheresse a évolué de 37 % sur la période 1982-2021 à 52 % de la sinistralité Cat Nat non auto cumulée des 10 dernières années [14]. L'année 2022 est d'ailleurs un bon exemple puisque la Fédération Française des Assurances a estimé, le 21 novembre 2022, le coût de la sécheresse entre 1,9 et 2,8 milliards d'euros. La sécheresse de 2003, record en terme de sinistralité jusqu'alors, ayant coûté 1,94 milliard d'euros.

Bien que les assureurs puissent bénéficier de la réassurance de la CCR et de la garantie illimitée de l'état, il demeure essentiel pour eux de mieux comprendre, modéliser et provisionner ce risque afin de respecter le SCR (*Solvency Capital Ratio*) imposé par Solvabilité 2, de limiter la dégradation de leur résultat et d'éviter une revalorisation à la hausse de leurs contrats de réassurance.

Enfin, il est important de souligner que tous les sinistres sécheresse liés au retrait-gonflement des sols argileux sont indemnisés au titre de la garantie Cat Nat dans des contrats MRH (les autres sinistres sécheresse -sécheresse agricole par exemple - sont indemnisés au titre d'autres garanties).
Cette garantie est obligatoire et fixée par l'État à 12 % de la prime dommages aux biens. Son déclenchement est conditionné par la déclaration d'une catastrophe naturelle au niveau de la commune, sans quoi l'assuré ne peut prétendre à aucune indemnisation. Cette déclaration est valable à condition de respecter les critères météorologique et géologique. Ce fonctionnement rend donc la modélisation de la sinistralité sécheresse atypique et il doit être pris en compte dans la modélisation.

L'objectif de ce mémoire est alors de construire un modèle prédictif de la sinistralité sécheresse en multirisque habitation (MRH). L'étude concernera à la fois la modélisation d'indicateurs d'humidité et la probabilité de déclaration de l'état de catastrophe naturelle au niveau des communes et à la fois la modélisation du risque subsidence au niveau des contrats. Elle se concentrera sur la région

Auvergne-Rhône-Alpes, sur laquelle les données ont été récoltées.

Les travaux réalisés sur le sujet traitent généralement de la modélisation de la sinistralité réelle de la sécheresse en créant par exemple de nouveaux indicateurs, ou en construisant des modèles de précipitation et de température comme il est possible de le constater dans les travaux publiés à l'Institut des actuaires sur le sujet en 2016, 2018 et 2021 ([27], [5], [21] et [8]). Cependant, il a été constaté que la modélisation, et donc la prise en compte de la déclaration de l'état de catastrophe naturelle au niveau de la commune, était minimisée. Il sera mis en évidence ci-après que le système d'indemnisation Cat Nat sécheresse français est particulier et conditionne très largement l'indemnisation réelle d'un assuré.
Afin d'avoir une vision la plus macroscopique possible, le mémoire s'organise autour de deux grandes parties.

1. Tout d'abord, la modélisation des indicateurs d'humidité des sols est traitée. Cette modélisation permet d'estimer la probabilité de survenance d'un des critères relatifs au déclenchement de la garantie Cat Nat sécheresse au niveau de la commune. De plus, cette modélisation peut être intégrée dans la suite de l'étude.
2. Ensuite, il est question de modéliser la sinistralité réellement observée chez l'assureur. C'est-à-dire qu'il est ici sujet de modéliser le risque sécheresse au niveau du contrat en fonction de ses caractéristiques endogènes (variables habituelles utilisées dans la tarification MRH) mais aussi de variables exogènes (exposition au risque de retrait-gonflement des argiles et indicateurs d'humidité estimés à partir de la deuxième partie).

La première et la deuxième partie font référence aux critères de déclenchement de catastrophe naturelle sécheresse revus en 2019 [24].
La modélisation du critère météorologique repose sur l'historique de données *Soil Wetness Index* (SWI) (indicateur défini par Météo-France [7]). Cet historique est fourni par Météo-France depuis 1969 [13]. Ainsi, en prenant en compte l'historique et en construisant un modèle prédicatif du SWI futur, il est possible d'extraire la probabilité de déclencher le critère météorologique au sein de chaque commune sur un horizon souhaité. Cette partie présente donc la construction d'un modèle prédictif du critère météorologique.
Les critères étant probablement amenés à être modifiés prochainement [24], l'étude présente également l'impact du passage potentiel de ces modifications [25].

La modélisation du risque observé chez les assureurs est réalisée à partir d'une base de données issue d'un portefeuille basé en Auvergne-Rhône-Alpes. C'est pourquoi l'étude est restreinte à cette zone géographique. Les données étant particulières, le modèle utilisé pour la fréquence est un XGBoost. L'un des enjeux de cette partie est principalement l'ajout d'une variable exogène aux variables caractéristiques du contrat. Cette variable exogène, appelée Score Argile, correspond à l'exposition à l'argile de l'habitation et est construite à partir des mêmes données d'exposition à l'argile présentées ci-dessuus fournies par Géorisques [10]. La prise en compte de cette variable permet une prédiction plus précise de la sinistralité des contrats.

Enfin, en dernière partie est proposée une application du modèle pour un assureur ainsi que les limites observées.

Première partie

Contexte et enjeu de la modélisation du risque sécheresse en assurance

Chapitre 1

Les catastrophes naturelles en France

Une catastrophe naturelle est un évènement imprévisible lié à un phénomène non généré par l'homme. Elle est souvent liée à un phénomène météorologique et génère beaucoup de dégâts et de victimes. Le changement climatique cause une augmentation de la fréquence et de l'intensité de ces catastrophes naturelles depuis quelques années.

1.1 Définition et tendance

Selon l'INSEE, une catastrophe naturelle est caractérisée par l'intensité anormale d'un agent naturel (inondation, coulée de boue, tremblement de terre, avalanche, sécheresse...) lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises.
Autrement dit, les catastrophes naturelles sont des évènements d'occurrence faible mais qui génèrent des coûts très élevés.
La tendance actuelle est une nette augmentation du nombre de catastrophes naturelles.

Voici un graphique issu des Ministères Écologie Énergie Territoires qui décrit le nombre d'événements ayant fait plus de 10 morts ou coûté plus de 30 M€ courants de dommages matériels depuis 1950.

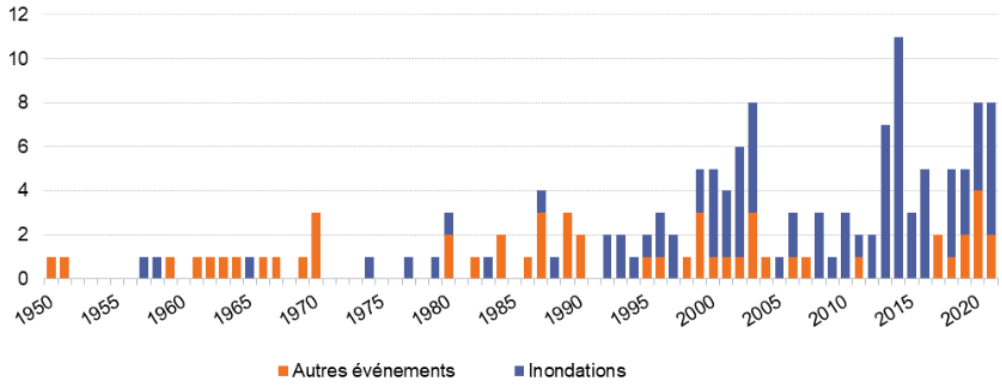


FIGURE I.1 – Occurrence des catastrophes naturelles graves (MTECT, 2022).

Ce graphique montre que les catastrophes naturelles graves sont de plus en plus fréquentes.

De plus, la France est un territoire particulièrement exposé aux catastrophes naturelles comme le montre ce graphique issu de "*The international disasters database*" :

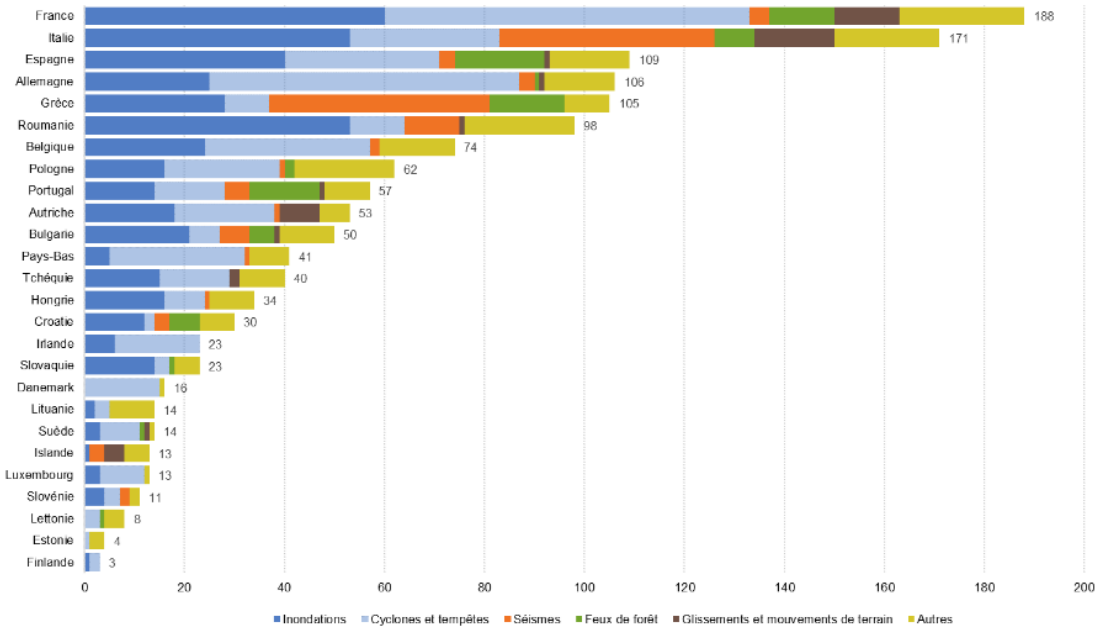


FIGURE I.2 – Nombre de catastrophes naturelles survenues en Europe entre 1900 et 2022 (The International Disaster Database, 2022).

Le pays est confronté simultanément aux risques d'inondations, de tempêtes et de sécheresses. À titre d'exemple, la France a subi les inondations de 1910, qui ont touché la Seine et ses environs, ainsi que celles de 1955 dans le sud de la France, principalement dans la région de Nice.
Les tempêtes de 1999, Martin et Lothar, ont également marqué l'histoire des événements météorologiques extrêmes en France. De plus, les canicules européennes de 2003, caractérisées par des températures exceptionnellement élevées et des décès liés à la chaleur, ainsi que celles de 2020, 2022 et 2023, moins meurtrières mais d'une intensité accrue, soulignent la variété des défis auxquels la France est confrontée en matière de catastrophes naturelles.
Historiquement, c'est le risque inondation qui coûte le plus cher aux assureurs (53 % de la sinistralité Cat Nat non-auto entre 1982 et 2021). La sécheresse représente 37 % de cette sinistralité sur cette période. Cependant, une nette augmentation de la sinistralité est constatée depuis 2018, avec des sécheresses plus ou moins intenses et longues tous les ans, exception faite en 2021. Ainsi, le pourcentage de sinistralité associé à la sécheresse est passé de 37 % à 52 % sur la période 2012-2022.

1.2 Le risque sécheresse en particulier

1.2.1 Différents types de sécheresse

Il convient avant tout de définir de quelle sécheresse il est question dans ce mémoire. Il existe en effet plusieurs types de sécheresse, qui ne sont pas indemnisées de la même manière : la sécheresse météorologique, la sécheresse agricole et la sécheresse hydrologique.

La sécheresse météorologique se manifeste par une période prolongée où les précipitations sont inférieures à la normale. Elle est caractérisée par un déficit pluviométrique persistant.

La sécheresse agricole survient lorsque les sols présentent un manque d'eau à une profondeur maximale d'un mètre, ce qui a un impact sur la croissance des végétaux. Même si les précipitations sont normales, cette forme de sécheresse peut se produire en raison de l'évaporation et de la transpiration des plantes. Ainsi, elle est influencée non seulement par les précipitations, mais également par des facteurs tels que la température de l'air, le vent, la composition des sols, les pratiques agricoles et la végétation.

La sécheresse hydrologique se manifeste lorsque le niveau des eaux souterraines, des rivières, des ruisseaux ou des lacs descend en dessous de la moyenne. Ce type de sécheresse dépend principalement des précipitations, mais il est également influencé par le type de sol, qui affecte l'infiltration et l'écoulement de l'eau. De plus, elle peut se produire même avec des précipitations normales ou supérieures à la moyenne en raison de la surexploitation des ressources en eau.

La sécheresse ici traitée est la sécheresse météorologique. C'est elle qui entraîne un retrait-gonflement des argiles et qui cause des dommages au bâti. Ces sinistres sont indemnisés au titre de la garantie catastrophe naturelle obligatoire dans les contrats Multirisque Habitation (MRH).

1.2.2 Le retrait-gonflement des argiles (RGA)

Les sols argileux possèdent la propriété de voir leur consistance se modifier en fonction de leur teneur en eau. Ainsi, en contexte humide, un sol argileux se présente comme souple et malléable, tandis que ce même sol desséché sera dur et cassant.
Des variations de volume plus ou moins conséquentes en fonction de la structure du sol et des minéraux en présence, accompagnent ces modifications de consistance.

Ainsi, lorsque la teneur en eau augmente dans un sol argileux, le volume de ce sol augmente. ce phénomène est appelé "gonflement des argiles".
Au contraire, une baisse de la teneur en eau provoquera un phénomène de rétractation ou "retrait des argiles".

Ces variations de volume fragilisent les sols et peuvent donc causer des fissures sur les bâtis. Et c'est précisément ce sinistre dont il est question dans ce mémoire.
Voici ci-dessous une représentation du phénomène de retrait-gonflement des argiles issu du site public BRGM (Bureau de Recherches Géologiques et Minières) :

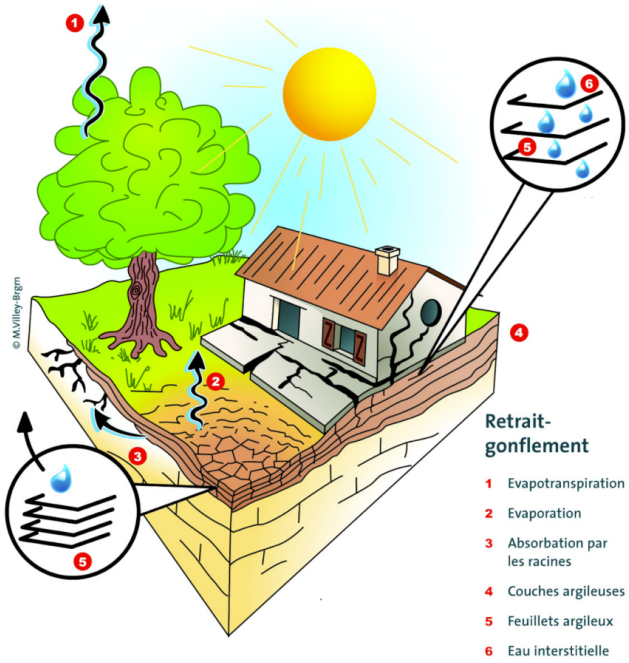


FIGURE I.3 – Mécanisme de fonctionnement du phénomène de retrait-gonflement des sols argileux (BRGM, 2016) [16].

Les fissures sur les maisons ne peuvent être indemnisées qu'au titre de la garantie catastrophe naturelle. Sinon, l'assuré peut essayer de se retourner contre le constructeur et faire jouer sa garantie décennale. Autrement dit, tout sinistre apparaissant hors période de catastrophe naturelle ou jugé non consécutif à la catastrophe naturelle ne sera pas indemnisé par l'assureur. Ces sinistres coûtent parfois cher (selon les données retenues dans cette étude, en moyenne 20 870 € selon les données récoltées dans le cadre de ce mémoire).
De plus, il est parfois difficile de modéliser ces sinistres car ils peuvent apparaître longtemps après l'épisode de sécheresse (plusieurs mois). Ainsi, passé un certain délai, il faudra justifier que les dégâts sont consécutifs à une sécheresse antérieure.
Quelques changements réglementaires, présentés ci-dessous en 2.3.4, devraient faciliter les démarches et l'indemnisation des assurés.

1.2.3 Exposition de la France au RGA

Le centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement (Cerema) estime que près de la moitié des sols en France sont concernés par des phénomènes de RGA et que 10 millions de maisons individuelles sont exposées.
C'est dans ce contexte qu'en 2019, le BRGM a accompagné la Direction Générale de la Prévention des Risques du Ministère chargé de l'Environnement dans la mise en place de l'article 68 de la loi Élan portant sur la prévention des risques liés au phénomène de retrait-gonflement des formations argileuses en contribuant à la définition du zonage réglementaire et à la rédaction des décrets afférents. Une représentation de ce zonage est présentée en I.4. Les données

ont également pu être récupérées à une échelle plus fine (Auvergne-Rhône-Alpes), sur le site Géorisques. Elles sont montrées en I.5.

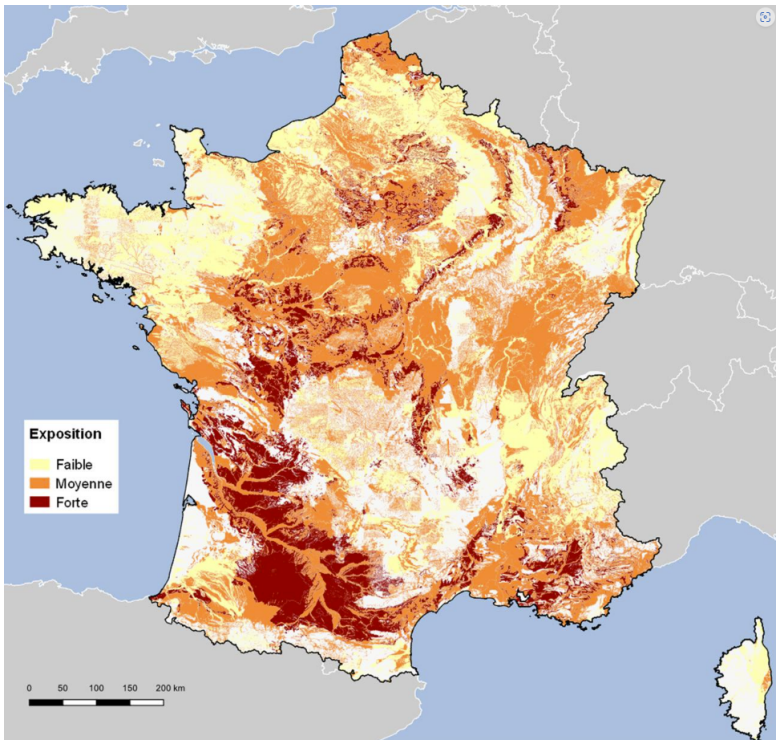


FIGURE I.4 – Carte d'exposition au phénomène de retrait-gonflement des sols argileux (BRGM, 2020).

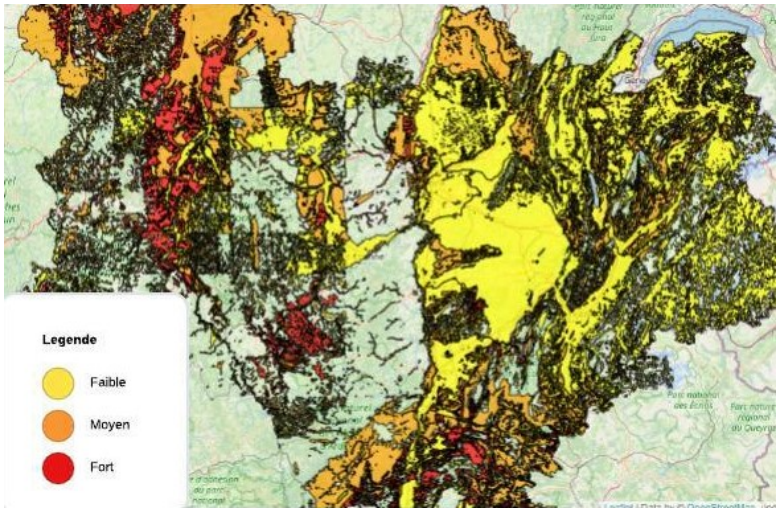


FIGURE I.5 – Carte d'exposition au phénomène RGA en Auvergne-Rhône-Alpes (BRGM, 2021).

Chapitre 2

Régime Cat Nat

Les assureurs peuvent prendre en charge les périls consécutifs aux catastrophes naturelles au titre de la garantie catastrophe naturelle. L'étude propose alors d'étudier le régime d'indemnisation des catastrophes naturelles, et ses spécificités dans le cas de la sécheresse.

2.1 Fonctionnement et catastrophes naturelles concernées

L'Etat impose en France une garantie obligatoire appelée garantie catastrophe naturelle dans les contrats d'assurance de dommages. Les assurés peuvent bénéficier de cette garantie seulement si la commune a fait l'objet d'un arrêté publié au Journal Officiel reconnaissant l'état de catastrophe naturelle.
Les évènements suivants peuvent être considérés comme des catastrophes naturelles : *inondations, coulées de boue, tremblements de terre, éruptions volcaniques, mouvements de terrain, sécheresses, raz-de-marée, avalanches et affaissements de terrain*. Il est nécessaire de noter que la tempête, la grêle et la neige font l'objet d'une garantie particulière et n'entrent pas dans le cadre de la garantie catastrophe naturelle.
Le fonctionnement détaillé de la déclaration catastrophe naturelle est décrit par la CCR dans le schéma ci-dessous :

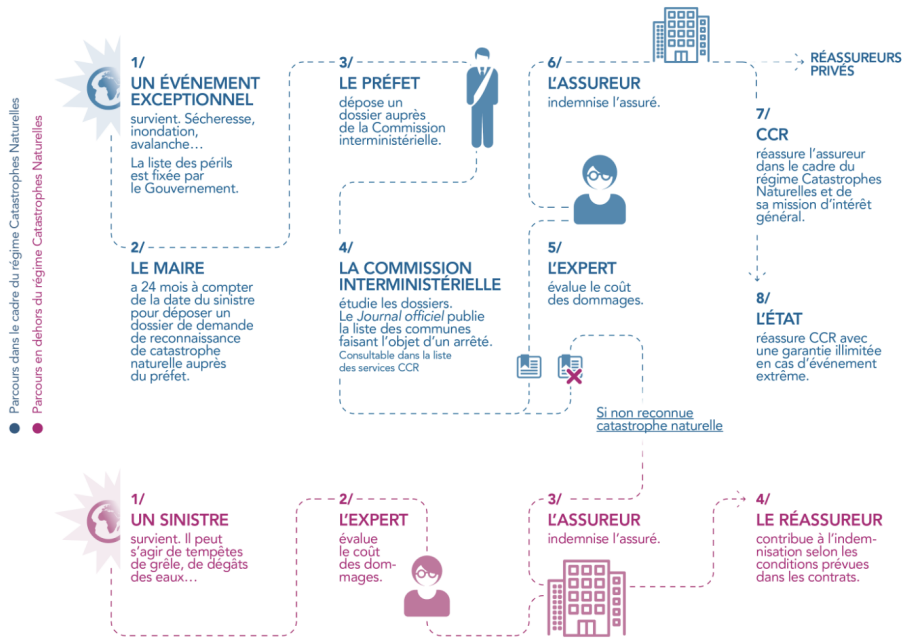


FIGURE I.1 – Fonction général du régime Cat Nat en France (CCR, 2022).

2.2 Garantie Cat Nat dans les contrats MRH

Pour les contrats Multirisque Habitation (MRH) (le fonctionnement de la garantie catastrophe naturelle pour les contrats auto ne sera pas détaillé), l'Etat fixe une surprime correspondant à 12 % de la prime afférente aux garanties dommages du contrat de base afin de financer la garantie Cat Nat. Autrement dit, le régime catastrophe naturelle est basé sur la mutualité des assurés. Aucune discrimination n'est faite entre les contrats et les régions et chacun contribue à part égale au régime d'indemnisation des catastrophes naturelles.
Il convient cependant de noter plusieurs évolutions du régime depuis 1982. En effet, le taux de la prime garantie catastrophe naturelle (par rapport à la prime dommages aux biens) est passé de 5,5 % à 9 % en 1983 et de 9 % à 12 % en 2000. Ce taux pourrait être amené à évoluer dans un horizon proche pour faire face à l'augmentation des catastrophes naturelles constatée depuis 2009.
De plus, les primes catastrophe naturelle sont soumises à un prélèvement qui alimente le Fonds de Prévention des Risques Naturels Majeurs (FPRNM, dit Fonds Barnier). Ce prélèvement a régulièrement augmenté pour atteindre 12 % à partir de 2009.
Le FPRNM a été créé par la loi n°95-101 du 2 février 1995 relative au renforcement de la protection de l'environnement. Il était initialement destiné à financer les indemnités d'expropriation de biens exposés à un risque naturel majeur. Son utilisation a été élargie aujourd'hui à d'autres dépenses et il est intégré depuis 2011 au budget de l'État.
L'évolution des primes Cat Nat entre 1982 et 2021 est clairement visible sur le bilan Cat Nat 1982-2021 de la CCR :

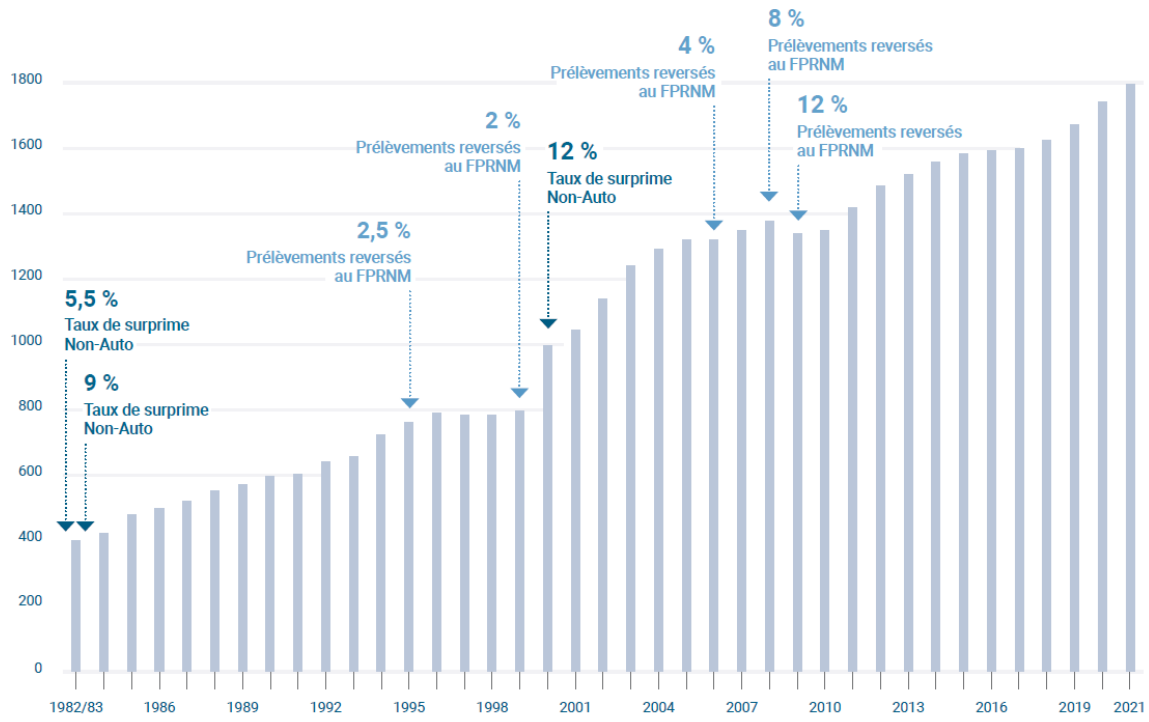


FIGURE I.2 – Evolution prime Cat Nat - Bilan Cat nat CCR 1982-2021 (CCR, 2023).

2.3 Règles d'indemnisation pour la sécheresse

Pour déclarer une catastrophe naturelle sécheresse dans une commune, plusieurs critères doivent être respectés. Les critères actuellement en vigueur, récemment modifiés par la circulaire du 10 mai 2019 "Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels" [24] sont les suivants :

— critère météorologique : basé sur l'indicateur d'humidité des sols SWI ;
— critère géologique : basé sur l'exposition de la commune au phénomène RGA selon la carte du BRGM.

2.3.1 Critère météorologique

Le critère météorologique vise à caractériser la teneur en eau du sol superficiel (les premiers mètres). Il s'appuie sur le recueil et le traitement par Météo-France de nombreuses données météorologiques et hydrologiques permettant de caractériser la teneur en eau des sols.
A partir de ces données, un indicateur d'humidité des sols superficiels est calculé pour 8981 points répartis sur le territoire et analysé pour chacune des quatre saisons de l'année.
Les valeurs de cet indicateur sont comparées à celles obtenues sur les cinquante dernières années pour évaluer le caractère exceptionnel de l'intensité de la sécheresse.

Le SWI Uniforme

Le SWI Uniforme représente, sur une profondeur d'environ deux mètres, l'état de la réserve en eau du sol par rapport à la réserve utile (eau disponible pour l'alimentation des plantes). Il s'agit donc bien de l'état hydrique du sol superficiel pour des sols argileux (noté entre 0 et 1) et non du remplissage des nappes phréatiques. Si le SWI uniforme est égal à zéro, le sol est très sec et les végétaux ne peuvent plus absorber d'eau, tandis que si le SWI uniforme est égal à un, le sol est saturé d'eau et a atteint sa réserve utile. Une définition précise du SWI est proposée par Météo-France [7].

Le SWI Uniforme est fourni par Météo-France pour 8981 mailles de 8 km^2 en France [13]. Il serait techniquement possible de le mesurer mais le nombre de mailles conséquent a contraint Météo-France à utiliser le modèle SIM (Safran-Isba-Modcou) [28].
Il utilise des données météorologiques (température de l'air, niveaux de précipitations, niveaux de rayonnement, vents, etc.) recueillies par le réseau d'observation de Météo-France qui comprend plusieurs milliers de stations de mesure régulièrement réparties sur le territoire.

Chacune des mailles ainsi définie est numérotée et recouvre tout ou partie d'une commune. Ce maillage est fixe et n'évolue pas d'une année sur l'autre.

Voici un exemple pour la commune de Loches (Indre-et-Loire) qui est recouverte par les mailles 4390, 4391, 4506 et 4507 :

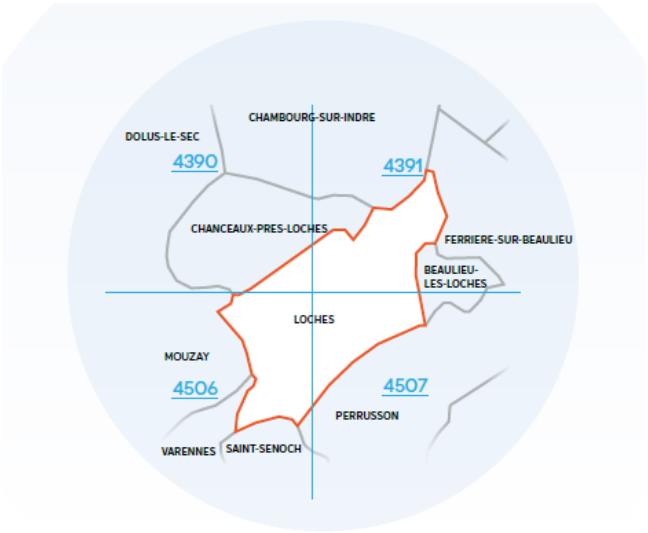


FIGURE I.3 – Zoom maillage Loches [7].

Indicateurs mensuels

Des indicateurs mensuels sont calculés et analysés pour chaque saison à partir du SWI uniforme.
Sur chacune des 8981 mailles, l'indice SWI uniforme d'humidité des sols est calculé quotidiennement. Pour chaque mois et chaque maille, Météo-France établit un indicateur d'humidité des sols superficiels dit « mensuel » en calculant la moyenne des indices SWI uniformes journaliers au cours de ce mois et des deux mois précédents. Ainsi, l'indicateur d'humidité des sols superficiels du mois de juillet est établi en s'appuyant sur la moyenne des indices quotidiens d'humidité des sols des mois de mai, juin et juillet. L'utilisation d'une période « glissante » de trois mois permet de tenir compte de la cinétique lente des phénomènes de sécheresse météorologique qui se manifestent sur plusieurs mois.
La commission interministérielle CATNAT considère une année civile comme découpée en quatre saisons qui sont en fait des trimestres complets :

- l'hiver correspond aux mois de janvier, février et mars ;
- le printemps correspond aux mois d'avril, mai et juin ;
- l'été correspond aux mois de juillet, août et septembre ;
- l'automne correspond aux mois d'octobre, novembre et décembre.

Pour chacune de ces saisons, trois indicateurs d'humidité des sols superficiels mensuels sont donc définis.

Par exemple, pour la saison de l'hiver, les indicateurs sont calculés de la manière suivante :

- indicateur de janvier : données de novembre de l'année précédente à janvier de l'année considérée ;
- indicateur de février : données de décembre de l'année précédente à février de l'année considérée ;

• indicateur de mars : données de janvier à mars de l'année considérée.

Règle de caractère anormal : la durée de retour

Le caractère anormal de l'intensité de la sécheresse est évalué à partir de la durée de retour. La durée de retour est le temps statistique entre deux occurrences d'un événement naturel d'une intensité donnée. L'autorité administrative considère qu'un épisode de sécheresse est anormal dès lors que son intensité, évaluée par l'indicateur d'humidité des sols superficiel correspond à une durée de retour supérieure ou égale à 25 ans.
Pour déterminer la durée de retour d'un épisode de sécheresse, Météo-France compare l'indicateur d'humidité des sols superficiels établi pour un mois donné avec les indicateurs établis pour ce même mois au cours des cinquante dernières années. Cette période « glissante » sur cinquante ans, qui intègre les années les plus récentes, permet de tenir compte de l'évolution du climat.
Par exemple, sur le graphique suivant pour la maille 1497 au mois de janvier, il y a caractère anormal en 2018 et 2019, mais pas en 2020. En effet, sur la période 1968-2018, la valeur du SWI du mois de septembre 2018 est la première plus basse valeur. De même, sur la période 1969-2019, la valeur du SWI du mois de septembre 2019 est la première plus basse valeur.

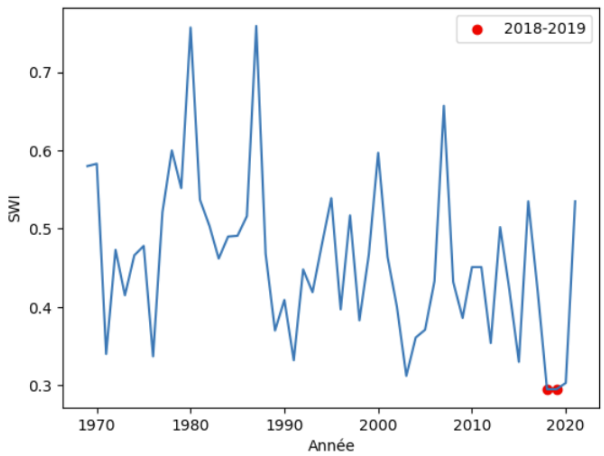


FIGURE I.4 – Historique SWI 1969-2021 mois de septembre, maille 1497 (chevauchant la commune Belles-forêts, département 57, Grand-Est).

Quelles communes remplissent le critère météorologique ?

Le critère météorologique est retenu pour chaque saison (hiver, printemps, été, automne). L'autorité administrative retient l'indicateur mensuel d'humidité des sols superficiels présentant la durée de retour la plus élevée pour chaque saison.
Exemple : pour l'hiver, si l'indicateur de mars présente une durée de retour supérieure à 25 ans mais pas les indicateurs de janvier et février, c'est l'indicateur de mars qui est retenu pour qualifier l'intensité de la sécheresse de l'hiver. Les communes respectent le critère météorologique même si une partie seulement (une maille) de leur territoire est touchée par un épisode anormal.
Autrement dit, une commune peut très bien bénéficier d'un arrêté Cat Nat si une seule maille respecte le critère anormal de sécheresse sur un seul mois de la saison.

2.3.2 Critère géologique

Le critère géologique est rempli si au moins 3 % de la commune est exposée au risque RGA défini par la cartographie BRGM du paragraphe précédent. Dans le cas où moins de 3 % de la commune est exposée, et où le critère météorologique est respecté, une étude géotechnique peut être réalisée afin de faire une demande exceptionnelle de reconnaissance de catastrophe naturelle.

En se basant sur la carte d'exposition réelle fournie par Géorisques I.5 et sur la base de données présentant les découpages des communes en France, disponible en ligne [6], il est possible d'observer, à l'aide de la bibliothèque GeoPandas de python, si chaque commune est touchée à au moins 3% par un niveau d'exposition RGA faible, moyen ou fort.

Le critère géologique est ainsi reconstruit sur python à partir de ces données. Voici les résultats obtenus, représentés sur la carte d'Auvergne-Rhône-Alpes I.5.

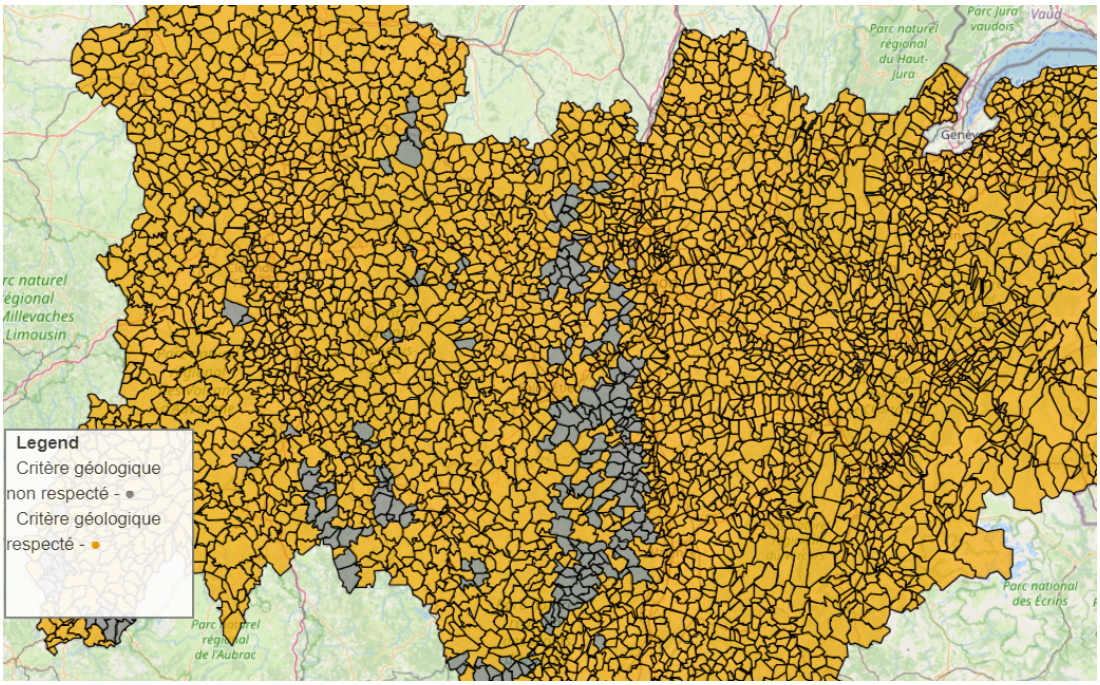


FIGURE I.5 – Critère géologique Auvergne-Rhône-Alpes.

Pratiquement toutes les communes d'Auvergne-Rhône-Alpes respectent le critère géologique. Par ailleurs, il a été négligé que certaines communes peuvent demander des études géologiques approfondies si leur exposition au phénomène RGA est inférieure à 3 %, dans le but de se conformer au critère géologique.

2.3.3 Plans de Prévention du Risque Sécheresse et moyens de prévention

Dans le but de prévenir les risques naturels et de réduire les dommages causés par les catastrophes, le gouvernement a instauré en 1995 les Politiques Publiques de Prévention des Risques Naturels (PPRN). Les préfets et les maires des zones à risque peuvent demander la mise en place d'un PPRN auprès de l'État. Ces plans visent à diminuer la probabilité de destruction et la vulnérabilité des bâtiments lors d'événements naturels en proposant des mesures de prévention et de protection. Cependant, ces mesures ne s'appliquent qu'aux maisons individuelles.

Les Plans de Prévention du Risque Sécheresse (PPRS), quant à eux, ont été mis en place dix ans après les PPRN, avec les premières approbations en 2004. Les PPRS définissent des normes de construction obligatoires pour les nouvelles maisons, telles que des fondations adaptées au sol, des chaînages internes dans les murs pour prévenir les fissures, et une distance suffisante par rapport aux sources d'humidité pour réduire les risques de retrait-gonflement des argiles. Les PPRS comprennent également des plans d'urbanisation et d'aménagement du territoire, ainsi que des zones interdites à la construction, entre autres mesures.
Un exemple de PPRN est fourni en référence [23].
De plus, l'article 68 de la loi Élan rend désormais obligatoire la réalisation d'études géotechniques dans les zones exposées au phénomène afin de prévenir et limiter les risques liés à d'éventuels sinistres. Les études géotechniques obligatoires sont des études effectuées par des experts pour évaluer la stabilité du sol et des terrains environnants afin de déterminer les risques potentiels de mouvements de terrain.

Enfin, parmi les moyens de prévention, il existe différents dispositifs :

— Les barrières anti-racines sont des dispositifs installés pour empêcher les racines des arbres et des plantes de s'infiltrer dans le sol et de causer des dommages aux structures et aux fondations des bâtiments.

— Les trottoirs et les drains périphériques sont des aménagements qui permettent de limiter les risques d'infiltration d'eau dans les sols environnants et de prévenir les érosions. Ils contribuent également à la protection des bâtiments contre les risques de mouvements de terrain en facilitant l'écoulement de l'eau et en limitant l'accumulation de pression dans le sol.

— Enfin, limiter la végétation autour de la maison est une mesure de précaution courante pour réduire les risques de mouvements de terrain. La végétation peut affecter la stabilité du sol en absorbant l'eau et en exerçant une pression sur les fondations. En limitant la croissance des racines et en éliminant les plantes trop proches de la maison, il est possible de réduire les risques de dommages causés par la végétation aux structures environnantes.

Tous les points évoqués permettent de réduire la sinistralité causée par le retrait-gonflement des argiles.

2.3.4 Évolutions réglementaires et loi Rousseau

Au sein de ce mémoire, les implications des évolutions réglementaires récentes sur les procédures de déclaration des arrêtés Cat Nat sont examinées. Ces changements ont rendu la déclaration de catastrophes naturelles plus aisée pour les communes. De ce fait, il est essentiel de considérer leur incidence sur la sinistralité des assureurs et de les intégrer dans l'analyse.
Certaines évolutions réglementaires restent encore en attente, et l'étude de ce mémoire s'arrête au contexte de début 2023, en attendant les clarifications éventuelles qui seront fournies par les instances gouvernementales par décret dans les mois à venir.

Dans cette optique, les réformes et lois suivantes seront abordées.

1. Réforme du régime des catastrophes naturelles, 2021

Parue au Journal Officiel le 28 décembre 2021, la loi relative à l'indemnisation des catastrophes naturelles [17] présente des modifications à application immédiate et d'autres applicables au 1er janvier 2023.

Parmi les modifications à application immédiate, il est à noter que le délai de dépôt d'un dossier de reconnaissance de l'état de catastrophe naturelle par les communes passe de 18 à 24 mois après la survenance de l'évènement. À l'inverse, le délai de publication au Journal Officiel de l'arrêté de reconnaissance de l'état de catastrophe naturelle se retrouve abaissé de trois à deux mois à compter du dépôt des demandes des communes.
Également, la transparence du processus décisionnel est améliorée : l'arrêté devra être motivé et les communes et/ou les sinistrés pourront solliciter la communication des documents ayant permis la prise de décision. Les recours gracieux, en cas de refus de reconnaissance, seront facilités.
De plus, un délégué à la reconnaissance de l'état de catastrophe naturelle est crée au niveau départemental et un référent Cat Nat est nommé dans chaque préfecture pour accompagner les communes dans leurs démarches. Par ailleurs, une commission nationale consultative des catastrophes naturelles est créée. Elle est chargée de rendre un avis chaque année sur la pertinence des critères retenus pour déterminer la reconnaissance de catastrophe naturelle ainsi que sur les conditions effectives de l'indemnisation des sinistrés.
Enfin, les sinistrés peuvent demander l'indemnisation jusqu'à cinq ans après la catastrophe naturelle contre deux auparavant.
Sans prise en compte de changement sur le risque sécheresse, ces différents éléments pourraient tout de même amener la sinistralité à augmenter lors de ces prochaines années.

D'autres modifications seront applicables au 1er janvier 2023.
L'assuré dispose désormais d'un délai de 30 jours (et non plus 10) à compter de l'arrêté pour déclarer le sinistre. A compter de la réception de la déclaration de sinistre (ou de la date de publication de l'arrêté si elle est postérieure), l'assureur doit prendre position et informer l'assuré sur la mise en jeu de la garantie Cat Nat et l'éventuelle mission d'expertise dans un délai d'un mois. Il doit, dans le mois suivant la réception de l'état estimatif transmis par l'assuré, ou du rapport d'expertise, proposer une indemnisation ou une réparation en nature. L'assureur devra communiquer le rapport d'expertise à l'assuré. À partir de l'accord de l'assuré sur la proposition d'indemnisation, l'indemnité doit être versée dans les 21 jours.

Les sinistrés qui résident dans des collectivités territoriales n'ayant pas adopté de Plan de Prévention des Risques Naturels (PPRN) ne seront plus pénalisés par une modulation de la franchise. Son montant sera fixé par décret. Auparavant, pour les communes vulnérables, il était crucial de faire une demande de PPRN car, depuis 2001, l'absence de PPRN pouvait entraîner une modulation des franchises en cas de sinistres répétés liés au même risque naturel.

Les frais de relogement d'urgence devront être pris en charge par tous les assureurs, tout comme les frais d'architecte ou de maîtrise d'ouvrage.

Il n'est plus sujet d'indemnisation des seuls dommages matériels directs mais d'indemnisation permettant de mettre un terme aux désordres existants. Tout refus d'assurance (ou résiliation compagnie) en raison de l'importance du risque de catastrophe naturelle qui pèse sur le bien, pourra être contesté devant le bureau central de tarification, lequel pourra imposer le contrat

à l'assureur.

A risque sécheresse équivalent, ceci devrait également augmenter fortement la sinistralité lors des prochaines années.

2. Ordonnance du 8 février 2023

L'ordonnance [22] vise à améliorer l'indemnisation des assurés subissant des sinistres consécutifs au retrait-gonflement des sols argileux.

Elle modifie le code des assurances afin :

— d'ajouter un nouveau mécanisme permettant la reconnaissance "Cat Nat" de communes ayant subi une succession anormale de sécheresses d'ampleur significative, mais dont l'intensité mesurée année par année ne remplit pas les critères actuels ;

— de préciser les conditions d'indemnisation des sinistres ;

— d'exclure du droit à la garantie "Cat Nat" les bâtiments construits sans permis de construire et les constructions neuves ne respectant pas la loi Élan du 23 novembre 2018 ;

— d'encadrer les conditions de réalisation de l'expertise désignée par les assureurs ;

— de fixer une obligation pour les assurés d'affecter l'indemnité perçue au titre d'un sinistre reconnu "Cat Nat" à la réalisation effective des travaux de réparation durable de leur habitation.

L'ordonnance sera applicable le 1er janvier 2024, sauf certaines dispositions qui entreront en vigueur au plus tard au 1er janvier 2025.

Ces évolutions complètent la réforme sur l'indemnisation des catastrophes naturelles, en vigueur depuis le 1er janvier 2023. Cette réforme améliore la transparence de la procédure de reconnaissance de l'état de catastrophe naturelle, favorise une indemnisation meilleure et plus rapide des sinistrés, et renforce les efforts de prévention face à ces phénomènes.

3. Loi Rousseau, 2023

Le 6 avril 2023, la loi Rousseau [25] a été votée à l'Assemblée nationale avec 115 voix pour et 9 voix contre. Cette loi est actuellement en cours de processus législatif au Sénat en vue de sa promulgation prochaine. Son objectif principal est de rééquilibrer la relation entre assureurs et assurés, ainsi que d'améliorer l'indemnisation des sinistres liés au retrait-gonflement des argiles. Voici un résumé des points clés de cette loi.

— La durée de la déclaration de catastrophe naturelle passe d'une saison à 12 mois. Cette modification est due au fait que certains sinistres peuvent survenir longtemps après une catastrophe naturelle.

— La durée de retour, qui était précédemment fixée à 25 ans, est désormais réduite à 10 ans. Bien que la loi ne spécifie pas explicitement l'historique à prendre en compte, dans le cadre de ce mémoire, il est supposé qu'il sera tenu compte des 5 valeurs minimales sur les 50 dernières années, il aurait également été possible de tenir compte des 2 valeurs minimales sur les 10 dernières années.

— La loi vise à rendre la procédure d'expertise plus efficace et impartiale, ainsi qu'à rééquilibrer les relations entre assureurs et assurés. Elle établit une présomption simple de causalité, selon laquelle lorsque l'état de catastrophe naturelle sécheresse est déclaré, il est présumé que le retrait-gonflement de l'argile est la cause déterminante des dommages. Cela entraînera une diminution significative du nombre de sinistres classés sans suite, c'est-à-dire des sinistres survenus lors d'une catastrophe naturelle dans une commune mais jugés non imputables à la catastrophe naturelle par l'assureur.

— Une expertise du sol, d'un coût estimé entre 1000 et 2000 euros selon le Figaro Immobilier [11], sera obligatoirement financée par l'assureur à la demande de l'assuré. Cette étude vise à fournir des informations approfondies sur les caractéristiques du sol afin de soutenir les demandes d'indemnisation.

— En cas d'impossibilité de réparer un bâtiment, l'assureur devra également prendre en charge les frais de reconstruction. Cette disposition peut représenter des coûts considérables pour les assureurs.

2.4 Évolutions réglementaires de la déclaration de catastrophe naturelle sécheresse

Depuis 1989, plusieurs évolutions réglementaires ont eu lieu concernant la déclaration de catastrophe naturelle due à la sécheresse.
À partir de décembre 2000, le "caractère Catastrophe Naturelle" éventuel des sécheresses a été apprécié selon une méthode d'analyse plus fine dite du « bilan hydrique à double réservoir » exigeant que soit établi, en plus du rapport géotechnique précédemment requis, un bilan hydrique destiné à mesurer la variation de la teneur en eau du premier mètre de sol et à déterminer si cette variation revêt un caractère d'intensité anormale.
À partir de septembre 2010, la Commission interministérielle a utilisé de nouveaux outils de mesure pour le calcul des critères de reconnaissance sécheresse basé sur l'indice d'humidité du sol (SWI – Soil Wetness Index) mesuré sur le maillage SAFRAN de 8 x 8 km et produit par Météo-France. Ces nouveaux outils de mesure sont utilisés par la Commission pour le traitement des dossiers depuis la sécheresse de l'année 2009.
Depuis 2019, ce sont les critères présentés dans la section précédente qui valent pour la déclaration de catastrophe naturelle sécheresse.
De plus, jusqu'en 2000, la sécheresse n'était pas considérée comme une catastrophe naturelle et était rattachée aux mouvements de terrains.

Malgré ces évolutions, le graphique ci-dessous met en évidence les principaux épisodes de sécheresse : celui des années 1989 à 1996, les sécheresses de 2003, 2011, 2017 à 2020, et dans une moindre mesure, celle de 2005.

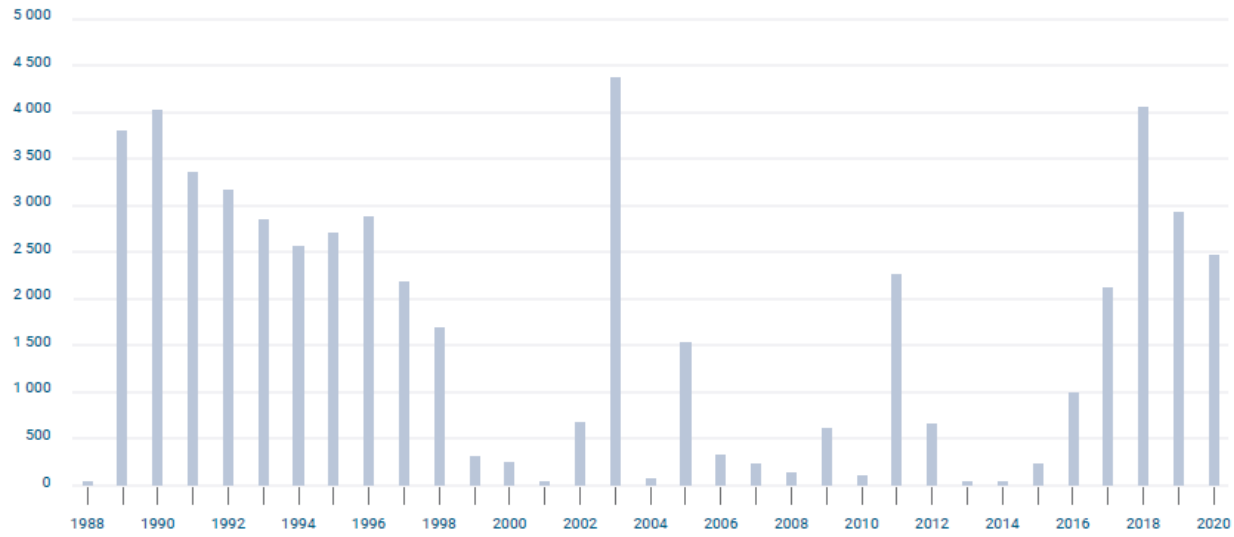


FIGURE I.6 – Nombre de déclarations catastrophes naturelles sécheresse entre 1982 et 2020, Bilan Cat Nat CCR 1982-2021 (CCR, 2023).

Sur la période 1989-2019, le coût moyen d'une reconnaissance sécheresse pour une commune s'élève à 261 K€.
À partir de 2015, le coût global de la sécheresse n'est pas encore consolidé, ce qui explique la marge d'erreur sur le coût moyen.
En comparaison aux années avant 2000, le coût moyen d'une reconnaissance Cat Nat sécheresse des 20 dernières années est en nette augmentation, vraisemblablement en raison de la mise en place des critères de reconnaissance à partir de décembre 2000.

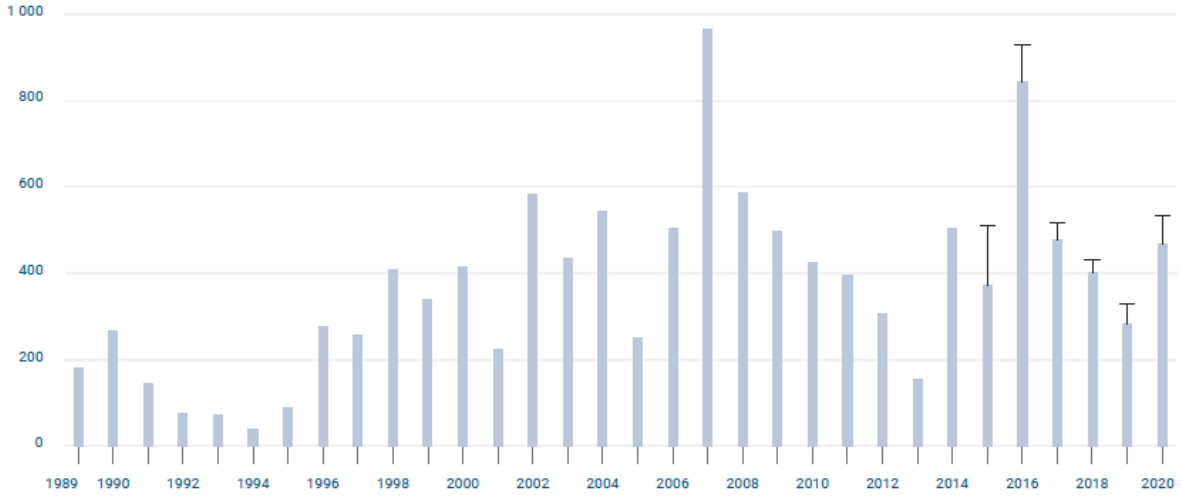


FIGURE I.7 – Coût moyen d'une reconnaissance catastrophe naturelle sécheresse entre 1982 et 2020 (en milliers d'euros), Bilan Cat Nat CCR 1982-2021 (CCR, 2023).

Chapitre 3

Étude proposée

Ici, un modèle de prévision de la sinistralité est développé pour le risque de subsidence en Auvergne-Rhône-Alpes. L'objectif est d'estimer le coût des sinistres pour chaque contrat d'assurance sur une période d'un an.
Ce travail s'articule autour de deux parties : la prédiction d'état de catastrophe naturelle, ainsi que la prédiction du coût observé sur le portefeuille d'un assureur.

Dans le cadre de cette recherche, il a été choisi de se concentrer sur la région Auvergne-Rhône-Alpes, en raison de l'ampleur considérable du problème et des contraintes liées aux ressources de calcul nécessaires pour une résolution globale. De plus, les données utilisées dans le cadre de cette étude sont principalement localisées en Auvergne-Rhône-Alpes, justifiant ainsi l'approche ciblée sur cette région.

L'étude réalisée dans le cadre de ce mémoire tente de modéliser, dans un premier temps, les SWI et la probabilité de respecter les critères de déclaration de catastrophe naturelle, et dans un second temps, le coût réellement observé chez un assureur. Le coût réel ne peut être observé que lorsque la catastrophe naturelle est observée, c'est pourquoi il est aussi possible de parler de coût conditionnel au respect des critères de déclaration catastrophe naturelle.

3.1 Probabilité de respecter les critères de catastrophe naturelle

Il sera dans la deuxième partie de ce mémoire proposé d'observer la probabilité de l'évènement catastrophe naturelle sécheresse. En effet, pouvoir estimer cette probabilité pourrait donner une information cruciale pour l'assureur car s'il n'y a pas de déclaration d'état de catastrophe naturelle, alors la garantie Cat Nat ne sera pas activée.

Estimer cette probabilité revient à estimer le fait les critères météorologique et géologique soient respectés à la date t pour la commune x. Le critère géologique est déterministe. Ainsi, c'est la probabilité que le critère météorologique soit respecté en t, où t représente ici une saison et un horizon (par exemple : "hiver 2024"), qui sera étudiée.
Autrement dit, comme vu dans la première partie, c'est la probabilité que le SWI d'au moins une des mailles ait une durée de retour de 25 ans au cours des mois de janvier, février et mars 2024 qui sera étudiée.

$$\begin{aligned}\mathbb{P}(\text{Cat Nat}(t,x)) &= \mathbb{P}(\text{Critère\_météorologique\_validé}(m,y)) \cdot 1_{\text{Critère\_géologique}}(x) \\ &= \mathbb{P}\left(\bigcup_{(\text{mailles } y \text{ de la commune } x)} \; \bigcup_{(\text{mois } m \text{ de la saison } t)} \text{SWI}(m,y) < \text{seuil}(m,y) \right) \\ &\cdot 1_{\text{Critère\_géologique}}(x)\end{aligned}$$

Où seuil(m,y) est la deuxième plus petite valeur du SWI sur la période de $t-50$ à $t-1$ pour la maille y pour le mois m. Dans le cas où la loi Rousseau serait promulguée, le seuil serait ici la cinquième plus petite valeur sur la période.
Et où $1_{\text{Critère\_géologique}}(x)$ représente si le critère géologique est respecté ou non. Il sera vu par la suite que ce critère est déterministe et ne dépend que de x.

Une comparaison entre les seuils actuels et les seuils proposés par la loi Rousseau sera effectuée afin de voir l'impact de la loi Rousseau sur cette probabilité en 3.2.3.

3.2 Estimation du coût conditionnel

Afin d'estimer l'espérance du coût conditionnellement au fait que les critères de catastrophe naturelle soient vérifiés, les données sinistres d'un assureur présent en Auvergne-Rhône-Alpes ont été étudiées.
Le caractère temporel et spatial de la sécheresse a conduit à une modélisation de $C(t,x)|\text{Cat Nat}, X$, la fonction du coût de l'assureur pour :

— une date $t \in \mathbb{R}^+$ (en pratique : un mois, une saison ou une année) ;

— à une position $x \in \Omega \subset \mathbb{R}^2$ (en pratique : la commune) ;

— pour un profil de risque $X = (X1, X2, ..., Xn)$ décrivant les caractéristiques du contrat (nature de l'habitation, superficie, localisation) ;

— et conditionnellement au déclenchement d'une catastrophe naturelle dans la commune.

Comme il sera vu dans la troisième partie, la base de données très déséquilibrée conduit à la modélisation de la fonction de coût de l'assureur suivante :

$$\begin{aligned}\mathbb{E}[C(t,x)|\text{Cat Nat}, X] &= \mathbb{E}[C(t,x)|\text{Cat Nat}, X, C(t,x) > 0] \cdot \mathbb{P}(C(t,x) > 0|\text{Cat Nat}, X) \\ &+ \mathbb{E}[C(t,x)|\text{Cat Nat}, X, C(t,x) = 0] \cdot \mathbb{P}(C(t,x) = 0|\text{Cat Nat}, X) \\ &= \mathbb{E}[C(t,x)|\text{Cat Nat}, X, C(t,x) > 0] \cdot (1 - \mathbb{P}(C(t,x) = 0)|\text{Cat Nat}, X)\end{aligned}$$

Différents modèles seront étudiés en III afin d'estimer :

$\mathbb{E}[C(t,x)|\text{Cat Nat}, X, C(t,x) > 0]$,
$\mathbb{P}(C(t,x) = 0)|\text{Cat Nat}, X)$.

Deuxième partie

Étude des SWI et du critère météorologique

L'étude des SWI sera menée dans cette partie. La projection des SWI des différentes mailles, tenant compte de leurs dépendances mutuelles, permettra de modéliser le critère météorologique, mais également d'être intégrée dans la troisième partie de ce mémoire.

Ce problème complexe (grande dimension) nécessitera de réaliser plusieurs hypothèses et présente certaines limites qui seront présentées en 1.3.

Il a été vu en 2.3.1 le fonctionnement du critère météorologique.
Il est important de souligner qu'**il est possible d'atteindre des niveaux de sécheresse relativement intenses sans toutefois déclencher le critère météorologique au niveau de la commune**. En effet, il suffit que la commune ait connu auparavant deux épisodes de sécheresse encore plus intenses pour ne pas remplir le critère météorologique. Par exemple dans une commune ayant connu une sécheresse en 2003 et en 2022, et ayant donc atteint des SWI extrêmement bas lors de ces deux étés, de l'ordre de 0,2 pour l'ensemble des mailles la composant. Il se pourrait qu'une sécheresse intense frappe la commune en 2024, et que les SWI atteints par les mailles soient de l'ordre de 0,25. Dans ce cas, certaines habitations connaîtraient sans doute des dégâts. Cependant, le critère météorologique ne serait pas respecté. Dans ce cas là, l'assuré ne bénéficierait d'aucune indemnisation au titre de la garantie Cat Nat.
Il est ainsi cohérent d'intégrer la modélisation du critère météorologique dans le cadre de ce mémoire car il conditionne l'indemnisation versée par l'assureur.

Il est alors proposé d'estimer la probabilité à horizon un an que les critères de déclaration de catastrophe naturelle sécheresse (météorologique et géologique) soient respectés.
Le problème de la modélisation du critère météorologique équivaut mathématiquement à modéliser la fonction de répartition jointe des SWI pour les mailles composant la région Auvergne-Rhône-Alpes comme vu en 3.1.

Chapitre 1

Grande dimension et réduction de la dimension des mailles

Avant de se lancer dans une modélisation, il est utile de remarquer que dans le cadre de la modélisation du critère météorologique, les projections basées sur les scénarios du GIEC ne sont pas exploitables.

Des données de SSWI (*Standardized Soil Wetness Index*) sont mises à disposition sur le site DRIAS par Météo-France [4], issues de l'expérience CLIMSEC selon différents scénarios du GIEC :

— RCP *(Representative Concentration Pathway)* 2.6 : scénario visant à réduire les concentrations de CO2 par des politiques climatiques ;

— RCP 4.5 : scénario visant à stabiliser les concentrations de CO2 par des politiques climatiques ;

— RCP 8.5 : scénario sans politiques climatiques.

Cependant, ces données représentent des écarts à la moyenne (standardisée) par rapport à la période 1958-2008.
Ces valeurs sont aujourd'hui moins fiables, car le paysage des sinistres a considérablement évolué depuis cette période. Il serait néanmoins possible d'utiliser le scénario 4.5 ou 8.5 pour pallier ce problème.
Toutefois, le problème majeur ici est que ces projections fournissent des valeurs probables à horizon souhaité, mais ne fournissent pas de distribution des variables, ni d'intervalle de confiance. Autrement dit, les données ne fournissent qu'une seule trajectoire par maille géographique, ce qui ne permet pas d'estimer la probabilité conjointe souhaitée mentionnée précédemment 3.1.
Bien que le projet fasse référence à diverses sources d'incertitude et caractérise leur variabilité spatio-temporelle, aucune information chiffrée sur cette question n'est disponible.
Ainsi, dans le cadre de ce mémoire, les données du projet CLIMSEC de Météo-France ne sont pas utilisées. Il a donc été choisi de développer un modèle de projection dans cette étude.
Ces données pourraient être exploitées dans le but d'expliquer la sinistralité du risque sécheresse observé en MRH comme dans le mémoire d'Alexandre Delorme [8]. Cependant, pour expliquer la probabilité de déclenchement du critère météorologique, il est nécessaire de recourir à la modélisation.

Afin de modéliser les SWI, il est important de souligner que les données proposées par Météo-France sont les données de 8981 mailles. Cependant les observations de SWI se font sur une période de 53 années comportant chacune 12 mois, soient 636 points. Bien que l'étude soit réduite à la région d'Auvergne-Rhône-Alpes, il demeure toujours plus de 3000 mailles pour lesquels il faut prédire le SWI, qui sont observés seulement 636 fois. Ainsi, la modélisation fait ici face à un problème de grande dimension.

Un des moyens de faciliter le problème est avant toute chose de réduire la dimension des mailles.

L'analyse en composantes principales (ACP) émerge alors comme une technique puissante et polyvalente pour explorer et compresser l'information contenue dans des ensembles de données complexes. Dans ce chapitre, l'importance cruciale de la réduction de dimension sera mise en lumière dans le contexte de l'analyse en composantes principales, en mettant en évidence les avantages et l'utilité de cette approche.

1.1 Statistiques descriptives sur les mailles

Afin de visualiser le problème, la figure II.1 montre une représentation du SWI pour les mailles 20, 300, 400, 500, 1000 et 1200, qui sont relativement éloignées géographiquement.

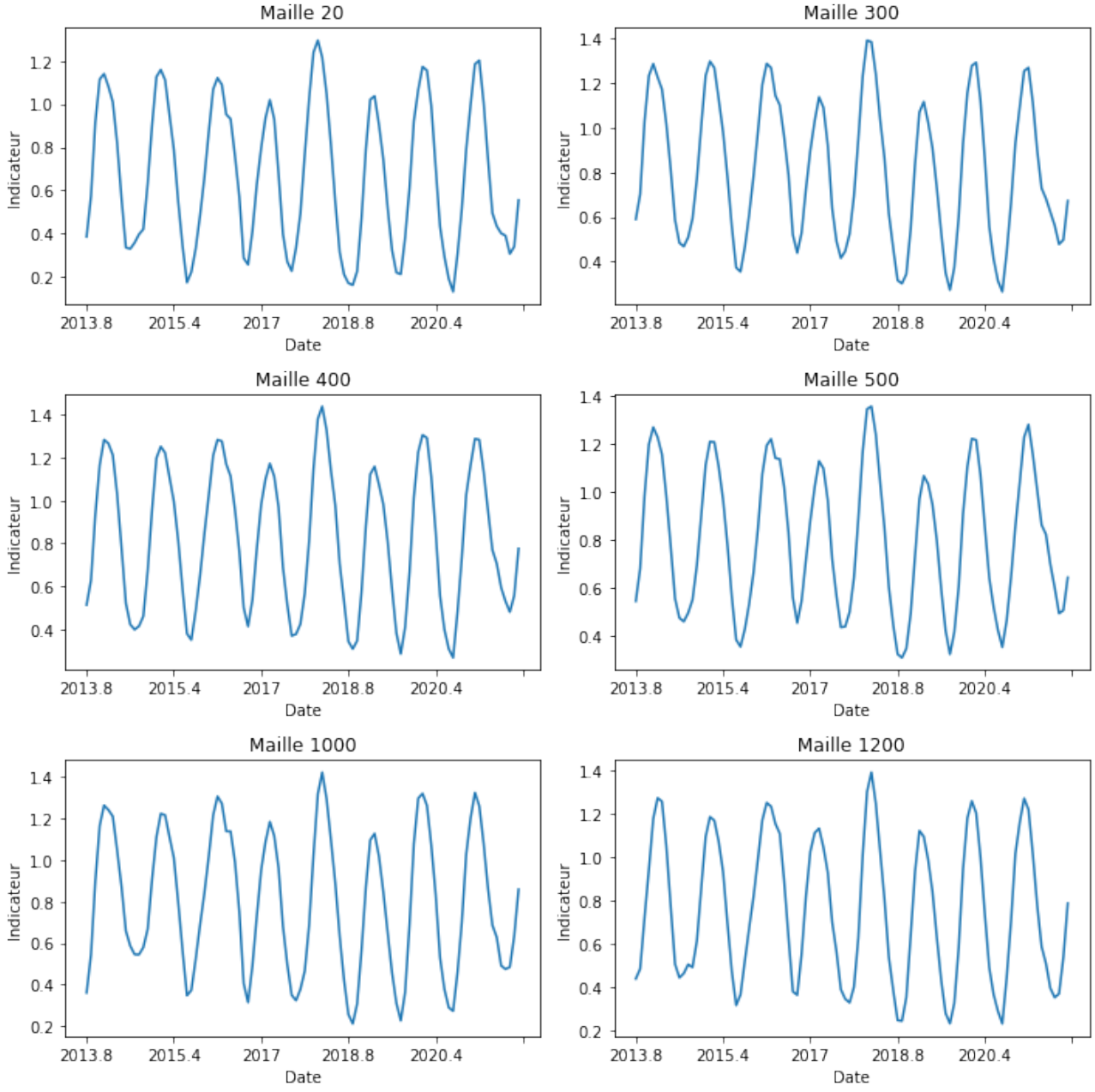


FIGURE II.1 – Représentation du SWI en fonction du temps (année.mois) pour les mailles 20, 300, 400, 500, 1000 et 1200.

Conjoncture

La première conjoncture qui peut en être tirée est que le phénomène est assez régulier et il est très similaire d'une maille à l'autre. Il semble aussi que les perturbations soient assez globales. Par exemple, pour le printemps 2018, il semble que toutes les mailles aient connu une période anormalement humide.
Il conviendrait alors de classer ces mailles par groupe.

1.2 Analyse en composantes principales

Il s'est avéré que la prédiction du SWI dans le contexte expliqué ci-dessus est un problème de grande dimension. La méthode d'analyse en composantes principales est donc utilisée. Cette méthode, décrite en annexe A, permet d'expliquer des variables aléatoires en les projetant dans des espaces de dimensions inférieures.

Tout d'abord, la figure II.2 représente la variance expliquée en fonction des composantes principales. Elle montre que l'utilisation d'un seul axe principal permet de restituer environ 60 % de la variance de la variable aléatoire.

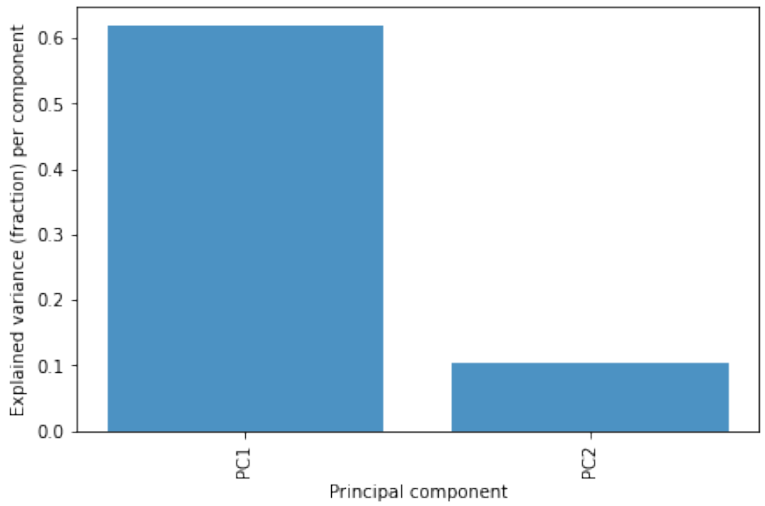


FIGURE II.2 – Variance expliquée sur les deux premiers axes principaux.

La valeur des SWI de chaque maille peut alors être représentée sur une droite représentant une période, combinaison linéaire des 636 périodes initiales. Il est possible de voir cette droite comme un indicateur d'humidité globale de la maille sur une année entière.
Afin de conserver une interprétabilité, il a été choisi d'utiliser la valeur entière de la projection sur ce premier axe principal pour réussir à faire du regroupement (*clustering*). En effet, le premier axe principal est divisé en 21 morceaux de tailles égales, et ainsi 21 groupes de mailles différents sont établis. Ce choix sera confirmé par la suite lorsque les mailles seront représentées géographiquement.
Les groupes les plus petits (0,1,2...) sont des groupes de mailles plutôt secs, tandis que les groupes de mailles plus grands (17,18,19,20) sont des groupes de mailles plutôt humides.

Plusieurs représentations graphiques permettent de visualiser cette modélisation. Par exemple, la figure II.3 permet de visualiser le décalage entre la maille représentée et le groupe à laquelle elle appartient. La figure II.4 présente toutes les mailles qui ont été regroupées sous le groupe 1 et les SWI moyens de chaque groupe en fonction du temps sont ensuite représentés dans la figure II.5.

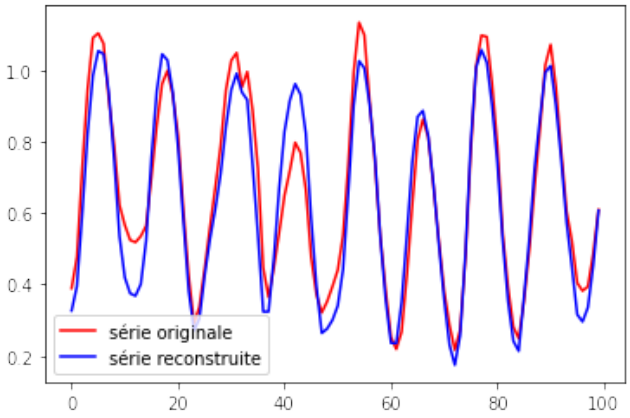


FIGURE II.3 – Comparaison entre série originale et série du groupe de mailles attribué sur les cent dernières dates.

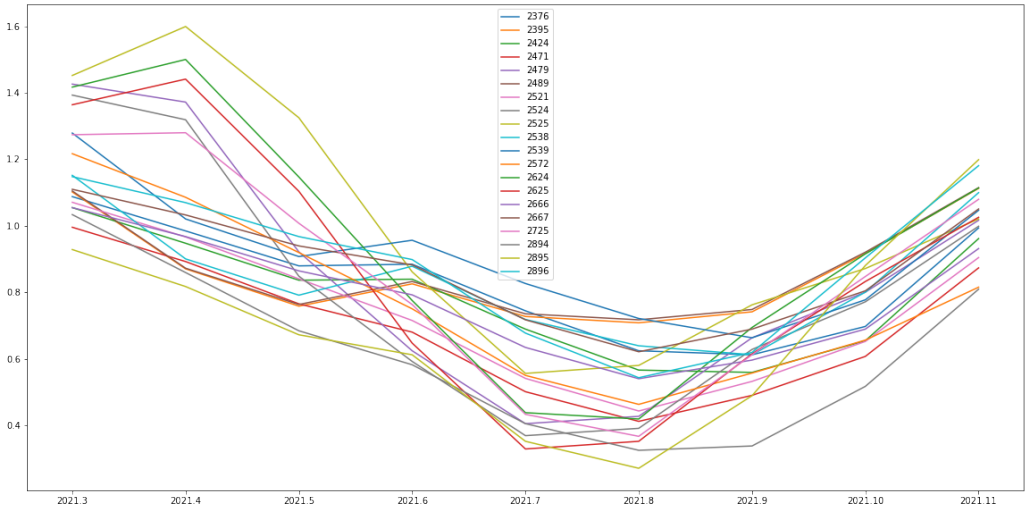


FIGURE II.4 – Représentation SWI en fonction du temps des mailles appartenant au groupe de mailles 1.

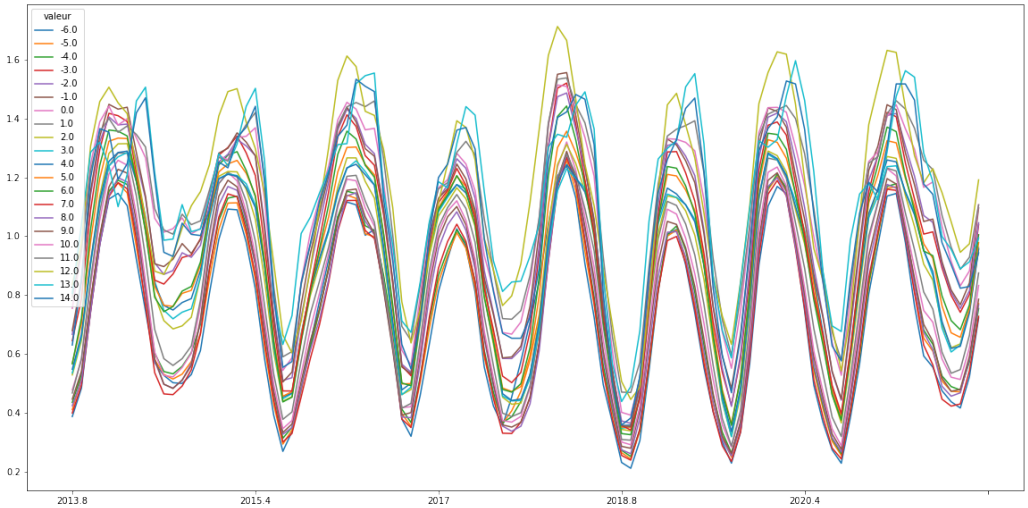


FIGURE II.5 – Représentation du SWI moyen des 21 groupes construits en fonction du temps.

L'interprétabilité du premier axe principal comme un axe qui représente l'humidité globale d'une maille au cours d'une année entière semble vérifiée.

De plus, voici les groupes de mailles obtenus en Auvergne-Rhône-Alpes :

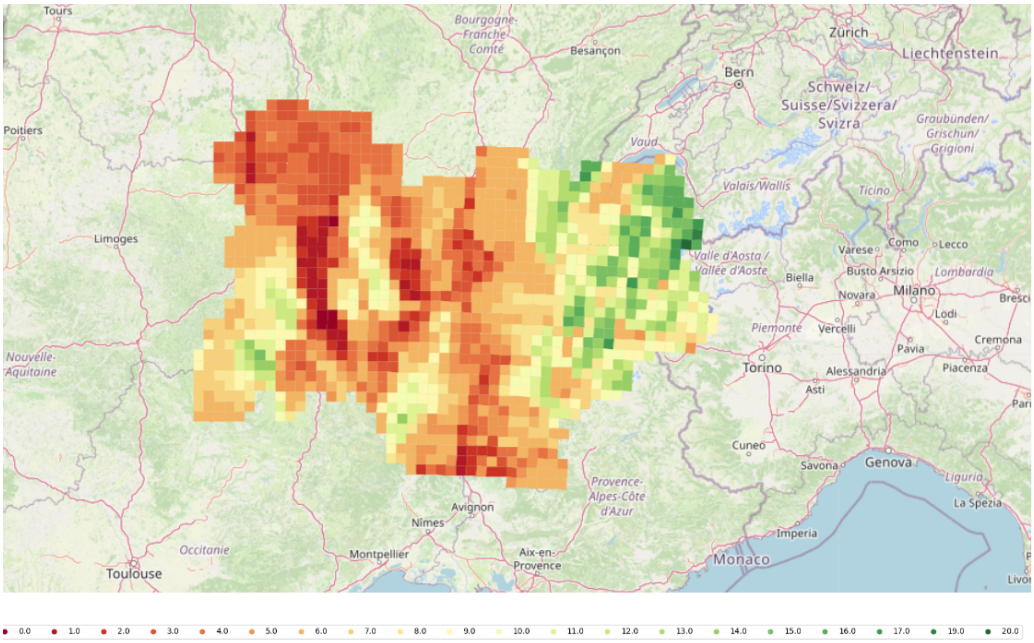


FIGURE II.6 – Représentation spatiale des 21 groupes.

Le regroupement effectué semble alors avoir du sens, puisqu'il réunit des mailles proches géographiquement dans des groupes identiques ou relativement proches.
De plus, certaines similitudes avec la carte d'exposition BRGM présentée en I.5 sont retrouvées. En effet, l'Allier et le Puy-de-Dôme semblent être classés dans des groupes de mailles plutôt secs, tandis que les départements de Savoie et Haute-Savoie semblent être classés dans des groupes de mailles plutôt humides.

Enfin, les corrélations entre les séries temporelles de quelques groupes de mailles formés avec l'analyse en composantes principales sont observées ci-dessous :

Groupe mailles	**1**	**2**	**3**	**4**	**5**
1	1,000	0,995	0,988	0,981	0,969
2	0,995	1,000	0,998	0,993	0,986
3	0,988	0,998	1,000	0,997	0,991
4	0,981	0,993	0,997	1,000	0,997
5	0,969	0,986	0,991	0,997	1,000

De fortes corrélations entre les groupes sont observées. Il est donc impossible de prédire les SWI des différents groupes de mailles de manière indépendante. Il est essentiel que le modèle retenu soit capable de rendre compte de ces corrélations, sans quoi les projections seront biaisées.

1.3 Simplification du problème

Au final, le problème initial est simplifié de la manière suivante :
Les plus de 3000 mailles d'Auvergne-Rhône-Alpes sont classées en 21 groupes différents. Pour

chaque groupe, la moyenne du SWI de ce groupe est calculée et le seuil est défini de la même manière qu'en première partie (durée de retour de 25 ans).
L'approximation d'un comportement identique pour toutes les mailles appartenant à un même groupe est faite.
Par exemple, si une maille du groupe 1 passe sous le seuil de durée de retour 25 ans, alors toutes les autres mailles du groupe 1 passeront en même temps sous ce seuil.

Conclusion

Ainsi, pour chaque commune, il suffit de regarder les différents groupes de mailles qui s'intersectent et d'observer la probabilité qu'au moins un des groupes passe sous le seuil afin d'estimer la probabilité que la commune respecte le critère météorologique.

Le problème initial est réécrit de la manière suivante :

$$\mathbb{P}(\text{Cat Nat}(t,x)) = \mathbb{P}\left(\bigcup_{(\text{groupes\_mailles } i \text{ commune } x)} \; \bigcup_{(\text{mois } m \text{ saison } t)} SWI_{groupe_i}(m) < seuil_{groupe_i}(m) \right) \cdot 1_{\text{Critère géologique validé}}(x)$$

Désormais, la dimension du problème est réduite, la question de la modélisation des variables aléatoires $SWI_{groupe0}$ à $SWI_{groupe20}$ se pose. Par ailleurs, il convient de rappeler que cette modélisation doit tenir compte de la dépendance mutuelle de ces variables.

Chapitre 2

Modèle Long Short-Term Memory (LSTM)

2.1 Motivations

Il existe des méthodes classiques pour la modélisation des séries temporelles, connues sous le nom de processus autorégressifs. Ces processus peuvent être généralisés au cas multidimensionnel, connus sous le nom de vecteurs autorégressifs. Cependant, ces approches sont limitées à la détection de motifs linéaires. Dans le contexte de données météorologiques, telles que le SWI, qui sont sujettes à des phénomènes irréguliers, en particulier dans le contexte du changement climatique actuel, les méthodes traditionnelles pourraient ne pas suffire à capturer ces phénomènes complexes sous-jacents. C'est pourquoi une transition vers des modèles plus performants, tels que ceux basés sur l'apprentissage automatique, a été envisagée.
Par ailleurs, les fonctions régressives peuvent être vues comme un cas particulier de *Recurrent Neural Network* (RNN). Une idée de la preuve est donnée en annexe B. Ainsi, l'emploi d'un RNN ne serait pas moins performant qu'un modèle "classique".
Les réseaux de neurones récurrents (RNN) offrent une solution pour surmonter les limitations souvent rencontrées par les méthodes classiques dans la modélisation des séries temporelles. Par exemple, contrairement aux méthodes classiques qui nécessitent souvent la stationnarisation des données en entrée, les RNN peuvent traiter des données non stationnaires sans perte d'information, évitant ainsi le phénomène de perte de mémoire présenté dans cet article [20]. De plus, les RNN sont capables de capturer des relations non linéaires qui échapperaient aux modèles autorégressifs traditionnels.

Cependant, les réseaux de neurones récurrents peuvent rencontrer un problème appelé le "*gradient exploding*" lors de l'entraînement. Ce problème survient lorsque les gradients deviennent très grands au fur et à mesure qu'ils sont rétropropagés dans le temps. Lorsque cela se produit, les poids du réseau peuvent être mis à jour de manière significative, entraînant des instabilités dans l'apprentissage. C'est ce qui a été constaté dans le cadre de ce mémoire.
Les LSTM sont conçus pour pallier aux problèmes de *gradient exploding*, ainsi qu'aux problèmes de *gradient vanishing*. Le mécanisme de portes dans les LSTM, y compris la porte d'oubli, est conçu pour réguler le flux d'informations à travers la cellule mémoire, permettant ainsi un meilleur contrôle sur la propagation des gradients.

Pour ces raisons, le choix d'un modèle LSTM semble donc être le plus approprié.
Le fonctionnement de ce modèle est présenté dans la section suivante 2.2.

2.2 Présentation d'un LSTM

Architecture d'une couche LSTM

Chaque couche LSTM est composée de plusieurs unités LSTM, ou cellules. Chaque cellule LSTM possède trois portes d'entrée (*Forget Gate*, *Input Gate*, *Output Gate*), une cellule mémoire c_t (aussi appelée *"cell state"*) et un état caché h_t (aussi appelé *"hidden state"*). Les portes régulent le flux d'informations dans et hors de la cellule mémoire, tandis que la cellule mémoire conserve des informations à long terme et que l'état caché représente l'information apprise par le modèle à l'instant t dans la séquence.

Porte d'oubli (*Forget Gate*) :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

où σ est la fonction sigmoïde, W_f une matrice de poids et b_f le biais.
La porte d'oubli décide quelles parties de la cellule mémoire c_{t-1} doivent être oubliées en fonction de l'état caché précédent h_{t-1} et de l'entrée actuelle x_t.

Porte d'entrée (*Input Gate*) :

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

où σ est la fonction sigmoïde, W_i une matrice de poids et b_i le biais.
La porte d'entrée régule la quantité d'informations à ajouter à la cellule mémoire.

Porte de sortie (*Output Gate*) :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

où σ est la fonction sigmoïde, W_o une matrice de poids et b_o le biais.
La porte de sortie contrôle la sortie de la cellule mémoire c_t vers la couche cachée h_t.

La mise à jour de la cellule mémoire se fait en deux étapes :

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t$$

où $g_t = \tanh(W_g \cdot [h_{t-1}, x_t] + b_g)$, σ est la fonction sigmoïde, W_g une matrice de poids et b_g le biais.

Voici une représentation d'une couche LSTM issue du la revue de l'IA [12] :

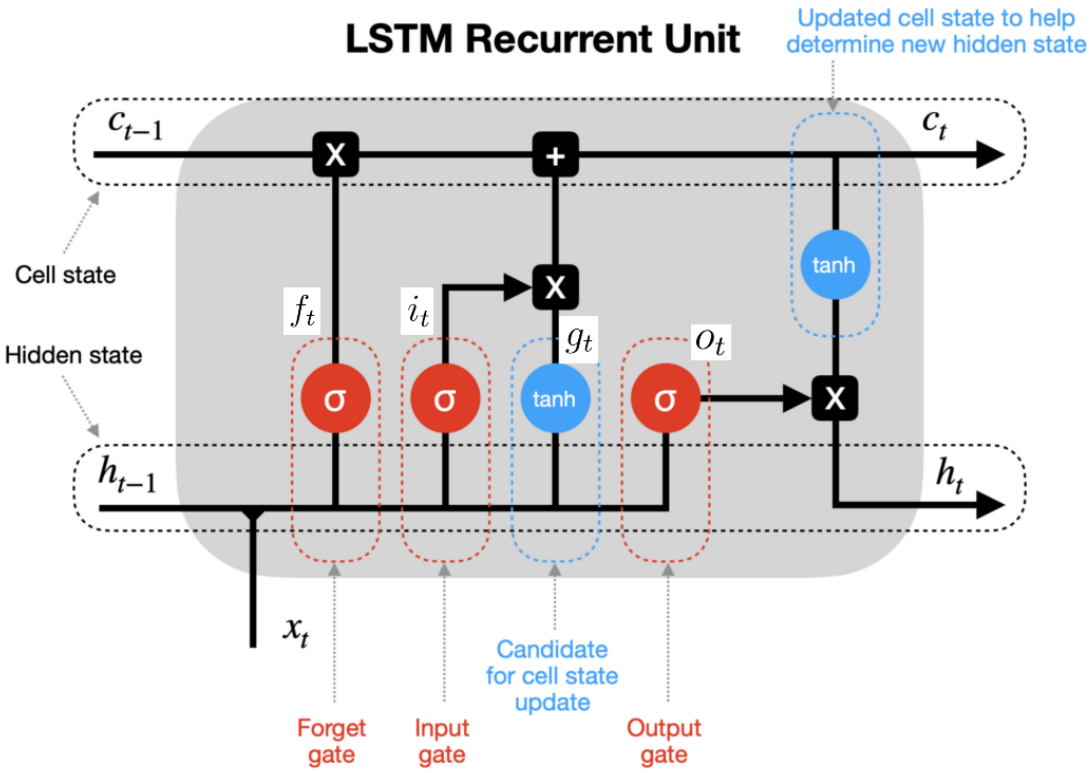


FIGURE II.1 – Schéma explicatif du fonctionnement d'une cellule LSTM.

Fonctions d'activation

Les fonctions d'activation régulent l'activation des portes et la mise à jour de la cellule mémoire. La fonction sigmoïde σ est souvent utilisée pour les portes, limitant les valeurs entre 0 et 1. La tangente hyperbolique tanh est utilisée pour la mise à jour de la cellule mémoire, produisant des valeurs entre -1 et 1.

Architecture multi-couche

Un modèle LSTM peut être composé de plusieurs couches empilées les unes sur les autres. Chaque couche prend en entrée les sorties de la couche précédente. L'architecture multi-couche permet au modèle d'apprendre des représentations hiérarchiques et complexes.
Les paramètres de chaque couche LSTM sont appris indépendamment lors de la phase d'entraînement et la rétropropagation du gradient se propage à travers toutes les couches lors de la phase d'optimisation.

Calcul de la Prédiction Finale dans un LSTM

La prédiction $\hat{y}_t$ issue d'un modèle LSTM est généralement calculée à partir de l'état caché final h_t à la dernière étape temporelle. Pour calculer $\hat{y}_t$, une couche entièrement connectée (*"fully connected"*) est souvent ajoutée au-dessus de l'état caché final. Mathématiquement, cela peut être écrit de la manière suivante :

$$\hat{y}_t = \text{Activation}(W \cdot h_t + b)$$

où W représente la matrice de poids, h_t est l'état caché final, b est le biais et Activation est la fonction d'activation appliquée à la sortie de la couche.

Cette opération permet d'obtenir la prédiction finale $\hat{y}_t$, qui est ensuite comparée à la vraie sortie y_t pour le calcul de la perte, contribuant ainsi à l'ajustement des paramètres du modèle au cours de l'entraînement.

Entraînement, Rétropropagation et Algorithme Adam *(Adaptive Moment Estimation)*

L'entraînement d'un modèle LSTM implique plusieurs étapes.

— **Propagation avant (*Forward Propagation*) :** Les données sont alimentées à travers le réseau et les sorties sont calculées.

— **Calcul de la perte (*Loss Calculation*) :** La différence entre les sorties prédites et les vraies sorties est mesurée à l'aide d'une fonction de perte. Par exemple, lors de problèmes de régression où la tâche consiste à prédire une valeur numérique, la fonction de perte moyenne quadratique (*Mean Squared Error* - MSE) est souvent utilisée. Elle mesure la moyenne des carrés des différences entre les valeurs prédites et les valeurs réelles.

La formule de la fonction de perte moyenne quadratique pour un *batch* de données est donnée par :

$$Loss = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2$$

où n est la taille du *batch*.

— **Rétropropagation du gradient (*Backpropagation*) :** Les gradients de la perte par rapport aux paramètres du réseau sont calculés. Cela se fait en utilisant la rétropropagation à travers le temps (BPTT, *Backpropagation Through Time*) pour prendre en compte les dépendances temporelles. Cet algorithme fondamental est décrit dans le papier [15]. Par ailleurs, c'est ici l'algorithme Adam [1] qui a été utilisé pour ajuster les paramètres du modèle de manière efficace.

2.3 Préparation des données d'entraînement

Parmi les 21 séries moyennes de SWI constituées à partir de l'ACP décrite en II.6, plusieurs échantillons seront constitués. Des fenêtres (parties des séries moyennes) de tailles précisées par la suite, représentées matriciellement afin de tenir compte de la dépendance entre les groupes de mailles, seront utilisées (valeurs en colonnes, groupes de mailles en lignes).
Afin d'obtenir une taille de fenêtre cohérente, il est proposé de prendre une taille de fenêtre égale à 12 (t-11 à t). En effet, il semble cohérent que les valeurs de SWI des 12 derniers mois aient un impact sur la prochaine valeur observée, mais au-delà, l'effet est supposé négligeable.
Ainsi, les données d'entrée sont des matrices de taille 20*12 correspondant aux SWI des 21 groupes de mailles sur 12 unités de temps consécutives.
10000 fenêtres sont ici tirées aléatoirement parmi les fenêtres constituées.
En réalité, les données d'entrée sont représentées sous forme de tenseur. Ce travail est réalisé à l'aide de la librairie pythons.

L'entraînement du réseau consiste donc à prédire un vecteur de taille 20, correspondant à la valeur des SWI de chaque groupe de mailles à l'instant t+1.

Les données sont ensuite normalisées entre 0 et 1 et sont divisées en un ensemble d'apprentissage, un ensemble de test et un ensemble de validation (80 % - 10 % - 10 %).

2.4 Choix des paramètres et de l'architecture du modèle

Configurer les réseaux de neurones est difficile car il n'existe pas de bonne théorie sur la manière de le faire.
La méthode utilisée a donc, ici, consisté à optimiser d'abord le nombre d'*epoch*, puis la taille du *batch* et enfin le nombre de neurones.
Dans cette section, les paramètres et l'architecture du modèle sont présentés.

Nombre d'*epoch*

Un passage complet du jeu de données d'entraînement par l'algorithme désigne une *epoch*.
L'erreur correspond à la somme des erreurs d'entraînement sur le *batch*. Le *batch* étant de grande taille, l'échelle est également grande.

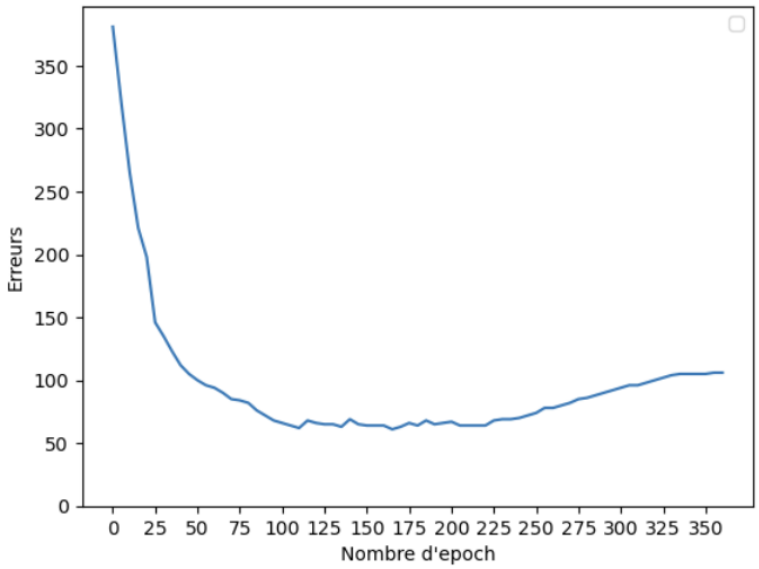


FIGURE II.2 – Erreur moyenne du LSTM sur l'ensemble de test en fonction du nombre d'*epoch*, *batch* fixé à 100, nombre de neurone fixé à 2.

Le nombre d'*epoch* optimal est ici égal à 112 environ.

Taille du *batch*

En apprentissage automatique, un *batch*, est un ensemble de données utilisé pour entraîner un modèle de manière itérative. Plutôt que de mettre à jour les poids du modèle après chaque exemple de données, le modèle est mis à jour après avoir parcouru un groupe d'exemples.
En fixant le nombre d'*epoch* à 112 et le nombre de neurones à 2, le nombre optimal du *batch* obtenu est de 200.

Nombre de neurones

Le nombre de neurones affecte la capacité d'apprentissage du réseau. En général, un plus grand nombre de neurones serait capable d'apprendre plus de structures à partir du problème au détriment d'un temps d'entraînement plus long. Une capacité d'apprentissage plus élevée crée également le problème du surapprentissage potentiel des données d'entraînement.
En fixant le nombre d'*epoch* à 112 et la taille du *batch* à 200, le nombre optimal de neurones

est de 200.

Pour résumer, voici la configuration optimale retenue :

Nombre d'*epoch*	Taille du *batch*	Nombre de neurones
112	200	200

TABLE II.1 – Optimisation du modèle LSTM.

Nombre de Couches (Profondeur)

Il est possible de choisir l'ordre du RNN à partir du diagramme *Partial Autocorrelation Function* (PACF). Il est, en effet, possible de faire le parallèle entre l'ordre du RNN et les fonctions autorégressives comme expliqué en annexe C.

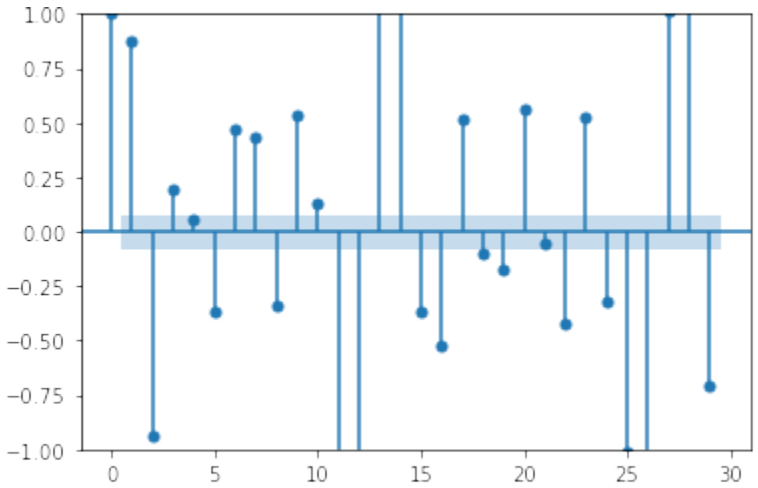


FIGURE II.3 – Diagramme des autocorrélations partielles (PACF).

Le nombre de couches est choisi à partir du graphique ci-dessus comme première valeur qui annule la fonction d'autocorrélation partielle comme l'ordre d'un processus autorégressif AR(p) serait choisi. La valeur 4 est alors retenue.

Connexions entre les Couches

Le schéma de n'annexe 2.2 présente les connexions entre les couches : fonction tangente hyperbolique et sigmoïde.

2.5 Évaluation

Afin de valider le modèle, le *Root Mean Square Error* (RMSE) du modèle naïf est comparé au RMSE du modèle LSTM sur l'ensemble de validation.

Le modèle naïf est le modèle qui prend comme meilleure estimation de la valeur du SWI à l'instant t, la valeur du SWI à l'instant t-1 : $SWI_t = SWI_{t-1}$.
Ce modèle "oublie" de tenir compte des phénomènes de saisonnalité et de "retour à la normale"

qui pourraient être pris en compte dans un modèle plus complexe. Ce modèle, volontairement basique, devrait donc être moins bon que les LSTMs.
Il permet de valider la cohérence des résultats obtenus avec ces derniers.

Le RMSE est estimé à partir d'une validation croisée réalisée en 5 itérations. Voici les résultats des RMSE finaux obtenus pour les différents modèles :

Modèle	RMSE
Naïf	0,10
LSTM	0,04

TABLE II.2 – RMSE des différents modèles.

Le modèle LSTM obtient un RMSE plus bas que le modèle naïf et le modèle autorégressif.

2.6 Conclusion

Il a été vu dans cette partie que la projection des SWI constituait un problème complexe. En effet, le nombre de mailles est plus élevé que le nombre d'instants où les SWI sont observés, ce qui fait de ce problème un problème de grande dimension. C'est pourquoi il a d'abord fallu réduire la dimension des cellules.
Par la suite, l'emploi d'un LSTM a été retenu pour modéliser le comportement des SWI des 21 groupes de cellules.

Cette modélisation permet désormais de projeter le SWI à horizon souhaité pour les 21 différents groupes de mailles.
Ainsi, en connaissant l'historique des 50 dernières années, il devient également possible d'estimer quand et avec quelle probabilité, le critère météorologique sera respecté pour chaque ville.
Le modèle propose un RMSE relativement faible. Cependant, le RMSE concerne les prédictions de SWI et non pas directement le critère météorologique.
Ainsi, les projections de SWI pourront être utilisées dans la troisième partie de ce mémoire mais il est important de valider la projection du critère météorologique dans la suite de cette partie.
Le prochain chapitre présente donc les probabilités de respecter le critère météorologique à partir des projections du LSTM et propose une comparaison entre les prédictions réalisées par le modèle et la réalitée observée ces dernières années.

Chapitre 3

Résultats

Il a ici été choisi d'utiliser les projections de SWI à horizon un an.
Les données de SWI fournies par Météo-France s'arrêtant au 31 décembre 2021 à l'heure de la création de ce modèle, les résultats proposés ci-dessous correspondent aux estimations de SWI de janvier à décembre 2022 pour chaque groupe de mailles présenté ci-dessus.
Un historique actualisé de données permettrait de projeter les SWI pour des dates plus récentes.
Également, le modèle serait théoriquement capable de projeter à horizon plus lointain, mais la fiabilité des résultats serait alors entachée.
Voici par exemple 10 projections proposées par le modèle pour le groupe de mailles 10 en II.1 et la moyenne de la prédiction confortée par un intervalle de confiance (quantile à 5% et quantile à 95%) en II.2.

Le LSTM semble avoir bien "capté" la saisonnalité en reproduisant une forme sinusoïdale pour tous les scénarios.
De même, les périodes inter-saisons semblent être assez similaires. En effet les augmentations ou diminutions sont relativement proches pour les mois de printemps et d'automne.
Cependant, certains scénarios semblent être plus ou moins extrêmes aussi bien l'été que l'hiver. En effet, entre juillet et septembre, il semble que la sécheresse arrive plus ou moins tard selon les scénarios et avec une intensité plus ou moins forte. De même, entre janvier et mars, les scénarios semblent donner des résultats assez différents, où la recharge en eau peut se faire plus ou moins fortement et avec une apparition plus ou moins délayée.

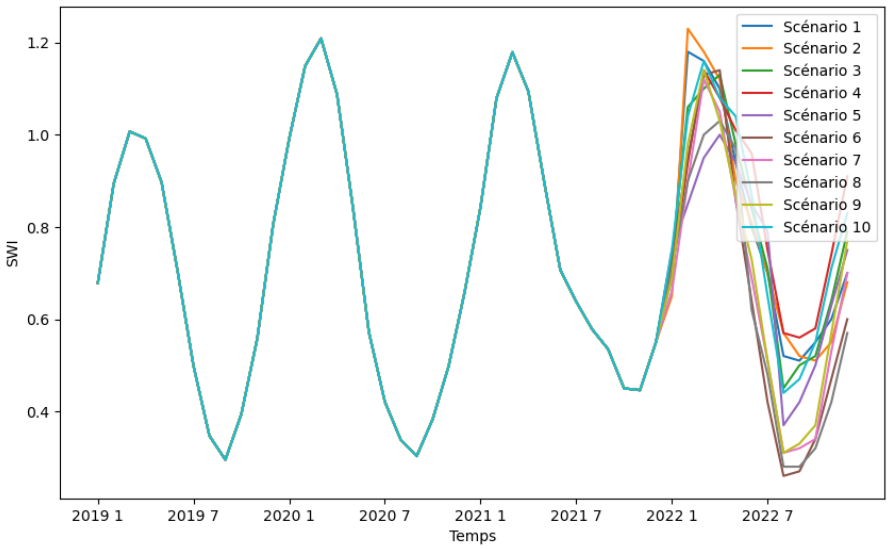


FIGURE II.1 – 10 projections pour l'année 2022, groupe de mailles 10.

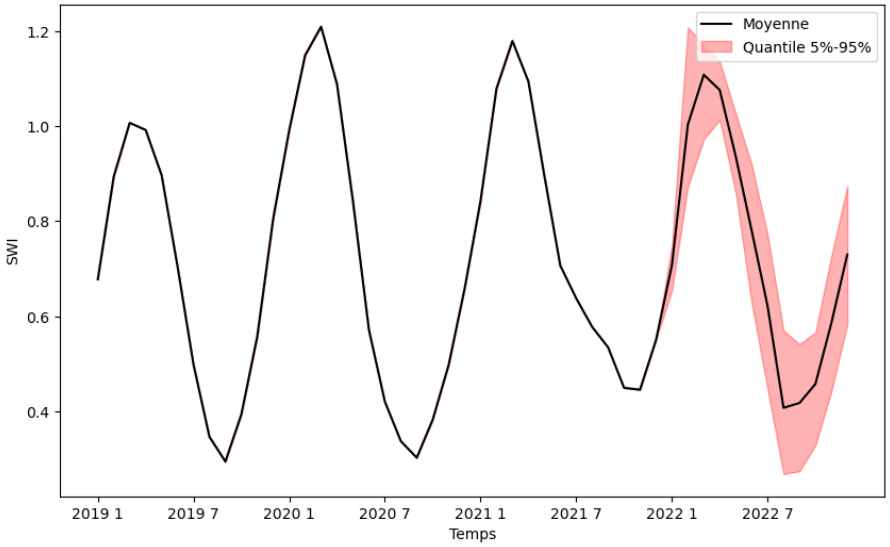


FIGURE II.2 – Moyenne et intervalle de confiance pour l'année 2022, groupe de mailles 10.

3.1 Projections SWI

Il est obtenu pour chaque entrée dans le réseau de neurones une projection de la forme :

Mois	Groupe Mailles					
	1	2	3	4	5	6
1	0.77	0.83	0.88	0.81	0.87	0.94
2	0.91	0.99	0.99	1.01	1.09	1.14
3	0.96	1.03	1.07	1.01	1.11	1.19
4	0.94	0.95	1.03	1.01	1.10	1.10
5	0.81	0.91	0.97	0.90	0.97	1.04
6	0.57	0.67	0.73	0.67	0.71	0.78
7	0.44	0.52	0.55	0.50	0.54	0.60
8	0.35	0.42	0.43	0.40	0.42	0.47
9	0.34	0.41	0.42	0.36	0.43	0.45
10	0.42	0.53	0.57	0.49	0.57	0.60
11	0.54	0.67	0.70	0.66	0.72	0.77
12	0.76	0.84	0.85	0.85	0.90	0.95

TABLE II.1 – Première projection des SWI à horizon t+1 groupes 1 à 6 (données perturbées).

D'abord, il est important de rappeler que les projections fournies sont des projections jointes, c'est-à-dire qu'elles prennent en compte l'ensemble des groupes de maille simultanément, plutôt que de considérer chaque groupe de mailles indépendamment.
De plus, il est à noter que les projections des indices SWI ont été ordonnées par ordre croissant d'humidité en fonction des groupes de mailles, ce qui confirme d'ailleurs que **l'ACP a bien regroupé les mailles selon leur niveau d'humidité**, de la moins humide à la plus humide.
De plus, le modèle démontre une bonne capacité à saisir la saisonnalité des données, se traduisant par des niveaux d'humidité des sols plus élevés en mars (fin de l'hiver) et plus faibles en septembre (fin de l'été). Ces périodes de l'année sont généralement associées à des conditions météorologiques favorables respectivement à une augmentation ou à une diminution de l'humidité des sols.

3.2 Probabilité de respecter le critère météorologique

3.2.1 Monte Carlo

Afin d'obtenir $\mathbb{P}(\text{Critère météorologique}(t, x))$, une méthode de type Monte Carlo détaillée ci-dessous est employée.

A partir d'une projection, en comparant les SWI estimés par le modèle au seuil préalablement défini, il est possible d'observer pour quels mois et pour quels groupes de mailles le critère météorologique est respecté :

Mois	Groupe Mailles					
	1	2	3	4	5	6
1	**True**	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	**True**	False	False	False
5	False	False	False	False	False	False
6	False	False	False	False	False	False
7	False	False	False	False	False	False
8	False	False	False	**True**	False	False
9	False	False	False	False	False	False
10	False	False	False	False	False	False
11	False	False	False	False	False	False
12	False	**True**	False	False	False	False

TABLE II.2 – Seuils du critère météorologique, horizon 1 an, groupes mailles 1 à 6 4, 1ère itération.

Pour connaître $\mathbb{P}(\text{Critère météorologique}(t, x))$, il suffit alors d'observer pour chaque sortie du réseau de neurones, si au moins un **True** se trouve dans la restriction du tableau aux mois appartenant à la saison t et aux groupes de mailles appartenant à la commune x.
Par exemple, pour la commune Saint-Pont en hiver 2022, composée des groupes de mailles 3 et 4, Il faut compter pour chaque simulation si au moins un **True** apparaît :

Mois	Groupe Mailles	
	3	4
1	**True**	False
2	False	**True**
3	False	False

TABLE II.3 – Seuils du critère météorologique, horizon hiver 2022, groupes mailles 3 et 4, 2ème itération.

Au final, $\mathbb{P}(\text{Critère météorologique}(t, x))$ est évalué comme :

$$\mathbb{P}(\text{Critère météorologique}(t, x)) = \frac{\text{Nombre de simulations comportant au moins un } \textbf{\textit{True}}}{\text{Nombre de simulations}}$$

En pratique, 1000 itérations ont été effectuées pour cette évaluation.

3.2.2 Critère actuel

Il est rappelé que, jusqu'à l'heure de la rédaction de ce mémoire, la loi Rousseau n'est pas entrée en application. Les probabilités de respecter le critère météorologique sont obtenues avec l'indicateur de durée de retour 25 ans. Les résultats obtenus avec cette durée de retour sont présentés ci-dessous.

Pour les 6 villes présentées en II.4, les probabilités de respecter le critère catastrophe naturelle semblent être comprises entre 2 % et 6 %. Les saisons du printemps et de l'automne semblent être légèrement moins propices à respecter ce critère météorologique.
Néanmoins, les probabilités de respecter le critère météorologique devraient être légèrement supérieurs. En effet, pour respecter le critère météorologique, il suffit que le SWI d'une seule maille de la commune dépasse la durée de retour de 25 ans pendant un seul mois de la saison. Les probabilités devraient donc être au moins supérieures à 4 %, plutôt autour de 10 %, sans compter l'impact du réchauffement climatique.

Une représentation globale des résultats en Auvergne-Rhône-Alpes est présentée en II.3.
Les probabilités globales sont comprises entre 0 % et 12 %. Contrairement à ce qui est observé ci-dessus, l'automne et le printemps ne semblent pas être interprétés comme moins probables de vérifier le critère météorologique que les autres saisons.
Cependant, il apparaît que l'été présente globalement de faibles probabilités.
De plus, la probabilité semble très inégalement répartie entre les territoires. Globalement, comme le montre le graphique du printemps 2022 ci-dessus, il est visible que les départements classés comme humides d'après le *clustering* 1.2 sont aussi ceux qui sont plus susceptibles de respecter le critère météorologique d'après le modèle. Par exemple, les départements de la Savoie et la Haute-Savoie ont des probabilités beaucoup plus fortes en hiver, automne et printemps de respecter le critère météorologique. Intuitivement, il est surprenant d'observer des régions "humides" comme plus susceptibles que des régions plus "sèches" de respecter le critère météorologique.

Cependant, plusieurs explicitations à ce phénomène sont suggérées :

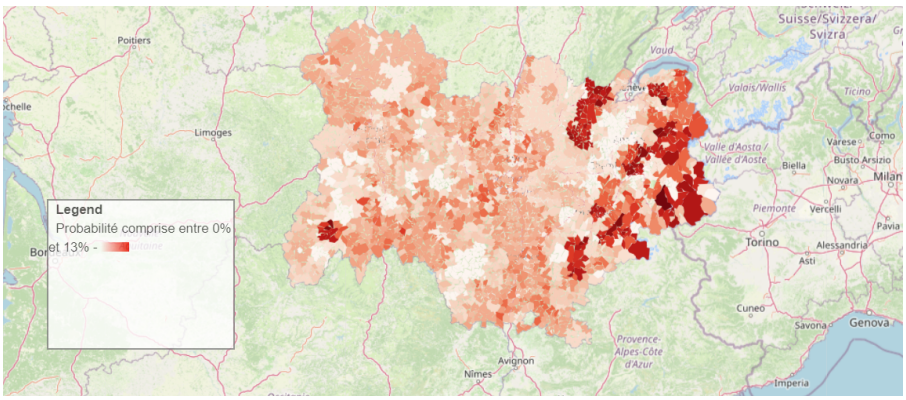
— les zones plus humides connaîtront des changements plus radicaux que les zones déjà plus sèches actuellement ;

— les critères pour les zones plus sèches sont trop bas à cause des sécheresses qui ont récemment eu lieu, et le modèle ne parvient pas à perturber suffisamment les SWI sur ces zones 3.4.

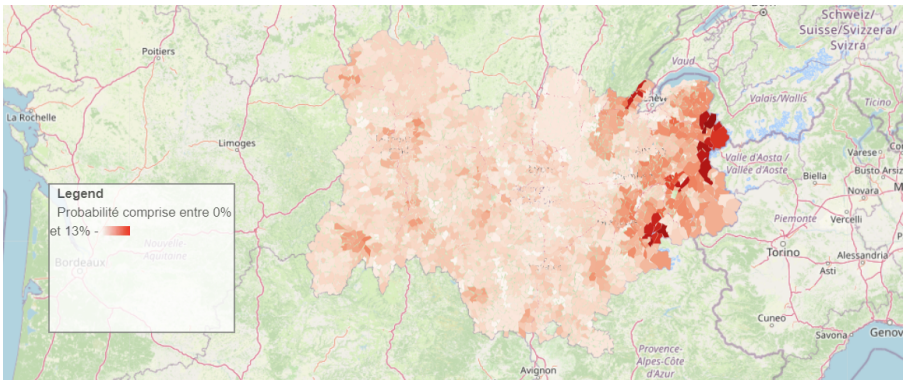
En pratique, il est possible que ces deux facteurs interviennent simultanément. Il est donc possible dans une certaine mesure de faire confiance à ces résultats, en particulier pour les zones encore relativement humides.

ville	groupes_mailles	proba hiver	proba printemps	proba été	proba automne
Lyon	1, 2, 3	0,054	0,027	0,057	0,019
Neschers	0, 1	0,026	0,025	0,053	0,019
Manglieu	0, 1, 5	0,043	0,024	0,058	0,016
Busséol	0, 1	0,026	0,025	0,053	0,019
Crest	0, 1	0,026	0,025	0,053	0,019
Saint-Jodard	1, 4	0,049	0,019	0,054	0,015

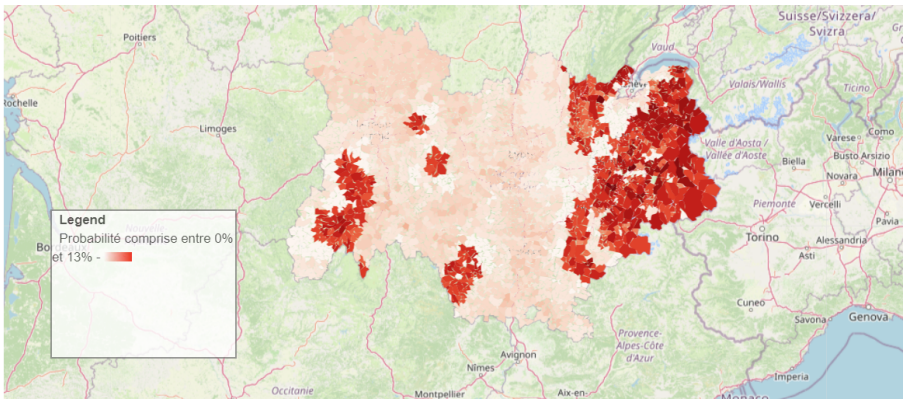
TABLE II.4 – Probabilité critère météorologique saison t+1 pour 6 villes.



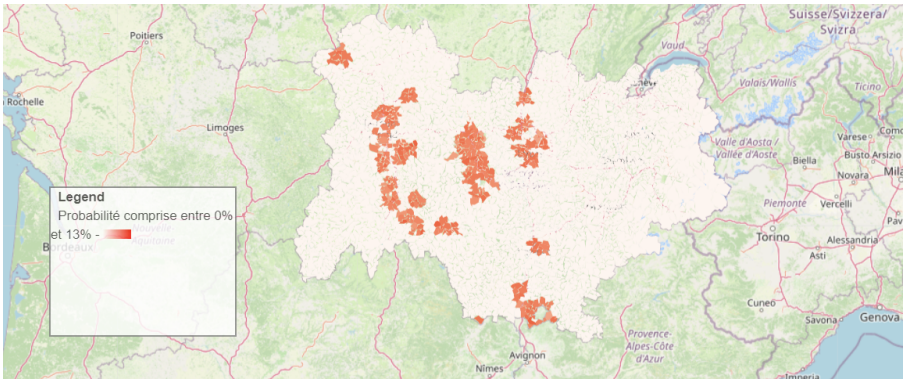
(a) Hiver 2022



(b) Printemps 2022



(c) Automne 2022



(d) Été 2022

FIGURE II.3 – Probabilités de respecter le critère météorologique en 2022 en Auvergne-Rhône-Alpes.

3.2.3 Critère Loi Rousseau

Dans un deuxième temps, les probabilités estimées par le modèle sont observées dans le cas où la loi Rousseau du 6 avril 2023 est effectivement promulguée. Elles sont donc estimées en utilisant l'indicateur de durée de retour 10 ans.

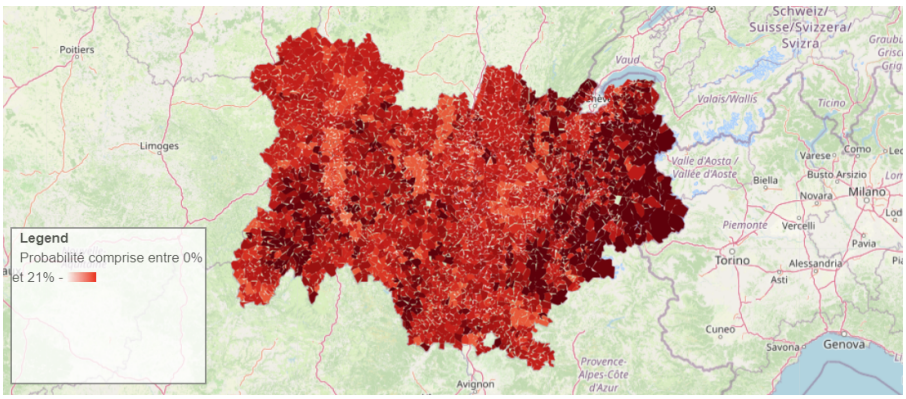
Les probabilités observées dans ce contexte sur les mêmes 6 villes que précédemment semblent être plus élevées II.5. En effet, des probabilités comprises entre 7 % et 12 % sont ici proposées. Aucune saison ne semble être significativement plus probable que les autres.

Ces hypothèses sont confirmées à l'échelle globale sur le graphique II.4.
Une hausse très significative de ces probabilités est observée avec des valeurs comprises entre 0 et 21 %.
Selon le modèle, la loi Rousseau permettrait de respecter les critères de déclaration catastrophe naturelle environ 11 % du temps, c'est-à-dire environ trois fois plus que précédemment.
Cependant, en raisonnant de la même manière que précédemment, dans ce cadre, pour respecter le critère météorologique, il suffit que le SWI d'une seule maille de la commune dépasse la durée de retour de 10 ans pendant un seul mois de la saison. Les probabilités devraient donc être au moins supérieures à 10 %.

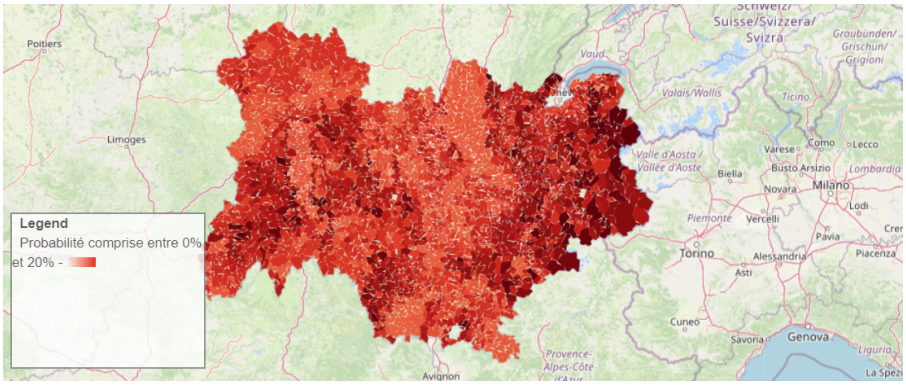
Au final, les résultats proposés par le modèle semblent légèrement sous-estimer les probabilités réelles.
Il est cependant intéressant de constater que la Loi Rousseau devrait tripler la probabilité de respecter le critère météorologique et donc grandement faciliter la déclaration de catastrophe naturelle dans les communes touchées par des sécheresses.

ville	groupes_mailles	proba hiver	proba printemps	proba été	proba automne
Lyon	1, 2, 3	0,095	0,089	0,115	0,087
Neschers	0, 1	0,071	0,075	0,073	0,097
Manglieu	0, 1, 5	0,119	0,106	0,099	0,105
Busséol	0, 1	0,071	0,075	0,073	0,097
Crest	0, 1	0,071	0,075	0,073	0,097
Saint-Jodard	1, 4	0,111	0,092	0,100	0,087

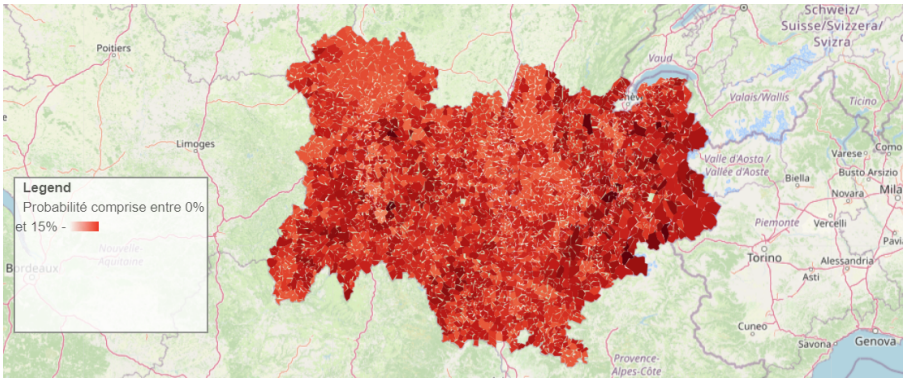
TABLE II.5 – Probabilité critère météorologique saison t+1.



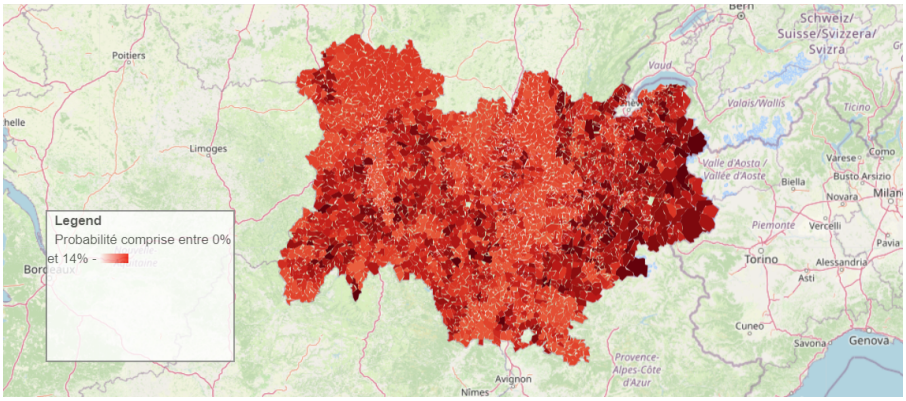
(a) Hiver 2022



(b) Printemps 2022



(c) Automne 2022



(d) Été 2022

FIGURE II.4 – Probabilités de respecter le critère météorologique en 2022, dans le contexte de la loi Rousseau en Auvergne-Rhône-Alpes.

3.3 Critère météorologique et géologique

Au final, pour $\mathbb{P}(\text{Cat Nat}(t,x))$, il suffit de multiplier les probabilités ci-dessus par 0 si le critère géologique n'est pas respecté, et par 1 s'il l'est à partir des données présentées en I.5.

ville	critère géologique	proba hiver	proba printemps	proba été	proba automne
Lyon	Oui	0,054	0,027	0,057	0,019
Neschers	Oui	0,026	0,025	0,053	0,019
Manglieu	Oui	0,043	0,024	0,058	0,016
Busséol	Oui	0,026	0,025	0,053	0,019
Crest	Oui	0,026	0,025	0,053	0,019
Saint-Jodard	Oui	0,049	0,019	0,054	0,015
Cournols	Non	0,0	0,0	0,0	0,0

TABLE II.6 – Critère géologique et probabilité respect critère météorologique saison t+1.

Voici les résultats finaux, qui correspondent à la probabilité de respect des deux critères météorologique et donc à la probabilité de déclaration de catastrophe naturelle sécheresse, représentés sur une carte (durée de retour de 25 ans) pour la saison du printemps 2022 :

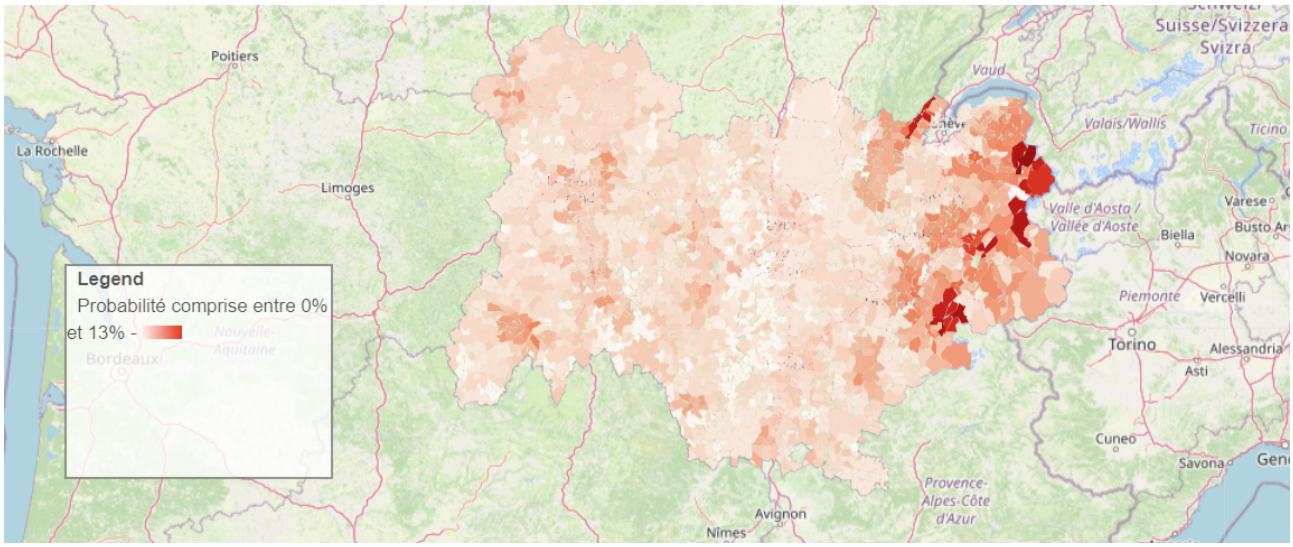


FIGURE II.5 – Probabilités de respecter les critères météorologique et géologique au printemps 2022 en Auvergne-Rhône-Alpes.

Sans surprise, les résultats sont semblables à ceux observés dans la section précédente II.3. Cependant, il est possible de distinguer une "zone blanche" correspondant aux communes qui ne respectent pas le critère géologique de la partie I.5. Pour résumer, ce résultat final est la superposition des figures II.3 et I.5.

3.4 Limites de la modélisation du critère météorologique

Dans cette étude, un modèle de prédiction des probabilités de dépassement du critère météorologique SWI (*Soil Wetness Index*) à un horizon temporel choisi à été développé. Cependant, malgré les résultats prometteurs obtenus, certaines limites dans cette modélisation ont été identifiées.
En analysant les résultats, il a été constaté que les probabilités prédites semblent être légèrement sous-estimées. Par exemple, une durée de retour de 25 ans devrait normalement correspondre, sans à priori, à une probabilité de 4 % par maille. De plus, lors de l'agrégation au niveau d'une commune, des probabilités légèrement supérieures étaient attendues, autour de 4,5 % voire 5%. Cependant, les probabilités obtenues sont plutôt centrées autour de 3 % en moyenne.
Une hypothèse explicative pourrait être que les seuils définis pour certaines années, notamment ceux de l'année 2020, étaient anormalement bas. En conséquence, la probabilité de dépasser ces seuils serait faible. Il est possible que le modèle n'ait pas suffisamment "appris" le réchauffement climatique observé au cours des dernières années, en particulier depuis 2018.
Dans ce sens, il est constaté que, ces dernières années, le critère météorologique est très souvent dépassé, comme le montre le graphique ci-dessous :

Saison	Année	Taux validation du critère constaté
Printemps	2018	0,68 %
Printemps	2019	43,14 %
Printemps	2020	56,58 %
Printemps	2021	6,39 %
	Moyenne Printemps	26,69 %
Été	2018	73,14 %
Été	2019	55,27 %
Été	2020	58,65 %
Été	2021	0,00 %
	Moyenne Été	46,74 %
Automne	2018	82,38 %
Automne	2019	13,19 %
Automne	2020	0,35 %
Automne	2021	0,00 %
	Moyenne Automne	23,9 %
Hiver	2018	12,38 %
Hiver	2019	39,19 %
Hiver	2020	5,03 %
Hiver	2021	0,76 %
	Moyenne Hiver	14,34 %
	Moyenne Globale	27,94 %

TABLE II.7 – Taux de validation du critère constaté par saison et année.

Bien que le réchauffement climatique joue un rôle indéniable, il reste donc difficile de dire si ces années ont été exceptionnelles en termes de sécheresse, ou s'il y a une tendance clairement à la baisse de l'humidité des sols.
Il faudra donc être prudent quant à l'utilisation des probabilités de survenance du critère météorologique de cette partie.
Certaines des conclusions "qualitatives" telles que l'impact de la Loi Rousseau restent cependant valables.

Il est important de souligner que prédire le comportement d'un indicateur complexe tel que le critère météorologique reste un défi. Idéalement, il faudrait disposer d'un modèle météorologique intégrant des projections probabilistes de SWI.
Néanmoins, en l'absence d'un tel modèle météorologique, cette approche reste une option qui permet de prédire les probabilités de respecter le critère météorologique.

Chapitre 4

Projections SWI moyens

Les projections de SWI produites par le LSTM peuvent néanmoins êtres utilisées dans la prévision des SWI moyens par commune conditionnellement au fait que les critères de déclaration de catastrophe naturelle sécheresse soient respectés à horizon un an. Ceci est une information majeure qui sera très utile dans la partie suivante : III.
En effet pour chaque commune, pour chaque mois de l'année, si au moins une de ses mailles dépasse l'indicateur, alors la moyenne du SWI des mailles associées à la commune est calculée. En effectuant la moyenne sur 10 000 projections, voici les résultats obtenus pour la ville de Lyon :

ville	groupes_mailles	saison	SWI moyen conditionnel
Lyon	1, 2, 3	hiver	0,86
Lyon	1, 2, 3	printemps	0,72
Lyon	1, 2, 3	été	0,32
Lyon	1, 2, 3	automne	0,39

TABLE II.1 – Projections SWI moyens à Lyon conditionnels au respect des critères de déclaration catastrophe naturelle à horizon un an.

Cette approche est géréralisée à l'ensemble des communes d'Auvergne-Rhône-Alpes et les données de SWI conditionnelles sont donc utilisables dans le modèle proposé dans la partie suivante sur la modélisation du coût conditionnel.

Chapitre 5

Conclusion

Dans la partie précédente, une méthode a été proposée afin de modéliser les SWI des mailles d'Auvergne-Rhône-Alpes.
Cette modélisation a d'abord nécessité une analyse en composantes afin de réduire la dimension des donnéees. Par la suite, l'emploi d'un LSTM a permis de projeter ces SWI. Les différentes valeurs de RMSE ont enfin permis de valider ce modèle.

Cependant, la modélisation du critère météorologique s'est avérée être compliquée. En effet, la modélisation des SWI n'a pas permis d'aboutir à une modélisation satisfaisante pour l'estimation de la probabilité de respecter le critère météorologique. Cela est certainement dû à plusieurs facteurs : de nombreuses hypothèses réalisées, le réchauffement climatique et son impact sur les SWI trop récent, complexité des données à modéliser en raison du phénomène sous-jacent physique compliqué.

Toutefois, le modèle peut déjà amener certaines conclusions quant au critère météorologique.

- Les seuils du critère météorologique ont déjà atteint des valeurs très basses en été pour la plupart des communes d'Auvergne-Rhône-Alpes.
- Les saisons d'automne, d'hiver et du printemps disposent encore d'une certaine marge concernant l'atteinte de valeurs extrêmement basses. Ceci se traduit par des probabilités de déclenchement du critère météorologique supérieures à celles de l'été. Il serait alors possible à l'avenir que le critère météorologique soit davantage atteint lors de ces saisons et que l'on assiste à davantage de sinistres notamment au printemps.
- Les communes de certaines régions encore relativement humides possèdent également une certaine marge : par exemple en Savoie et en Haute-Savoie. Ces communes n'ayant pas encore atteint de valeurs très basses, elles pourraient, à l'avenir, les dépasser plus facilement et donc connaître une augmentation plus forte du nombre de sécheresses relativement aux autres communes d'Auvergne-Rhône-Alpes.

Enfin, cette partie est également importante dans la modélisation du risque observé par les assureurs puisqu'elle a permis de déduire les SWI futurs pour chaque groupe de mailles. En effet, l'information qu'un sinistre ne peut survenir qu'au déclenchement d'une catastrophe naturelle, et donc au respect du critère météorologique requis, indique que **le SWI sera nécessairement l'une des deux valeurs les plus basses jamais enregistrées**. Cette information cruciale sera intégrée à la modélisation du coût de la prochaine partie.

Troisième partie

Coût conditionnel d'un sinistre sécheresse

Bien que l'importance du cadre réglementaire de la garantie Cat Nat sécheresse ait été montrée dans les deux parties ci-dessus, il demeure néanmoins nécessaire de comprendre le risque réellement observé chez les assureurs.

Ainsi, dans cette partie, il sera désormais question d'étudier la sinistralité sécheresse.
En conditionnant par rapport au fait que les critères de déclaration de catastrophe naturelle sécheresse soient respectés, **de l'information supplémentaire** est apportée : le SWI d'une des mailles de la commune dans laquelle se situe **le contrat a dépassé le seuil défini par le critère météorologique** pour un mois de la saison (ce sont les SWI moyens qui sont estimés en 4) et le contrat se trouve dans une **commune respectant le critère géologique**.

Des variables endogènes aux contrats sont également étudiées dans l'explication du coût d'un sinistre sécheresse.
En effet, il semblerait comme le rapport du Mission Risques Naturels (MRN) de 2018 le montre que certaines variables jouent un rôle important dans la sinistralité. Dans le cas d'un sinistre de type fissure sur le bâti, les principales composantes de la charge du sinistre sont :

— fondations/ouvrage enterré/sous-œuvre (52 % de la charge en moyenne) ;

— façade/véranda/revêtement extérieur (32 % de la charge en moyenne) ;

— embellissement/finition (13 % de la charge en moyenne).

L'étude menée dans cette partie est réalisée à partir d'une base de données sinistres de notre client.

Chapitre 1

Variables étudiées

Plusieurs variables peuvent influencer le coût d'un sinistre sécheresse.
Tout d'abord, il semblerait que les coûts soient relativement différents selon les départements.

Le rapport du MRN propose plusieurs explications :

- conditions économiques du BTP plus pénalisantes dans certains départements ;
- conditions géotechniques plus favorables à la réalisation de micropieux dans certains départements ;
- pratique régionale de reprises en sous-œuvre par micropieux sans mise en œuvre d'une longrine de répartition des charges en confortement des fondations défaillantes, plus économique mais susceptible d'accroître à terme le risque de désordres de 2ème génération.

Enfin, selon le mémoire présenté en pièce [5], les sinistres touchent beaucoup plus les maisons que les appartements (99 % de maisons pour 1 % d'appartements), et cela est certainement lié aux fondations qui sont plus profondes pour les immeubles.
C'est également ce qui est constaté sur la base de données sinistres utilisée dans ce mémoire.
Cependant, un sinistre qui se produit sur un immeuble coûte plus cher lors des réparations.
De plus, il semblerait que connaître la végétation aux alentours, les conditions locales de l'argile dans le sol et les éventuels moyens de prévention mis en place seraient utiles à la modélisation du coût.
Néanmoins, les données à la disposition des assureurs ne sont pas toujours aussi précises. Dans l'idéal, récolter autant d'informations que possible serait bénéfique mais poser trop de questions aux clients pourrait en dissuader certains ou même prendre trop de temps aux services de souscription.
Dans l'attente de ces précisions, l'étude sera réduite aux variables suivantes : SWI moyen (conditionnel à la déclaration de catastrophe naturelle), Score Argile, Nombre de pièces, Nature de l'habitation.
La variable Score Argile sera définie dans la partie suivante (2.2).

Statistiques descriptives

Avant toute modélisation, il demeure essentiel de mieux comprendre la base de données.

Variable à expliquer

Il faut d'abord remarquer que le coût moyen d'un sinistre (sinistres à zéros exclus) est de 20 870 €.
De plus, le coût conditionnel au déclenchement de catastrophe naturelle est totalement déséquilibré, présentant une forte concentration autour de 0 comme le montre le tableau III.2.
Il est donc à souligner que le modèle de classification utilisé devra chercher à prédire une probabilité très faible, de l'ordre de 10^{-4}.

Par ailleurs, il est possible d'observer la charge sinistre en fonction des variables Score Argile, Nombre de pièces et SWI moyen sur III.1.
La charge sinistre en fonction de la nature de l'habitation est également présentée en III.1.

Conjecture

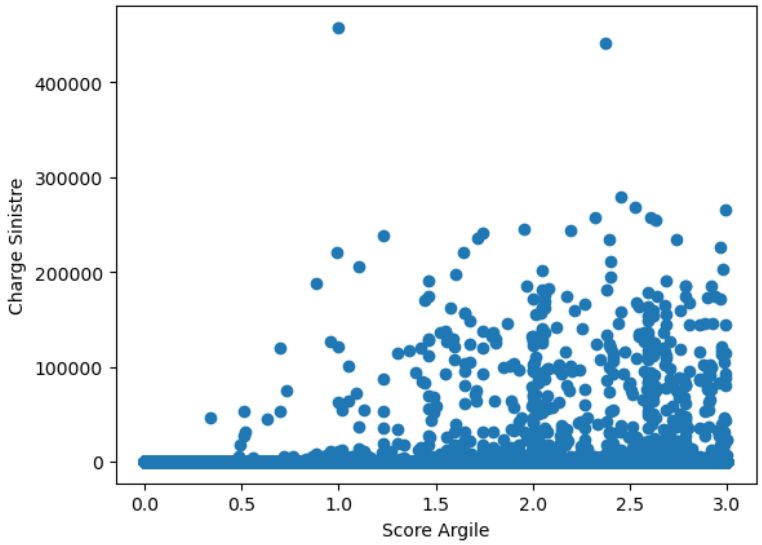
Il semble que les variables Nature Habitation, SWI moyen et Score Argile soient significatives.
La variable Nombre de pièces semble elle aussi, dans une moindre mesure, être significative.

Nature Habitation	**Charge Constatée (€)**
Appartement	0,59
Maison	33,7

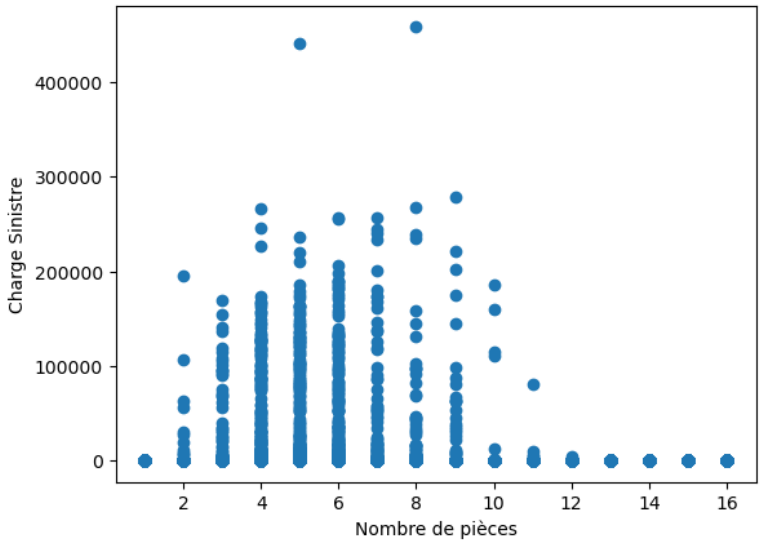
TABLE III.1 – Charge sinistre en fonction de Nature de l'habitation.

Réalisations de la variable $C(t,x)$|Cat Nat	Proportion dans la base de données en %
$C(t,x) = 0$|Cat Nat	99,9
$C(t,x) > 0$|Cat Nat	0,09

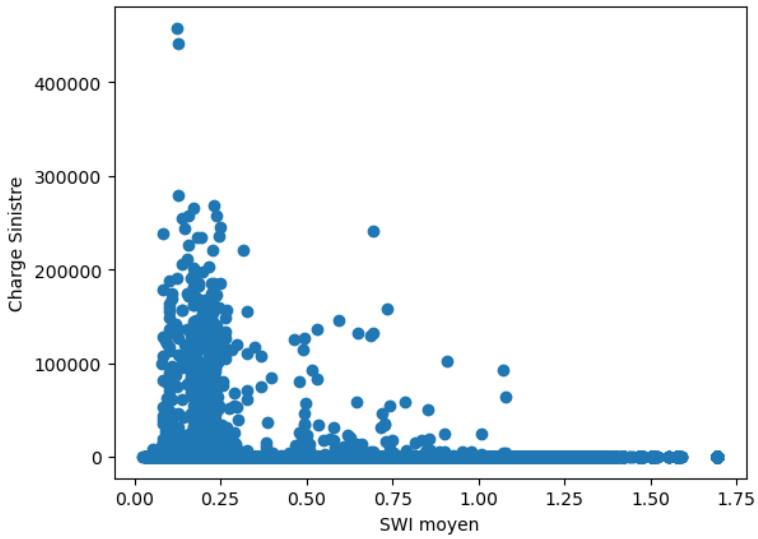
TABLE III.2 – Description des coûts conditionnels.



(a) Charge sinistre en fonction de Score Argile.



(b) Charge sinistre en fonction de Nombre de pièces.



(c) Charge sinistre en fonction de SWI moyen.

FIGURE III.1 – Charge sinistre en fonction des variables quantitatives.

Variables explicatives

Par la suite, la répartition des variables explicatives est observée, c'est-à-dire la répartition des variables SWI futur s'il y a état de catastrophe naturelle sécheresse, Nombre de pièces, Nature de l'habitation. La répartition des variables est présentée en III.2 et en III.3.

La répartition de la variable qualitative nature habitation est équilibrée.
La variable Nombre de pièces est répartie entre 1 et 16 pièces. La fréquence d'occurrence est croissante de 1 à 5 pièces, qui est la modalité la plus représentée ; puis décroissante de 5 à 16 pièces. Cette variable sera utilisée comme une variable quantitative.
Le SWI futur est lui distribué de manière relativement homogène entre 0 et 1,75.
Enfin, le Score Argile est distribué de manière homogène entre 0 et 3. Néanmoins, la distribution est plus dense autour de 1.

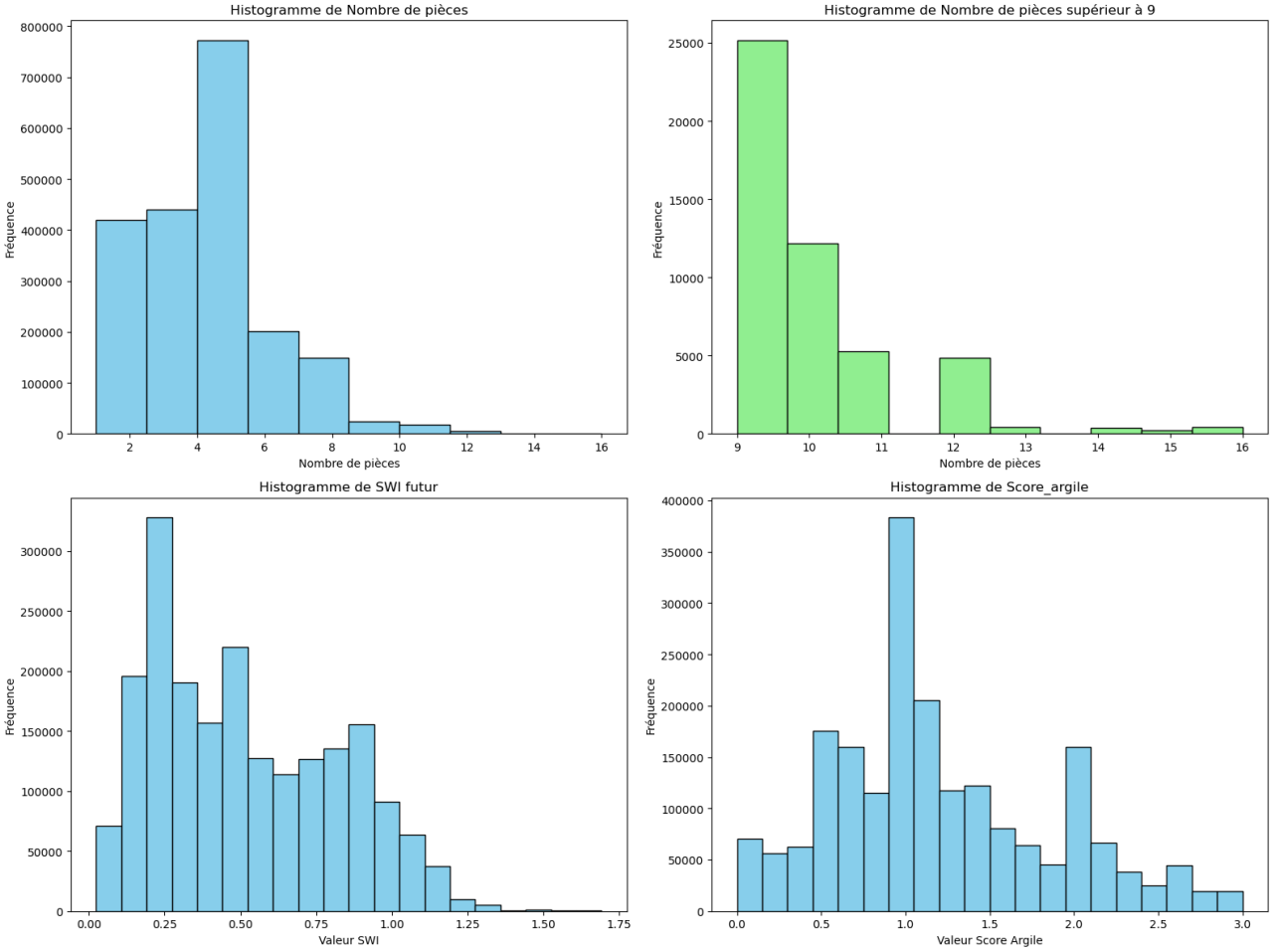


FIGURE III.2 – Histogrammes des variables quantitatives Nombre de pièces, SWI futur et Score Argile.

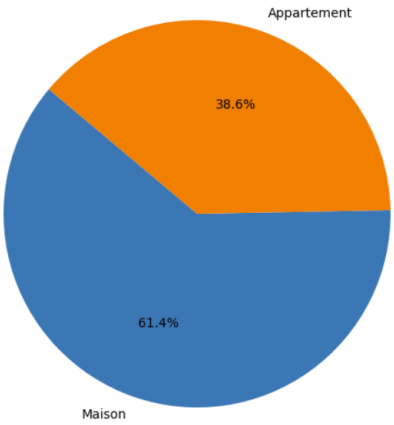


FIGURE III.3 – Répartition des modalités de la variable Nature habitation.

Ces variables explicatives, ne présentant pas d'anomalie statistique, pourront donc être utilisées dans le modèle proposé ci-dessous.

Par la suite, la dépendance entre les différentes variables est observée.

Comme le montre la matrice de corrélation III.4, les variables explicatives SWI futur et Score Argile sont légèrement corrélées (-18 %) tandis que les autres combinaisons de variables ne sont pas corrélées entre elles. En intégrant ces variables-ci dans le modèle qui sera proposé, les problèmes de colinéarité seront évités.

Il est ici admis que la nature de l'habitation n'est pas liée aux autres variables.

Enfin, les variables semblent effectivement bien expliquer le phénomène de survenance de sinistre comme le montrent les p-value des tests ANOVA (analyse de la variance) en III.3 (test de dépendance entre une variable qualitative et une variable quantitative) et la p-value de test du Chi-2 en III.4 (test de dépendance entre deux variables qualitatives).

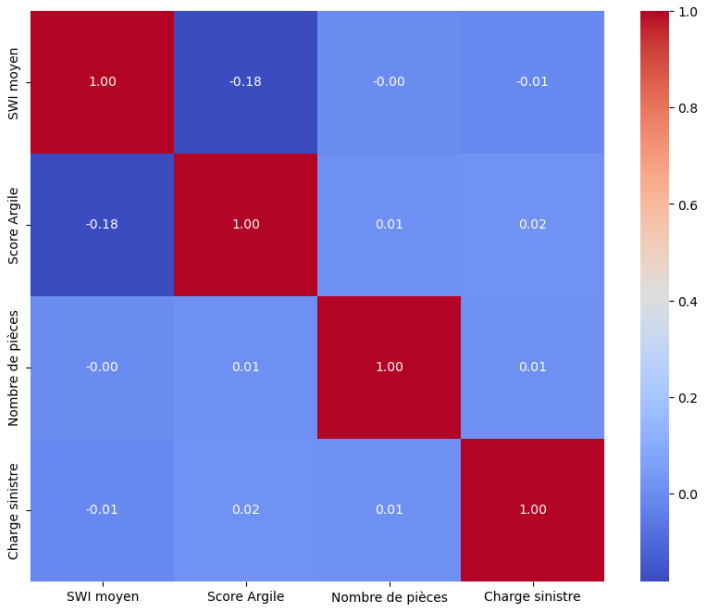


FIGURE III.4 – Matrice de corrélation des variables quantitatives.

Variables explicatives	Survenance d'un sinistre
SWI futur	2.94×10^{-203}
Score Argile	0.0
Nombre de pièces	1.11×10^{-129}

TABLE III.3 – p-value tests ANOVA.

	Survenance d'un sinistre
Nature Habitation	2.11×10^{-266}

TABLE III.4 – p-value test Chi-2.

Les variables ne sont donc pas liées entre elles et sont suffisamment explicatives du phénomène pour être utilisées par la suite.

En réalité, seulement quelques variables sont ici prises en compte : SWI moyen, Score Argile, Nombre de pièces et Nature habitation. La période de construction pourrait jouer un rôle mais trop peu d'observations sont proposées dans la base de données de cette étude. Cette connaissance pourrait néanmoins renseigner sur la profondeur des fondations et la solidité globale du bâti.

Chapitre 2

Zonier d'exposition au phénomène de retrait gonflement des argiles

2.1 Motivations

Initialement, les variables explicatives fournies par la base de données sont Nombre de pièces, nature Habitation et le SWI estimé en deuxième partie pour la commune. Cependant, il paraît cohérent de tenir compte de la teneur du sol dans la modélisation de la sinistralité. Il est donc proposé de construire une variable "endogène" appelée Score Argile, reflétant l'exposition au phénomène de retrait gonflement des argiles dans cette partie.

2.2 Construction de la variable Score Argile

D'après la carte BRGM d'exposition au phénomène RGA I.5, trois niveaux d'argile sont référencés dans les sols français : faible, moyen et fort. Cette carte permet d'abord de définir le critère géologique comme vu en partie I.5. L'enjeu ici est d'agréger ces données à l'échelle de la commune. En effet, une commune peut à la fois contenir des zones sans argile et à faible, moyenne et forte exposition.

La première hypothèse réalisée afin d'obtenir cette vision par commune est l'hypothèse pessimiste : la commune se voit toujours attribuée le niveau d'argile le plus fort qui la compose. Par exemple, une commune composée à 5 % d'une zone fortement exposée, 50 % d'une zone faiblement exposée et 45 % d'une zone sans argile est classée comme une commune fortement argileuse.
Cette approche, présentée en III.1, permet de constater que les communes classées comme ayant une forte exposition présentent en moyenne des coûts plus élevés que les zones à faible exposition. Toutefois, elle comporte un défaut majeur : des zones normalement faiblement argileuses comme la Savoie ou la Haute-Savoie sont représentées comme moyennement exposées.
Cette carte ne semble donc pas bien représenter la sinistralité réelle du client, qui est très localisée dans le Puy de Dôme, l'Allier, et l'agglomération Lyonnaise.

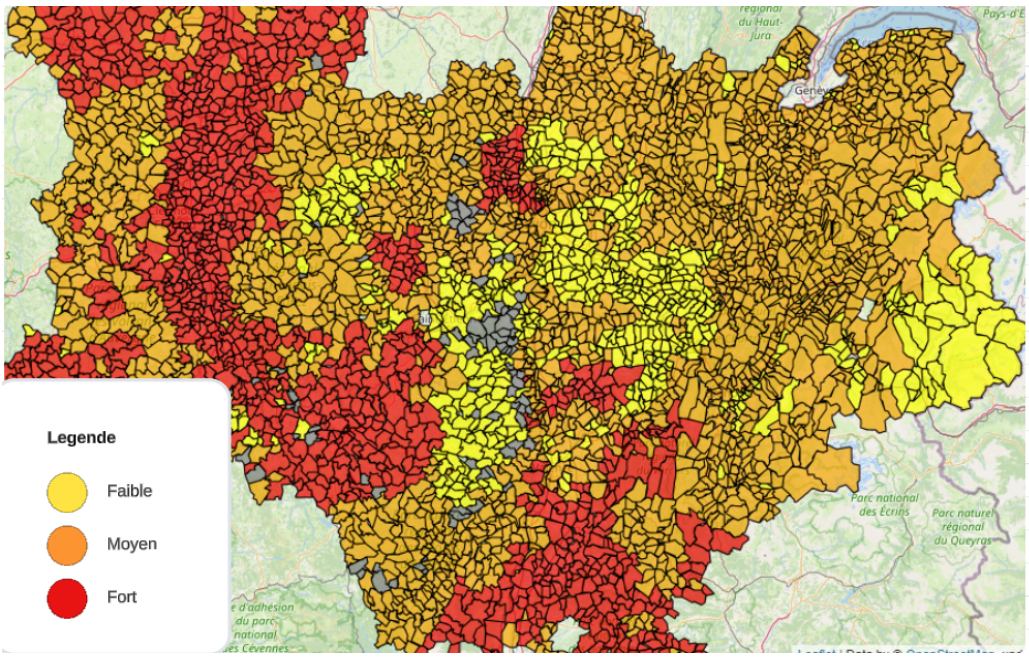


FIGURE III.1 – Modèle naïf d'exposition RGA agrégée par commune, hypothèse "pessimiste".

Suite aux manquements du modèle naïf présenté ci-dessus, une autre approche est proposée. En effet, un score d'exposition au phénomène RGA pour chaque commune est défini de la manière suivante :

si x % de la commune est exposée faiblement au phénomène selon la carte BRGM,
si y % de la commune est exposée moyennement au phénomène selon la carte BRGM,
et si z % de la commune est exposée fortement au phénomène selon la carte BRGM,
alors la commune se voit attribuée un score de $x\% \times 1 + y\% \times 2 + z\% \times 3$.

Les résultats obtenus sont présentés en III.2. Une couleur est attribuée à chaque commune suivant une échelle graduée allant du blanc (score d'argile = 1) au rouge foncé (score d'argile = 3). Ainsi, les communes représentées en rouge foncé sont fortement exposées au risque de retrait-gonflement des argiles tandis que les communes représentées en blanc sont faiblement exposées au risque de retrait-gonflement des argiles.
Cette carte semble plus pertinente au regard des remarques faites précédemment : en effet, les régions montagneuses semblent être moins exposées que certaines autres régions. Tandis que les régions de l'Allier, du Puy-de-Dôme et de la Haute-Loire (zones à forte sinistralité dans le portefeuille client) sont représentées comme fortement à risque.
Une comparaison est d'ailleurs proposée entre ce zonier d'exposition au RGA et le nombre de mois passés sous état de catastrophe naturelle sécheresse par commune depuis 1969, représenté en III.3. Ici, les communes représentées sont celles qui ont passé au moins un mois sous état de catastrophe naturelle sécheresse. Et plus elles ont passés de mois sous état de catastrophe naturelle sécheresse, plus leur couleur tend du blanc au rouge foncé.
Une forte similarité entre les deux cartes est remarquée : les zones identifiées comme à risque selon le zonier ont également tendance à être des zones ayant déclaré beaucoup de catastrophes naturelles sécheresse.

Cette comparaison confirme alors l'utilisation de ce zonier dans le modèle.

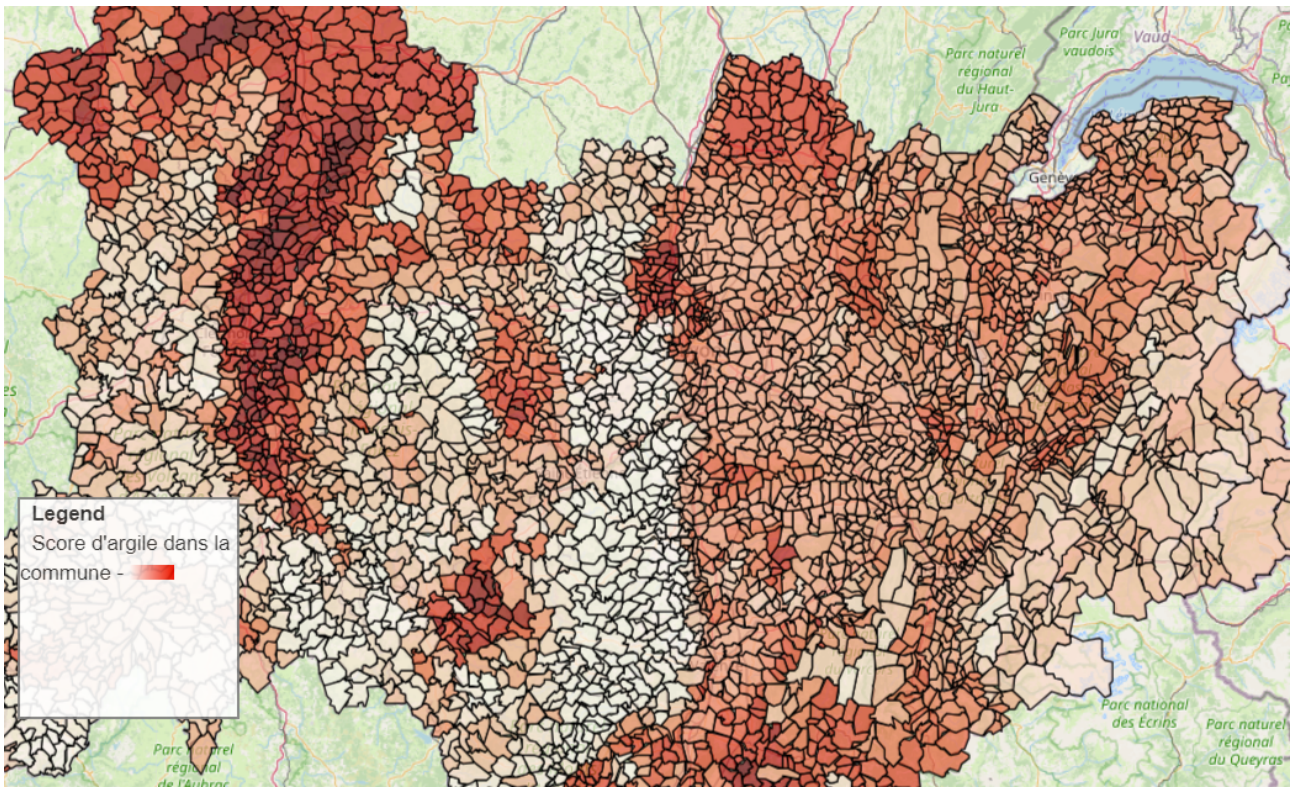


FIGURE III.2 – Zonier exposition au phénomène RGA en Auvergne-Rhône-Alpes.

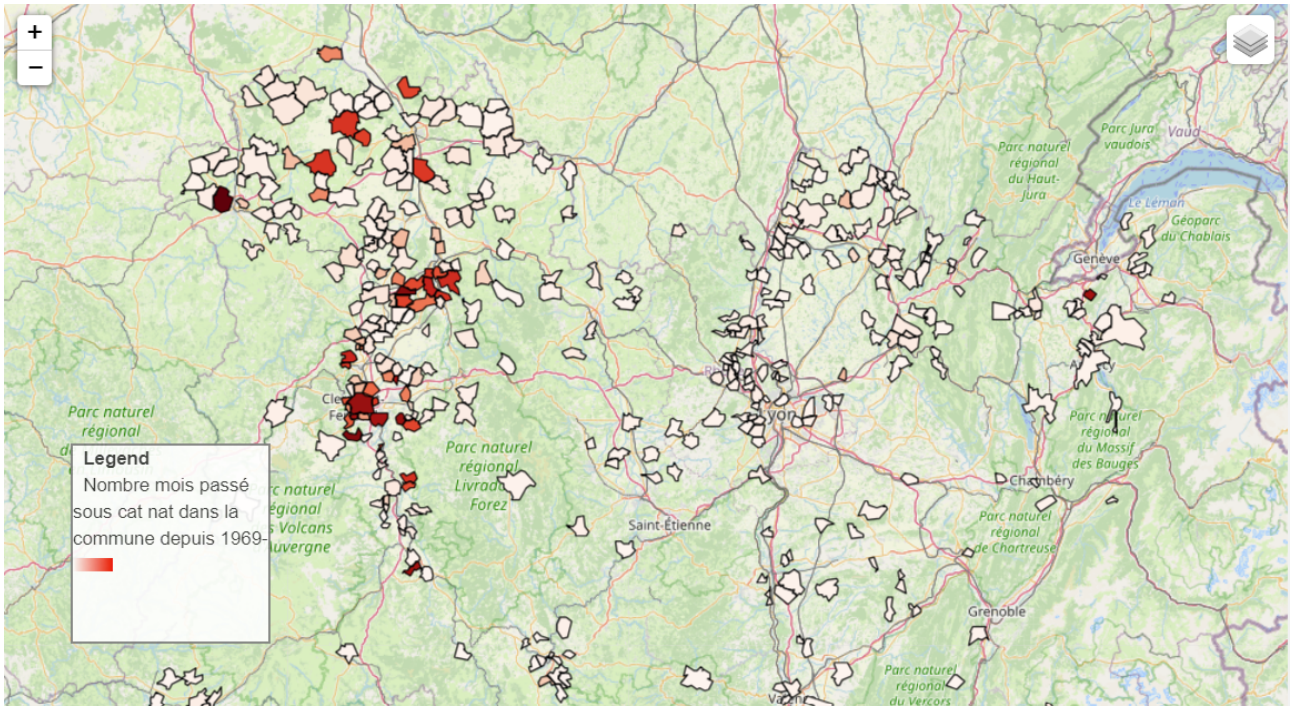


FIGURE III.3 – Nombre de mois passés sous état de catastrophe naturelle sécheresse sur la période 1969-2021 en Auvergne-Rhône-Alpes.

Chapitre 3

Estimation d'un sinistre sécheresse

Afin d'estimer le coût d'un sinistre sécheresse, une approche coût-fréquence est proposée comme en 3.2.

Il faut rappeler ici que le sinistre est estimé conditionnellement au respect du déclenchement de catastrophe naturelle. Il est supposé que cette information est connue (ou estimée) et permet à l'assureur de bénéficier de l'information supplémentaire apportée sur le SWI (dépassement d'un certain seuil).
Les variables propres au contrat (Nature de l'habitation, Nombre de pièces, et le Score Argile de la section précédente) sont également utilisées pour la prédiction.

La fréquence d'apparition de sinistre $\mathbb{P}(C(t,x)=0)|\text{CatNat}, X)$ est étudiée en 3.1.

La modélisation du coût de ces sinistres est également étudiée en 3.2 :
$\mathbb{E}[C(t,x)|\text{CatNat}, X, C(t,x)>0]$.

3.1 Probabilité de survenance d'un sinistre

3.1.1 Séparation de la base de données

La base de données est séparée en une base d'entraînement (80 % de la base complète) et une base de test (20 % de la base complète).
Les statistiques descriptives des variables à expliquer et explicatives sont semblables d'une base à l'autre, ce qui permet d'entraîner les modèles sur la base d'entraînement et de les valider en comparant les prédictions obtenues pour la base test.

3.1.2 Estimations de la probabilité

Dans le cadre de données déséquilibrées comme celles-ci III.2, l'utilisation de modèles linéaires classiques est caduque. Il est conseillé de se diriger vers des modèles de Machine Learning. XGBoost est connu pour être particulièrement efficace dans la gestion de données déséquilibrées grâce à sa capacité à gérer les poids des différentes observations. En utilisant des poids d'observations adaptés, XGBoost peut attribuer une plus grande importance aux observations de la classe minoritaire et ainsi améliorer sa performance sur cette classe. Ainsi, des modèles XGBoost sont utilisés afin d'estimer la probabilité de survenance d'un sinistre.
Précisément, $\mathbb{P}(C(t,x)=0|\text{CatNat}, X)$ représente la probabilité qu'aucun sinistre n'apparaisse dans la commune x, pendant la saison t (exemple : hiver 2022) sachant que les conditions de

catastrophe naturelle sont respectées dans la commune sur un contrat dont sont connues les variables :

— SWI futur s'il y a état de catastrophe naturelle sécheresse (estimation de 4) ;
— Score Argile ;
— Nombre de pièces ;
— Nature habitation.

En agrégeant sur chaque contrat et sur chaque saison, le coût total des sinistres constatés, le SWI moyen constaté pendant la saison, la nature de l'habitation du contrat, le score d'argile donné par le zonier ci-dessus dans la commune du contrat, et le nombre de pièces du contrat, le modèle XGBoost est entraîné.
Afin de simplifier les notations, la probabilité précédente sera notée $\mathbb{P}(Y = 0)$. Le modèle prédira $\mathbb{P}(Y > 0)$ et $\mathbb{P}(Y = 0) = 1$ - $\mathbb{P}(Y > 0)$.

Ainsi, dans cette partie, il sera considéré qu'il y a un sinistre si $\mathbb{P}(Y > 0) > 0.5$ et qu'il n'y a pas de sinistres si $\mathbb{P}(Y > 0) < 0.5$. Ceci est appelé le **seuil de classification**. Il est en général fixé ici à 50%.

Prédit/Observé	∅	Sinistre
∅	Vrai Négatif (TN)	Faux Négatif (FN)
Sinistre	Faux Positif (FP)	Vrai Positif (TP)

La démarche est ici de minimiser le nombre de Faux Négatifs, car dans une optique de souscription, un assureur préférera retirer tous les contrats risqués, quitte à retirer trop de Faux Positifs.
En effet, le tarif étant fixé par l'Etat pour la garantie Cat Nat, l'assureur ne s'expose pas à de l'anti-sélection dans ce cas.
Différentes métriques seront donc utilisées afin de valider la bonne classification du modèle.

D'abord, le ***recall*** du modèle, est donné par :

$$Recall = \frac{TP}{TP + FN}$$

Cette métrique permet de répondre à la question suivante : quelle proportion de sinistres réels a été identifiée correctement ? Elle est très importante ici car l'objectif premier est de bien détecter les sinistres.

Ensuite, l'***accuracy*** est donnée par :

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

L'*accuracy* est une mesure de performance couramment utilisée pour évaluer les modèles de classification. Elle représente le pourcentage total de prédictions correctes faites par le modèle parmi toutes les prédictions effectuées. En d'autres termes, l'*accuracy* mesure la capacité du modèle à prédire correctement les classes des échantillons. C'est la mesure généralement utilisée sur un modèle de classification mais elle est moins importante que le *recall* dans le cadre de cette étude.

Enfin, le ***F1-score*** est donnée par :

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Il mesure la parcimonie globale du modèle.

3.1.3 Modèle XGBoost

XGBoost, qui signifie « *eXtreme Gradient Boosting* », est une amélioration de l'apprentissage ensembliste basé sur les arbres de décision, principalement utilisé pour la classification et la régression. Sa force réside dans sa rapidité et son efficacité, obtenues grâce à des techniques d'optimisation comme le gradient boosting. Son fonctionnement repose sur la construction séquentielle d'une série d'arbres de décision, chaque nouvel arbre étant ajusté aux erreurs résiduelles du modèle précédent. Cette approche itérative permet au modèle de s'adapter progressivement aux structures de données complexes.
XGBoost est facilement implémentable sur python avec la librairie XGBoost. Cependant, il convient de réaliser une optimisation des paramètres généraux en utilisant la validation croisée.
Pour le modèle XGBoost, dans un cadre de classification, la métrique d'évaluation couramment utilisée est la fonction de perte d'une régression logistique. Cependant, dans la suite de cette étude, cette fonction de perte est modifiée.

Modification de la fonction de perte

Il est ici proposé de modifier la fonction de perte de XGBoost afin de pouvoir classifier le plus finement possible les assurés et donc diminuer le plus possible le nombre de Faux Négatifs (FN). Il a été choisi de pénaliser davantage lorsque qu'un sinistre est mal identifié que lorsque qu'un "non-sinistre" est bien identifié.

La log-vraisemblance de la fonction de perte pour une régression logistique, utilisée par défaut dans un modèle de classification XGBoost, est la suivante :

$$L = y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)$$

où $p_i = \frac{1}{1+e^{-\hat{y}_i}}$ et $\hat{y}_i$ est la prédiction avant application de la transformation logistique.

Elle est ici modifiée en introduisant des poids w_0 et w_1 qui vont pénaliser respectivement les erreurs de classification dans les classes 0 et 1 :

$$L = w_0 y_i \cdot \log(p_i) + w_1 (1 - y_i) \cdot \log(1 - p_i)$$

Il faut, pour implémenter XGBoost, calculer le gradient et la Hessienne de cette nouvelle fonction.
En effet, le gradient de la fonction de perte mesure comment chaque paramètre du modèle affecte la performance du modèle. Le calcul du gradient permet de mettre à jour les paramètres du modèle de manière à minimiser la perte.
La hessienne, ou matrice Hessienne, est la dérivée seconde de la fonction de perte par rapport aux paramètres du modèle. Elle mesure comment la perte change lorsque les paramètres du modèle sont modifiés. La hessienne est utilisée pour calculer la direction de descente de gradient la plus rapide et ainsi converger plus rapidement vers une solution optimale.
Dans le cas de XGBoost, le modèle utilise l'algorithme de descente de gradient stochastique qui consiste à mettre à jour les paramètres du modèle en fonction du gradient et de la hessienne de la fonction de perte. Le calcul du gradient et de la hessienne est donc essentiel pour la mise à jour efficace des paramètres du modèle et pour atteindre une solution optimale.
Le calcul est de la hessienne et du gradient sont proposés en annexe D.

Dans le cadre de ce mémoire $w_0 = \frac{\text{nombre d'observations totales}}{\text{nombre de sinistres}}$ et $w_1 = 1 - w_0$.

Cette pondération permet de pénaliser très fortement lorsque le modèle ne prédit pas correctement la survenance de sinistres. Ainsi, l'algorithme aura tendance à "surapprendre" la survenance de sinistre et à affecter des probabilités significatives aux contrats jugés comme risqués. Autrement dit, le *recall* pourra être maximisé.

Estimation des paramètres de XGBoost

Le tableau III.2 présente les configurations optimales obtenues pour le modèle, résultant d'une validation croisée à 5 blocs *(5-fold Cross-Validation)*. La métrique d'évaluation retenue est fonction de perte présentée dans la partie précédente. Ces configurations optimales sont obtenues après un processus méticuleux d'exploration des hyperparamètres, garantissant ainsi des performances optimales pour le modèle.

	Paramétrage optimal
eta	0.3
max_depth	4
min_child_weight	1
gamma	0
colsample_bytree	0.9
subsample	0.9
nround	20

TABLE III.1 – Hyperparamétrisation du modèle XGBoost.

Validation du modèle

Maintenant que les paramètres optimaux du modèle sont fixés, il faut vérifier la validité des résultats.
Les *recall*, *accuracy* et *F1-score* ainsi que la courbe *Receiver Operating Characteristic* (ROC) sont calculés sur l'ensemble de test et présentés ci-dessous :

	Valeur obtenue
Recall	88 %
Accuracy	88 %
F1-Score	93 %

TABLE III.2 – *Recall*, *Accuracy* et *F1* sur ensemble de test.

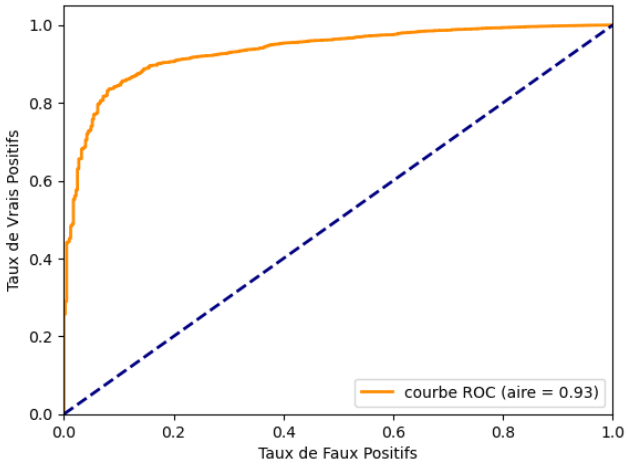


FIGURE III.1 – Courbe ROC.

Le *recall* permet de dire que la proportion de sinistres réels prédite correctement sur l'ensemble de test est de 88 % en moyenne.
L'*accuracy* est elle de 88 % en moyenne et le *F1-Score* de 93 % en moyenne.
La courbe ROC est une représentation graphique des performances d'un modèle de classification binaire pour tous les seuils de classification . Elle permet de comparer différentes méthodes de classification. Un modèle de classification est d'autant meilleur que la courbe est élevée. Par conséquent, plus l'aire sous la courbe est grande, meilleur est le classificateur. Cette aire est reflétée par la valeur AUC (*Area Under Curve*) [18], c'est-à-dire l'aire sous la courbe.
L'aire sous la courbe est, dans le contexte de cette étude, de 93 %.
Ainsi, ces métriques permettent de valider la bonne classification du modèle.

Il est également intéressant d'observer les variables les plus explicatives du modèle :

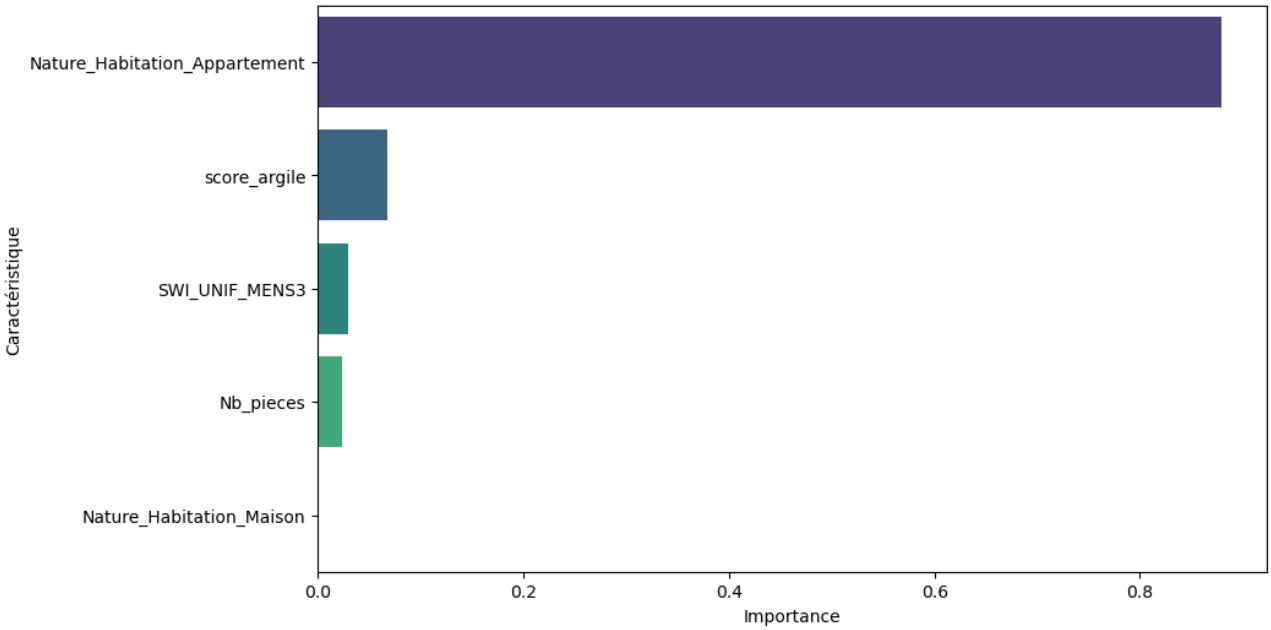


FIGURE III.2 – Importance des variables.

La variable la plus explicative du phénomène est donc la nature de l'habitation. En effet, la survenance de sinistre est bien plus probable sur une maison que sur un appartement. De plus,

la prise en compte du SWI de la deuxième partie est responsable dans une certaine mesure de l'explication de la survenance d'un sinistre. Enfin, la variable de score d'argile est, quant à elle, un peu plus responsable de l'explication de la survenance d'un sinistre. Le nombre de pièces, et donc la superficie, joue le rôle le plus mineur dans la survenance d'un sinistre.

3.2 Coût d'un sinistre sécheresse

Dans un premier temps, l'utilisation d'un GLM (*Global Linear Model*) classique est étudiée afin d'estimer le coût d'un sinistre sécheresse.
$\mathbb{E}[C(t,x)|\text{CatNat}, C(t,x) > 0, X]$ représente le coût moyen d'un sinistre, dont les caractéristiques connues du contrat appartenant à la commune x pendant la saison t (exemple : hiver 2022), sachant que les conditions de catastrophe naturelle sont respectées dans la commune, sont :

— SWI futur s'il y a état de catastrophe naturelle sécheresse (estimation de 4) ;

— Score Argile ;

— Nombre de pièces ;

— Nature habitation.

Voici les résultats obtenus pour différentes familles et fonctions liens :

Famille	**Fonction Lien**	**Log-vraisemblance**	**Déviance**	**R^2**	**AIC**	**BIC**
Gamma	Log	-6345	1075	4,6 %	12700	-2283
Gamma	Inverse	-6346	1077	4,5 %	12702	-2283
Gamma	Identité	-6342	1061	6,7 %	12694	-2283
Linéaire	/	-6716	/	3,8 %	13442	13463
Gaussienne	Log	-6715	2.09×10^{12}	4,1 %	13441	2.09×10^{12}
Gaussienne	Inverse	-6715	2.09×10^{12}	4,3 %	13441	2.09×10^{12}
Gaussienne	Identité	-6716	2.1×10^{12}	3,8 %	13442	2.1×10^{12}

TABLE III.3 – Performance des différents modèles.

Le meilleur modèle semble être le Gamma-Identité.
Cependant, le R^2 du modèle est relativement faible. En effet, le modèle explique seulement 6,7 % de la variabilité totale du problème.
La variabilité du coût doit être trop grande et il semble difficile de capter suffisamment d'informations au niveau d'échelle de la commune.
Les données manquent de précisions pour être en mesure de prédire ce coût. Il serait par exemple essentiel d'avoir une étude géotechnique des sols spécifique au contrat, et non pas à l'échelle de la commune. De même, connaître des informations plus précises sur le type de matériaux utilisés, la profondeur des fondations, l'environnement végétal aux alentours ou encore les éventuels moyens de prévention mis en place, aideraient grandement à améliorer ce R^2.
Néanmoins, la plupart des assureurs ne demandent pas ces informations lors de la souscription du contrat, il ne sera donc pas possible d'améliorer la modélisation dans le cadre de ce mémoire.

De plus, la normalité des résidus n'est pas vérifiée avec ce modèle GLM :

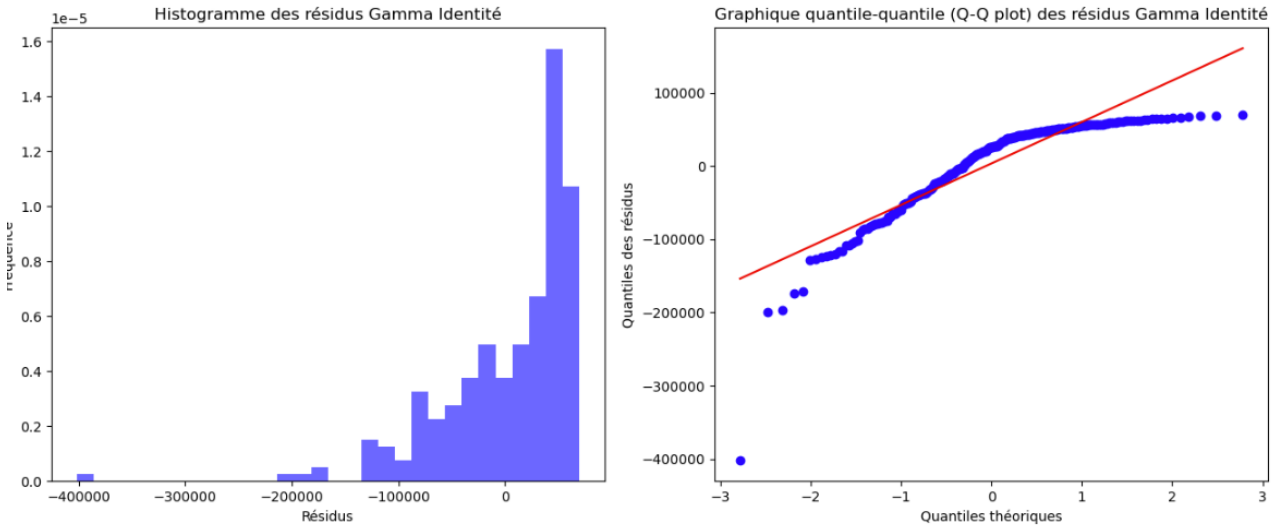


FIGURE III.3 – Tets de normalité des résidus du modèle Gamma Identité.

Il est cependant intéressant de constater que certaines variables proposées semblent être significatives dans l'explication du coût :

Variable	**Coefficient**	**Ecart type erreur**	**P-value**
Constante	3.29×10^4	7565	0.000
Score Argile	1.018×10^4	1226	0.000
SWI	-5.631×10^4	3883	0.000
Nombre de pièces	2484	1313	0.059
Nature_Habitation_Appartement	929	2990	0.756

TABLE III.4 – Significativité des variables du modèle Gamma Identité.

Des modèles plus simples sont alors testés : coût moyen par département et coût moyen global. Des comparaisons entre les RMSE des différents modèles sont présentées dans le tableau III.5.

Modèles	**RMSE**
Gamma identité	15 500
Coût moyen département	15 850
Coût moyen global	15 150

TABLE III.5 – RMSE des différents modèles.

Les RMSE sont relativement grands. En effet, il semble relativement difficile de prédire précisément le coût conditionnel.
Cependant, c'est le modèle Coût moyen global qui donne le plus petit RMSE. C'est donc celui-ci qui est utilisé.
De plus, l'étude MRN 2018 [26] semble aller dans ce sens puisqu'elle évoque qu'il est difficile d'expliquer le coût et préconise d'utiliser un modèle plus simple de type coût moyen par département. Il manque peut-être certaines données sur certains départements et il est choisi, conformément aux résultats ci-dessus, de prendre le coût moyen global.

Ainsi, il est au final supposé que :

$$\mathbb{E}[C(t,x)|\text{Cat Nat}, C(t,x) > 0, X] = 20\ 870{,}18$$

3.3 Sinistralité estimée pour chaque contrat

3.3.1 Présentation des résultats

Au final, il est possible d'estimer la sinistralité sécheresse de chaque contrat, conditionnellement au fait que les critères de déclaration d'état de catastrophe naturelle sécheresse soient vérifiés.
Le modèle initial est simplifié de la manière suivante :

$$\mathbb{E}[C(t,x)|\text{CatNat}, X] = 20870,18 \cdot (1 - \mathbb{P}(C(t,x) = 0)|\text{CatNat}, X)$$

Pour rappel, la sinistralité conditionnelle estimée est la sinistralité conditionnelle correspondant à une saison (exemple : hiver 2022).
Voici les résultats obtenus pour 5 contrats :

Pièces	Habitation	Score_argile	SWI	Sinistre constaté	Proba 0	Coût estimé
4,0	Maison	2,812	0,412	504,0	0,563	11 749
6,0	Maison	1,207	0,393	0,0	0,021	438
5,0	Maison	1,947	0,239	156,0	0,054	1 127
2,0	Appartement	1,707	0,450	0,0	0,0092	192
2,0	Appartement	1,981	0,238	0	0,0062	129

TABLE III.6 – Sinistralité conditionnelle estimée par contrat.

3.3.2 Validation du modèle

Le RMSE de ce modèle global vaut 1 625,86.
Ainsi, en moyenne, l'écart entre la sinistralité estimée et la sinistralité réelle est de 1 625 €.

Ceci peut paraître élevé, cependant, il faut rappeler que la distribution des sinistres est très inégalement répartie entre 0 et plus de 400 000 € :

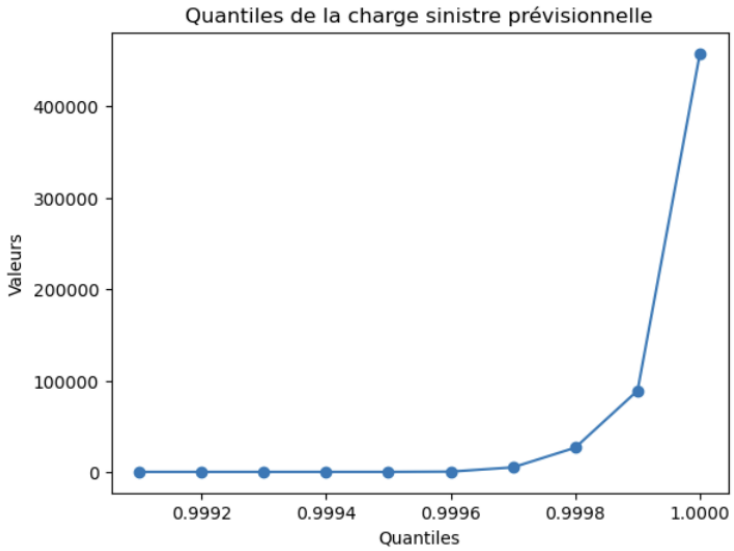


FIGURE III.4 – Distribution du coût réel - quantiles de 99,9 % à 100 %.

Il est donc difficile de donner une bonne estimation pour chaque contrat, et ceci explique cette grande valeur de RMSE.

Cependant, pour valider la pertinence du modèle, la base de données est divisée en différents *clusters* comme réalisé lors de l'étude de la mesure *lift*.
La mesure *lift* décrite dans le mémoire de Christian Chow sur l'"Utilisation des données télématiques pour l'analyse de la sinistralité automobile" (page 106) [3].
Le *lift* est défini ainsi :

$$lift_i = \frac{taux\ de\ survenance\ du\ groupe\ i}{taux\ de\ survenance\ moyen}$$

où les groupes sont construits de cette manière : la base de données est ordonnée par ordre croissant de la sinistralité estimée, en dix parts égales (pour la dernière part, elle est à nouveau divisée en cinq de la même manière).
Ainsi, pour chaque part constituée, la sinistralité estimée attribuée à la part et la sinistralité réelle constatée sur cette part sont comparées et permettent de juger de la qualité du modèle.

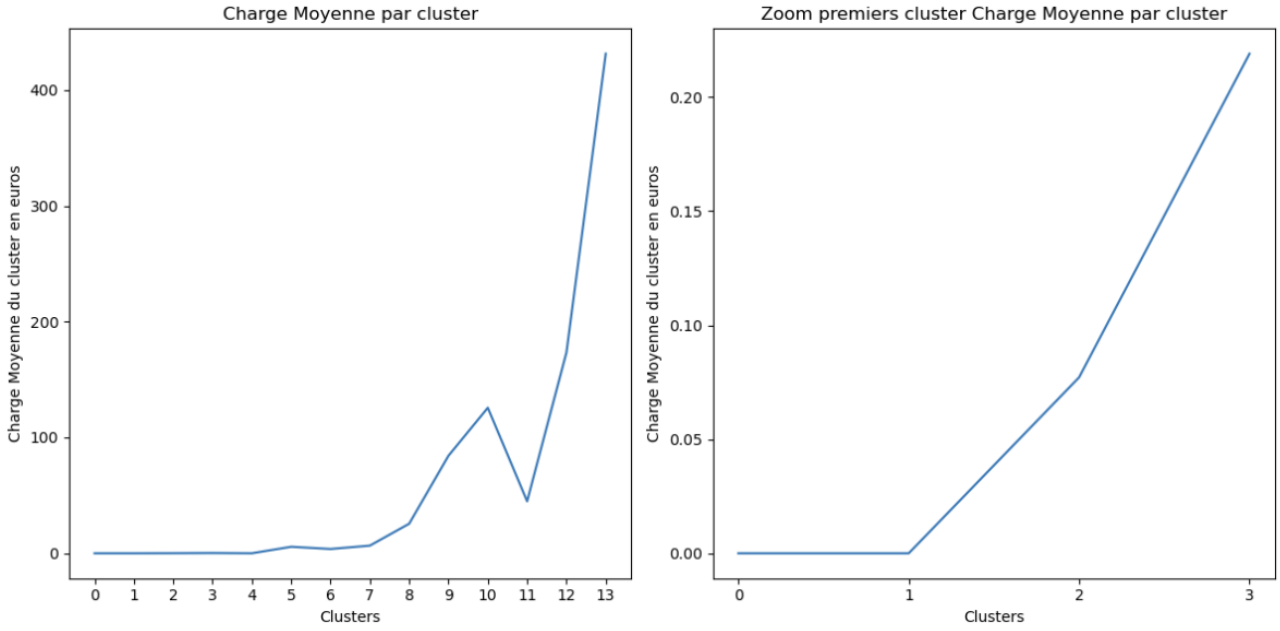


FIGURE III.5 – Charge moyenne pour chaque groupe constitué.

Il semble que le modèle attribue bien des coûts plus élevés aux contrats coûtant en moyenne plus chers.

Ci-dessous, sont représentées les moyennes des charges constatées en euros en fonction des coûts attribués pour chaque *cluster*. En théorie, le modèle parfait attribuerait exactement le même coût que la moyenne des charges constatées sur le *cluster*, il est alors représenté comme la fonction identité.

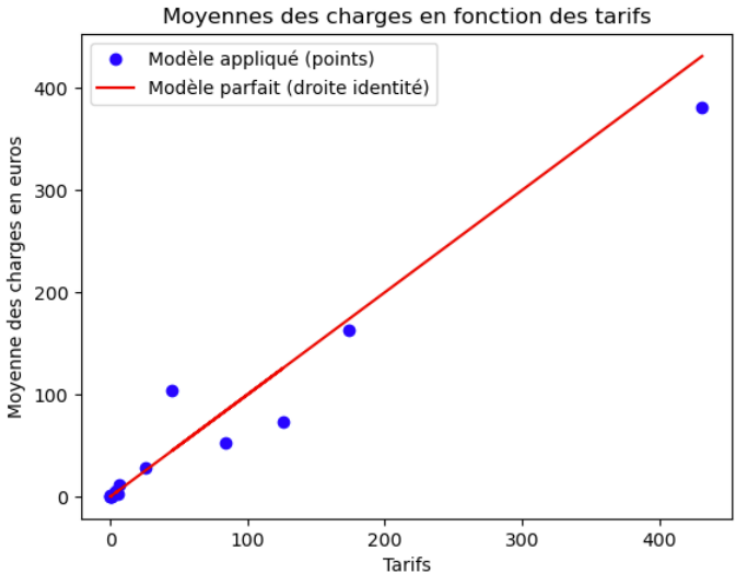


FIGURE III.6 – Coût conditionnel estimé moyen et charge moyenne par groupe constitué.

La sinistralité estimée attribuée au groupe considéré est en moyenne proche de la charge moyenne constatée sur le groupe.

Ces graphiques permettent, malgré le RMSE relativement élevé, de valider le modèle.

Compte-tenu de la répartition très inégale de la sinistralité des contrats dans les communes qui ont déclaré l'état de catastrophe naturelle sécheresse 3.3.2, il serait cohérent d'avoir une sinistralité estimée relativement discriminante en fonction des caractéristiques du contrat et surtout de sa localisation et donc de son exposition au phénomène de retrait-gonflement des argiles 2.2.
Il est constaté pour les 5 contrats en 3.3.1 une bonne discrimination.
Afin d'avoir un point de vue plus global, l'histogramme de la probabilité estimée est représenté ci-dessous :

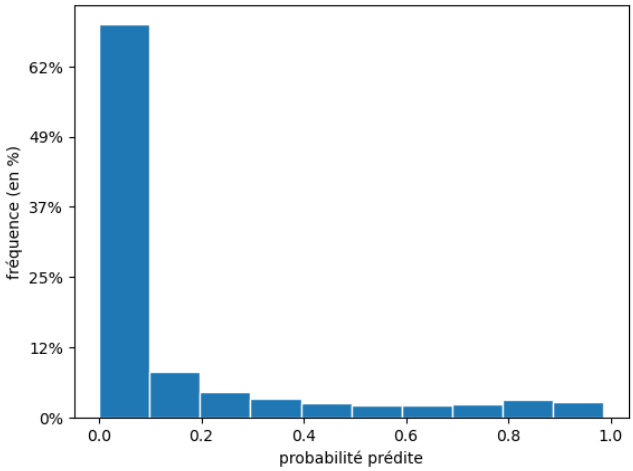


FIGURE III.7 – Histogramme probabilité prédite sur l'échantillon test.

La probabilité estimée semble être bien discriminante.

De plus, au global, sur l'échantillon de test (lignes tirées au hasard entre 2018 et 2022), les sinistres réels représentent 13 031 307 € tandis que l'estimation représente 12 376 969 €, ce qui veut dire que 94 % de la sinistralité globale est captée.
Enfin, les 5 % des contrats les plus coûteux, c'est-à-dire jugés les plus risqués par le modèle, sont représentés.
La référence est ici placée sur l'année 2022 (horizon +1 an lors de la rédaction de ce mémoire) pour la saison été.

Voici une représentation graphique des 5 % des contrats les plus risqués du portefeuille de notre client situé en Auvergne-Rhône-Alpes :

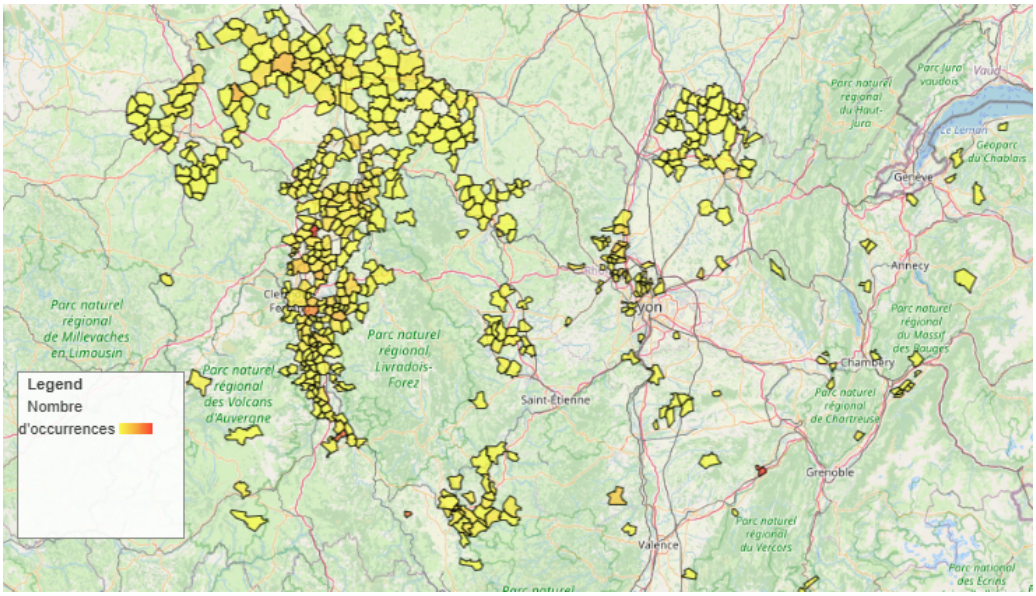


FIGURE III.8 – 5 % des contrats les plus risqués du portefeuille client.

D'abord, la ressemblance de cette carte avec le score d'exposition au RGA présenté en 2.2 est à souligner. En effet, l'exposition fournie par BRGM semble être un bon indicateur du risque lié au contrat. Cependant, le modèle intègre également les variables Nature habitation, SWI moyen estimé ainsi que le Nombre de pièces. L'impact de ces variables n'est pas visible sur ce graphique mais est présenté dans les résultats du modèle ci-dessus.

Chapitre 4

Conclusion

Dans cette partie, une modélisation coût-fréquence de la sinistralité par contrat a été proposée, conditionnellement au fait que les critères de déclaration de catastrophe naturelle soient respectés. L'information apportée par ce conditionnement permet d'affiner l'estimation du SWI futur ainsi que de réduire la base de données aux communes respectant le critère géologique. L'apport de cette information affine normalement les prédictions.

Faisant face à une base de données très déséquilibrée, l'emploi d'un modèle XGBoost avec fonction de perte modifiée a été proposé afin de modéliser la fréquence. Le but de ce modèle est de détecter les contrats qui auront des sinistres. En effet, cette détection pourrait s'avérer être très utile pour la sélection des risques d'un assureur. Plusieurs tests de validation ont permis de valider le modèle.
Les coûts étant également très inégalement répartis, allant de 0 à plus de 400 000 € dans la base de données de cette étude, l'emploi de modèles linéaires classiques ne donne pas de résultats satisfaisants. Le rapport MRN de 2018 [26] démontrant qu'il est difficile de prédire le coût d'un sinistre, il a été préféré ici de retenir le coût moyen comme approximation du coût d'un sinistre.

Enfin, l'agrégation de coût et de la fréquence ont permis d'obtenir une estimation de la sinistralité sécheresse par contrat, conditionnellement au déclenchement de catastrophe naturelle. Les mesures *lift* ont permis de valider le modèle, malgré un RMSE relativement grand.

Quatrième partie

Applications, limites, critiques

Chapitre 1

Applications

L'étude permet d'estimer en premier lieu l'impact de potentielles lois futures ainsi que du changement climatique. Elle permet aussi à l'assureur d'appliquer certaines stratégies afin de mieux piloter son risque.

1.1 Impact réglementaire et changement climatique

1.1.1 Loi Rousseau

La Loi Rousseau faciliterait la vérification du critère météorologique, en passant la durée de retour de 25 ans à 10 ans ; permettant ainsi la déclaration d'arrêtés Cat Nat plus facilement. Conformément à la partie 3.4, il a été observé qu'avec le seuil actuel, les critères sont très régulièrement validés, en particulier l'été. Il ne serait pas surprenant que la **probabilité de respecter le critère météorologique augmente significativement**.
Il serait d'ailleurs en théorie possible d'utiliser le modèle proposé en section 3.3, mais les limites de celui-ci sont exposées ci-dessus.

De plus, la loi Rousseau impacterait la **présomption de causalité**, autrement dit, il serait à la charge de l'assureur de démontrer qu'un sinistre déclaré par un assuré n'est pas causé par la sécheresse dont fait part l'arrêté Cat Nat en vigueur dans la commune. Jusqu'à présent, environ deux tiers des sinistres étaient classés sans suite et coûtaient à l'assureur uniquement les frais d'experts d'environ 500 €. Ainsi, l'espérance du coût d'un sinistre devrait augmenter.

1.1.2 Changement climatique et impact sur les SWI

Dans le cadre de ce mémoire, la prise en compte du réchauffement climatique pourrait ne pas avoir été perçue. En effet, l'impact à long terme du réchauffement climatique n'est pas explicitement inclus. Sa prévision demeure très incertaine, comme le met en évidence l'examen des scénarios du GIEC, ce qui rend difficile toute proposition à long terme.

Cependant, la section 4 démontre que les SWI intégrés dans le modèle de coût sont déjà ajustés de l'impact du réchauffement climatique. En d'autres termes, lorsque des conditions météorologiques particulières surviennent dans une localité, **il est possible d'anticiper que le SWI sera invariablement l'une des deux valeurs les plus basses jamais enregistrées sur la maille lors du mois en question. Ainsi, le SWI évolue naturellement vers des valeurs de plus en plus faibles en tenant compte des données historiques.**

1.2 Stratégies pour l'assureur

1.2.1 Souscription

A partir des études réalisées dans ce mémoire, certaines applications en assurance sont proposées.
Il semble dors et déjà pour lui, possible de mieux piloter son risque en mieux identifiant son risque. Il peut alors prendre des décisions stratégiques au niveau de la souscription, mais également mettre des moyens de prévention en place.

A partir des probabilités issues du modèle XGBoost, certaines stratégies de souscription peuvent être proposées.

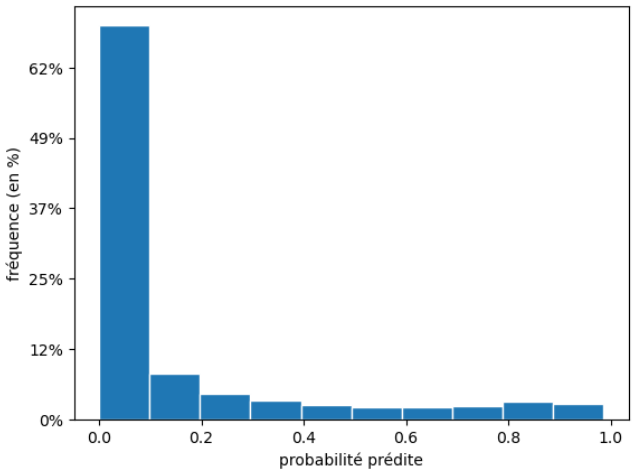


FIGURE IV.1 – Histogramme probabilité de survenance prédite par XGBoost.

Par exemple, en fixant un seuil de classification de 50 %, comme vu en 3.1.3, environ 15 % des contrats jugés les plus à risque par le modèle pourraient être identifiés. Un assureur plus averse au risque pourrait également choisir un seuil de classification de 20 %. Dans ce cas, c'est environ 25 % des contrats les plus risqués qui ne seraient pas souscrits.

Sous réserve que les communes n'aient pas fait la demande exceptionnelle de déclaration de catastrophe naturelle, l'assureur pourrait également souscrire tous les contrats des communes ne respectant pas le critère géologique :

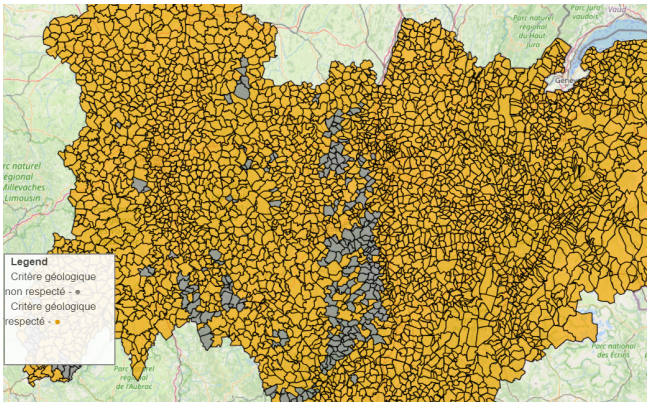


FIGURE IV.2 – Critère géologique Auvergne-Rhône-Alpes.

En effet, lorsque ce critère n'étant pas respecté, il est en théorie impossible de déclarer l'état de catastrophe naturelle au niveau de la commune. Ainsi, la garantie Cat Nat ne peut pas être déclenchée et les sinistres ne donneront pas lieu à une indemnisation.

Cette approche fait, néanmoins, un lien avec l'actualité de l'**inassurabilité** en assurance. En effet, les sinistres sécheresse coûtant de plus en plus chers aux assureurs, sous l'effet du changement climatique, de plus en plus d'assureurs réfléchissent à durcir leur politique de souscription. Cette proposition de réponse présente donc des enjeux sociétaux. En effet, certaines habitations ne seraient plus assurées, en particulier les maisons exposées dans les zones à risque et des assurés ne seraient plus couverts face à ces risques de plus en plus coûteux.

1.2.2 Critères de sélection pour les assureurs

En l'absence de contrôle sur les tarifs, les assureurs doivent se baser sur des critères de sélection appropriés pour leurs contrats. Résilier ou ne plus souscrire certains contrats devient alors la seule option. Les recommandations qui peuvent être tirées de ce mémoire incluent d'éviter les communes les plus exposées aux phénomènes de retrait-gonflement des argiles (RGA), comme décrit en section 2.2. De plus, privilégier l'assurance pour les appartements peut être judicieux, car ces derniers présentent une fréquence de sinistres beaucoup plus faible III.2. Par ailleurs, même si les données actuelles ne le montrent pas, il serait préférable de privilégier les habitations avec des fondations solides, construites plus récemment (conformément à la loi Elan de 2018, [19]).

Enfin, il est à l'issue de ce mémoire suggéré de continuer à rassembler les informations les plus détaillées possibles telles que la taille du bien, le type de construction, les études de sol avant la souscription et l'environnement végétal, afin d'affiner les évaluations des coûts, d'améliorer le modèle (XGBoost). De plus, bénéficier d'une localisation très précise (coordonnées GPS) permettrait d'affiner la variable "Score argile", qui pour l'instant reflète l'exposition de la commune au phénomène de retrait-gonflement des argiles et non pas l'exposition spécifique du contrat.

1.2.3 Moyens de prévention

L'assureur dispose également de la possibilité de mettre en œuvre des mesures préventives. En effet, il pourrait exiger, pour tous les contrats à risque (maisons, dans les communes répondant à des critères géologiques spécifiques), la fourniture d'études géotechniques des sols. De plus, l'assureur pourrait demander au futur assuré de démontrer qu'il limite la présence de végétation autour de son habitation. Enfin, il pourrait envisager de proposer des "contrats durables", consistant à financer une partie des mesures préventives de l'assuré telles que l'injection hydraulique des trottoirs ou le renforcement des fondations.

Chapitre 2

Limites du régime Cat Nat

Cette étude a permis de mettre en lumière certaines limites du système d'indemnisation des sinistres sécheresse liés au retrait gonflement des argiles.

2.1 Critère météorologique

Le régime d'indemnisation Cat Nat sécheresse actuel présente deux défauts principaux.

1. D'abord, un arrêté de Cat Nat est déclaré saison par saison à partir des critères de définition de caractère anormal du SWI eux-mêmes donnés mois par mois. Cependant, il n'est presque jamais observé de sinistre se produisant en dehors de la période estivale. Un critère avec une vision estivale ou annuelle pourrait être plus approprié.

2. De plus, la catastrophe naturelle ne peut être déclenché que lorsque le SWI dépasse un certain seuil historique (2 valeurs les plus basses). Or, il est possible d'atteindre une valeur proche de ce minimum, et donc de connaître une véritable sécheresse qui cause des dégâts. Ceci est particulièrement vrai dans le contexte du réchauffement climatique, où les seuils du critère météorologique seront toujours repoussés de plus en plus bas, et une année particulièrement sèche rentrera alors dans la norme, ne permettant plus aux communes de bénéficier d'arrêtés Cat Nat et donc aux assurés d'être indemnisés pour leurs sinistres.

Il serait alors possible de remettre en question la pertinence de ce critère.

2.2 Est-il toujours possible de parler de catastrophe naturelle ?

La question se pose de savoir s'il est possible de continuer à qualifier un événement de catastrophe naturelle s'il devient de plus en plus fréquent et presque courant à l'avenir. La Loi Rousseau vise à redéfinir le critère météorologique en prenant en compte une période de 10 ans. Autrement dit, un événement survenant tous les 10 ans pourrait-il encore être considéré comme une catastrophe naturelle ? Peut-on toujours parler de risque rare et extrême, tel que défini par l'INSEE 1.1, dans un contexte où des événements autrefois rares deviennent plus fréquents ?

En réalité, il semble que les assurés et les assureurs, par le biais de la garantie Cat Nat obligatoire sur les contrats MRH, contribuent progressivement au financement du renouvellement du parc immobilier et à son adaptation à un climat de plus en plus sec. Ne faudrait-il pas déléguer cette responsabilité à l'État et aux politiques d'aménagement du territoire, afin de réviser les

approches traditionnelles de gestion des risques et de promouvoir activement des politiques d'aménagement plus résilientes face aux évolutions climatiques ?

Cinquième partie

Conclusion

Afin d'élaborer un modèle prédictif de la sinistralité sécheresse en Multirisque Habitation (MRH), l'étude menée s'est articulée autour de quatre grandes parties.
D'abord, la première partie a présenté le contexte de l'indemnisation des sinistres sécheresses en MRH et les enjeux concernant sa modélisation. **La compréhension du contexte réglementaire changeant et du régime d'indemnisation Cat Nat sécheresse** ont représenté les principaux défis de cette partie. Il a ensuite été possible de proposer un modèle rigoureux de prédiction de la sinistralité sécheresse qui tient compte à la fois du cadre réglementaire mais aussi du risque réellement constaté chez l'assureur au niveau du contrat.

Dans un deuxième temps, l'étude s'est portée sur la modélisation des SWI et du critère météorologique.
Bien que nécessitant une réduction de la dimension des données et l'utilisation de modèles plus complexes que ceux habituellement employés pour les séries temporelles, la modélisation des SWI a été achevée.
La projection des SWI a été le premier enjeu de cette deuxième partie. En effet, **avoir une estimation du SWI futur par commune et par saison permet d'avoir plus d'informations pour prédire la sinistralité**.
Le deuxième enjeu de cette partie a été la **modélisation du critère météorologique**. Elle est cohérente mais semble néanmoins sous-estimée. Cela est dû à plusieurs éléments dont un changement brutal de la sinistralité avec le réchauffement climatique, des limites liées au modèle utilisé et des hypothèses fortes. Cependant, la modélisation de ce critère a déjà permis d'aboutir à certaines conclusions et a aussi pour but de souligner l'importance de la prise en compte du cadre réglementaire pour les assureurs.
Enfin, cette partie a proposé de visualiser l'impact du changement proposé par la Loi Rousseau en 2023 sur le critère météorologique. **Le modèle prédit une hausse des déclarations de catastrophes naturelles au sein des communes**.

Dans un troisième temps, l'étude s'est dirigée vers la modélisation du risque réellement observé chez les assureurs en Auvergne-Rhône-Alpes. Pour cela, les données portefeuille de notre client ont été réduites uniquement aux contrats ayant connu une déclaration de catastrophe naturelle dans leur commune.
Le premier apport de cette partie a été la construction de la variable exogène au contrat Score Argile correspondant à l'exposition au phénomène retrait-gonflement des argiles. Cette variable semblait être corrélée aux déclarations de catastrophes naturelles sécheresses historiques et elle a alors été intégrée dans la base de données. **L'apport de cette variable ajoute alors de l'information pour la prédiction de la sinistralité**. Une deuxième variable exogène appelée SWI futur, issue de la deuxième partie, a également pu être intégrée dans les données. **Cette variable ajoute également de l'information, puisqu'à ce stade, il est connu que la catastrophe naturelle est déclarée et que le SWI a dépassé un certain seuil**.
Par ailleurs, il a été constaté que la distribution des sinistres était particulière. En effet, la survenance de sinistres est très rare (de l'ordre de 0,09 %). L'utilisation de modèles linéaires classiques était donc caduque dans ce cas. De plus, les coûts des sinistres sont également difficiles à prédire conformément au rapport MRN de 2018 [26]. Ce travail a donc proposé l'étude d'un modèle coût-fréquence.
Le coût, après mise en évidence de la difficulté de modélisation par un GLM classique, a été modélisé comme étant le coût moyen pour tous les contrats. La fréquence a, quant à elle, été estimée à l'aide d'un modèle XGBoost avec fonction de perte modifiée, utilisant les variables endogènes et exogènes du contrat.
Ce modèle, bien discriminant, permet d'identifier facilement les contrats les plus à risque.

Enfin, plusieurs tests ont été réalisés en fin de partie et ont validé la qualité du modèle.

La quatrième et dernière partie de ce mémoire a présenté les applications possibles du modèle et les limites actuelles.
Dans un premier temps, **deux indicateurs de sélection (probabilités XGBoost et critère géologique) ont par exemple été proposés**.
Dans un deuxième temps, les limites relatives au régime d'indemnisation Cat Nat sécheresse ont été présentées. Ces limites semblent être, à l'heure actuelle, assez importantes et la compréhension de ce régime est nécessaire pour faire évoluer notamment le contexte réglementaire. En effet, les règles, et en particulier le critère météorologique permettant la déclaration de la catastrophe naturelle au niveau de la commune, ne semblent pas être suffisamment adaptées au contexte de réchauffement climatique actuel.

Par ailleurs, la proposition de nouveaux critères, plus adaptés, pourrait être étudiée.
Une autre voie d'amélioration de ce travail pourrait être la mise en œuvre de l'agrégation de la deuxième et de la troisième partie. En effet, si les probabilités de déclaration de catastrophe naturelle sécheresse de la deuxième partie peuvent être plus précises et si le portefeuille permettant de simuler les coûts conditionnels au déclenchement de l'état de catastrophe naturelle peut intégrer plus de variables, alors le modèle pourrait être mis en œuvre et le coût du risque par contrat (prime sécheresse) pourrait alors apparaître pour l'assureur. Il serait également possible d'étudier plus en détails la modélisation du coût d'un sinistre car les modèles linéaires classiques ne semblent pas aboutir à une modélisation satisfaisante.

Bibliographie

[1] *Algorithme ADAM*. URL : https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/.

[2] CCR. *BILAN CAT NAT CCR 1982-2021*. Rapp. tech. CCR, 2022.

[3] Christian CHOW. "Utilisation des données télématiques pour l'analyse de la sinistralité automobile". Mém. de mast. ISUP, 2019.

[4] *CLIMSEC Project by Météo-France*. https://climserv.meteo.fr/climsec/. Accessed on : [8 setpembre 2022]. 2011.

[5] Daphné LE CORNEC. "Étude d'un modèle de coût du risque sécheresse en France". Mém. de mast. Dauphine Université Paris, 2017.

[6] Gregoire DAVID. *Communes GeoJSON for France*. https://github.com/gregoiredavid/france-geojson/blob/master/communes.geojson. Accessed on : [6 décembre 2022].

[7] *Définition swi uniforme météo France*. URL : https://www.rhone-mediterranee.eaufrance.fr/sites/sierm/files/content/migrate_documents/indicateur_swi.pdf.

[8] Alexandre DELORME. "Assurances contre la sécheresse au XXIe siècle : perspectives d'évolution". Mém. de mast. Dauphine Université Paris, 2023.

[9] Matthew F DIXON, Igor HALPERIN et Paul BILOKON. *Machine learning in Finance*. T. 1406. Springer, 2020.

[10] *Données Géorisques exposition Retrait-gonflement des argiles*. URL : https://www.georisques.gouv.fr/donnees/bases-de-donnees/retrait-gonflement-des-argiles.

[11] *Etude de sol : prix, obligation et déroulement, Figaro Immobilier, 2022*. URL : https://immobilier.lefigaro.fr/faire-construire/guide-de-construction-immobilier/1022-etude-de-sol-prix-obligation-et-deroulement/.

[12] *Fonctionnement LSTM*. URL : https://larevueia.fr/quest-ce-quun-reseau-lstm/.

[13] Météo FRANCE. *Données publiques Météo France historique SWI 1969-2021*. URL : https://meteofrance.fr.

[14] *Garantie Cat Nat et périls couverts*. URL : https://catastrophes-naturelles.ccr.fr/perils-couverts.

[15] Jiang GUO. "Backpropagation through time". In : *Unpubl. ms., Harbin Institute of Technology* 40 (2013), p. 1-6.

[16] *https ://infoterre.brgm.fr/rapports/RP-54862-FR.pdf, (Fonctionnement RGA), consulté le 8 septembre 2022.*

[17] *https ://www.legifrance.gouv.fr/loda/id/JORFTEXT000044589864/ (Réforme régime indemnisation Cat Nat, 2021), consulté le 5 janvier 2023.*

[18] *L'AUC ROC, Kobia*. URL : https://kobia.fr/classification-metrics-auc-roc/.

[19] *Loi Elan, 23 Novembre 2018*. URL : https://www.legifrance.gouv.fr/jorf/id/JORFARTI000037639713.

[20] Marcos Lopez de Prado. "Stochastic Flow Diagrams Add Topology to the Econometric Toolkit". In : *Available at SSRN 2381301* (2014).

[21] Leo LOVISOLO. "Construction d'un générateur climatique permettant l'évaluation d'options basées sur un indice sécheresse en France". Mém. de mast. Euro-Institut d'Actuariat, 2021.

[22] "Ordonnance n° 2023-78 du 8 février 2023 relative à la prise en charge des conséquences des désordres causés par le phénomène naturel de mouvements de terrain différentiels consécutifs à la sécheresse et à la réhydratation des sols, LégiFrance". In : (2023).

[23] *Plans de prévention des risques naturels prévisibles (PPR) Mouvements différentiels de terrain liés au phénomène de retrait-gonflement des argiles*. Rapp. tech. Direction départementale de l'équipement du Tarn, 2008.

[24] "Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels". In : *Légifrance* (2019).

[25] *Proposition de loi Rousseau, 6 avril 2023*. URL : https://www.vie-publique.fr/loi/289228-maisons-fissurees-retrait-gonflement-argile-proposition-de-loi-rousseau.

[26] *Rapport MRN : Sécheresse Géotechnique 2018*. Rapp. tech. Mission Risques Naturels, 2018.

[27] Jana Friederike SCHULTE. "Modélisation du risque subsidence en France métropolitaine". Mém. de mast. Institut de Statistique de Sorbonne Université, 2016.

[28] Jean-Michel Soubeyroux et al. "Safran-Isba-Modcou (SIM) : Un outil pour le suivi hydrométéorologique opérationnel et les études". In : *La Météorologie* 63 (2008), p. 40-45.

Table des figures

Liste des tableaux

Annexes

Annexe A

Théorie sur l'analyse en composantes principales

L'individu i est représenté par le point A_i de $\mathbb{R}^p$ de coordonnées $x_i^1, \cdots, x_i^p$ qui constituent le vecteur $\underline{x}_i$:

$$i \mapsto A_i\left(\underline{x}_i\right).$$

Le poids p_i est affecté au point A_i. Le barycentre G des points (A_i, p_i) est à l'origine car les caractères sont centrés : $\overrightarrow{OG} = \sum_{i=1}^n p_i \overrightarrow{OA}_i$ a pour j-ième composante $g^j = \sum_{i=1}^n p_i x_i^j = 0$. Une métrique M dans $\mathbb{R}^p$ est associée à la distance euclidienne d. Pour deux points $A_i\left(\underline{x}_i\right)$ et $A_{i'}\left(\underline{x}_{i'}\right)$, il est possible d'écrire la distance ainsi :

$$d^2\left(A_i, A_{i'}\right) = \left\|\underline{x}_i - \underline{x}_{i'}\right\|^2 = \left(\underline{x}_i - \underline{x}_{i'}\right)' M \left(\underline{x}_i - \underline{x}_{i'}\right).$$

Dans ce mémoire, la métrique euclidienne est retenue. Les SWI des individus (mailles numérotées de 1 à 3000) sont représentés dans l'espace des périodes (exemple : janvier, année 1969).

$$d\left(A_i, A_{i'}\right) = \sqrt{\sum_{j=1}^p \left(x_i^j - x_{i'}^j\right)^2}$$

Un sous-espace de projection F_r, de dimension fixée r, est déterminé tel que la projection du nuage de points A_i sur ce sous-espace, soit l'image la plus fidèle possible du nuage initial.
Avec P_i la projection de A_i sur F_r, il vient :

$$d\left(P_i, P_{i'}\right) = \left\|\overrightarrow{P_i P_{i'}}\right\| \leq \left\|\overrightarrow{A_i A_{i'}}\right\| = d\left(A_i, A_{i'}\right)$$

Donc :

$$\underbrace{\sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} \left\|\overrightarrow{P_i P_{i'}}\right\|^2}_{\text{dépend du sous-espace } F_r} \leq \underbrace{\sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} \left\|\overrightarrow{A_i A_{i'}}\right\|^2}_{\text{nombre fixe}}$$

Pour que, dans leur ensemble, les distances entre les points P_i soient les plus proches possibles des distances entre les points A_i, F_r est déterminée de telle manière à ce la somme $\sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} \left\|\overrightarrow{P_i P_{i'}}\right\|^2$ soit maximale. F_r est un sous-espace de dimension r qui rend maximale la somme $\sum_{i=1}^n \sum_{i=1}^n p_i p_{i'} \left\|\overrightarrow{P_i P_{i'}}\right\|^2$.
Enfin, le théorème suivant est introduit.

Théorème 1. *La qualité globale de représentation du nuage de points A_i par projection sur F_r est définie comme le rapport :*

$$\frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} p_i p_{i'} \left\|\overrightarrow{P_i P_{i'}}\right\|^2}{\sum_{i=1}^{n}\sum_{i=1}^{n} p_i p_{i'} \left\|\overrightarrow{A_i A_{i'}}\right\|^2} = \frac{\sum_{i=1}^{n} p_i \left\|\overrightarrow{GP_i}\right\|^2}{\sum_{i=1}^{n} p_i \left\|\overrightarrow{GA_i}\right\|^2} = \frac{\textit{inertie des projections } P_i \textit{ par rapport à } G}{\textit{inertie des points } A_i \textit{ par rapport à } G}.$$

Elle est aussi appelée pourcentage d'inertie expliquée par le sous-espace F_r. La qualité globale de représentation par projection sur F_r est égale à $\frac{\lambda_1+\lambda_2+\cdots+\lambda_r}{\lambda_1+\lambda_2+\cdots+\lambda_p} = \frac{\lambda_1+\lambda_2+\cdots+\lambda_r}{Trace\ (X'DXM)}$. C'est la somme des qualités globales de représentation par projection sur les axes principaux qui engendrent F_r.

Annexe B

Liens RNN et AR (AutoRégressif), les RNN comme des AR non linéaires

Les processus autorégressifs auraient pu être envisagés pour la modélisation des séries temporelles SWI. Cependant, ces approches ne peuvent pas être utilisées dans le cadre de ce mémoire. Dans cette perspective, une représentation des Réseaux de Neurones Nécurrents (RNN) en tant que généralisations des processus AutoRégressifs AR est proposée dans un cas simple. Le cas général n'est ici pas démontré.
Cette approche vise à garantir que l'adoption d'un modèle plus complexe de type réseau de neurones n'entraîne pas nécessairement une diminution des performances par rapport à un modèle plus simple.
Le contenu de cette annexe est inspiré du livre [9] qui présente notamment les modèles classiques, puis les liens avec le deep-learning.

Le cas le plus simple d'un RNN avec une unité cachée ($H = 1$), sans fonction d'activation et où la dimension du vecteur d'entrée ($P = 1$) est considéré.

Soit x l'input, y :l'output et z l'état caché.
Il est supposé que $W_z^{(1)} = \phi_z, |\phi_z| < 1, W_x^{(1)} = \phi_x, W_y = 1, b_h = 0$ et $b_y = \mu$.

Il est alors possible montrer que $f^{(1)}_{W(1),b(1)}(X_t)$ est un modèle autorégressif, $AR(p)$, d'ordre p avec des coefficients autorégressifs à décroissance géométrique $\phi_i = \phi_x \phi_z^{i-1}$:

$$\begin{aligned} z_{t-p} &= \phi_x x_{t-p} \\ z_{t-T+2} &= \phi_z z_{t-T+1} + \phi_x x_{t-T+2} \\ \dots &= \dots \\ z_{t-1} &= \phi_z z_{t-2} + \phi_x x_{t-1} \\ N\hat{x}_t &= z_{t-1} + \mu \end{aligned}$$

puis

$$\hat{x}_t = \mu + \phi_x \left(L + \phi_z L^2 + \cdots + \phi_z^{p-1} L^p\right)[x_t]$$

$$= \mu + \sum_{i=1} \phi_i x_{t-i}$$

Il est à noter que si l'architecture est modifiée de sorte à que les poids de récurrence $W^{(1)}_{z,i} = \phi_{z,i}$ dépendent du retard, la couche cachée non activée est alors la suivante :

$$z_{t-i} = \phi_{z,i} z_{t-i-1} + \phi_x x_{t-i}$$

ce qui donne :

$$\hat{x}_t = \mu + \phi_x \left(L + \phi_{z,1} L^2 + \cdots + \prod_{i=1}^{p-1} \phi_{z,i} L^p \right) [x_t]$$

et donc les poids dans ce modèle $AR(p)$ sont : $\phi_j = \phi_x \prod_{i=1}^{j-1} \phi_{z,i}$ ce qui permet une présentation plus flexible de la structure d'autocorrélation que le RNN simple qui est limité aux poids à décroissance géométrique. Enfin, il est à noter qu'un RNN linéaire avec un nombre infini de retards et sans biais correspond à un lisseur exponentiel, $z_t = \alpha x_t + (1 - \alpha) z_{t-1}$ lorsque $W_z = 1 - \alpha, W_x = \alpha$, et $W_y = 1$.

Il sera accepté qu'il est possible de généraliser le RNN du AR(p) au VAR(p) (*Vectorial Autoregressive*).

L'accent a ici été mis sur les modèles RNN simples pour des raisons de concision de notation mais il sera accepté que la preuve est généralisable aux RNN complexes.

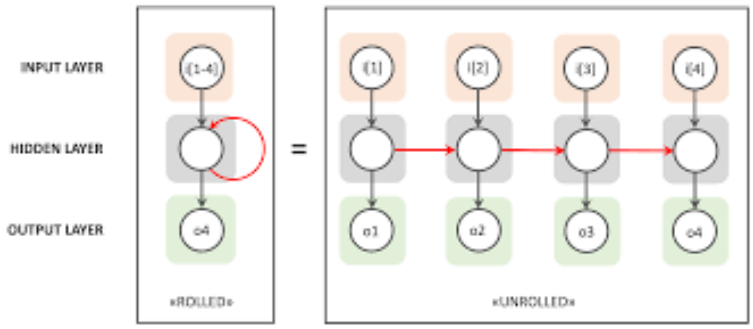


FIGURE V.1 – Schéma d'un réseau de neurones récurrent.

Annexe C

Choix de l'ordre du RNN, lien avec processus autorégressifs

L'objectif de cette annexe est de montrer, qu'à l'instar d'un modèle AR(p), le diagramme PACF permet d'estimer l'ordre du RNN. *Au vu de l'étude qui précède, il est possible, par exemple, de choisir d'expliquer la valeur de la série à l'instant t avec les 5 valeurs précédentes.* Cette annexe a également pour objectif de permettre de mieux comprendre les RNN.

La fonction d'autocovariance partielle apporte ici un éclairage supplémentaire. Un processus RNN(1) est d'abord considéré. L'autocovariance lag-1 est :

$$\tilde{\gamma}_1 = \mathbb{E}\left[y_t - \mu, y_{t-1} - \mu\right] = \mathbb{E}\left[\hat{y}_t + \epsilon_t - \mu, y_{t-1} - \mu\right],$$

et en utilisant le modèle RNN(1) avec, pour simplifier, un seul poids de récurrence, ϕ :

$$\hat{y}_t = \sigma\left(\phi y_{t-1}\right)$$

donne :

$$\tilde{\gamma}_1 = \mathbb{E}\left[\sigma\left(\phi y_{t-1}\right) + \epsilon_t - \mu, y_{t-1} - \mu\right] = \mathbb{E}\left[y_{t-1}\sigma\left(\phi y_{t-1}\right)\right],$$

où il a été supposé $\mu = 0$ dans la deuxième partie de l'expression. En vérifiant que la covariance $AR(1)$ est récupéré, $\sigma := Id$ est fixé de sorte à ce que :

$$\bar{\gamma}_1 = \phi\mathbb{E}\left[y_{t-1}^2\right] = \phi\mathbb{V}\left[y_{t-1}\right].$$

En continuant avec l'autocovariance lag-2, il vient :

$$\bar{\gamma}_2 = \mathbb{E}\left[y_t - P\left(y_t \mid y_{t-1}\right), y_{t-2} - P\left(y_{t-2} \mid y_{t-1}\right)\right]$$

et $P\left(y_t \mid y_{t-1}\right)$ est approximé par le RNN(1) :

$$\hat{y}_t = \sigma\left(\phi y_{t-1}\right).$$

En substituant $y_t = \hat{y}_t + \epsilon_t$ dans ce qui précède, il vient :

$$\tilde{\gamma}_2 = \mathbb{E}\left[\epsilon_t, y_{t-2} - P\left(y_{t-2} \mid y_{t-1}\right)\right].$$

L'approximation de $P\left(y_{t-2} \mid y_{t-1}\right)$ avec le RNN est alors :

$$\hat{y}_{t-2} = \sigma\left(\phi\left(\hat{y}_{t-1} + \epsilon_{t-1}\right)\right).$$

Il est possible d'observer que $\hat{y}_{t-2}$ dépend de ϵ_{t-1} mais pas de ϵ_t.
$y_{t-2} - P\left(y_{t-2} \mid y_{t-1}\right)$ dépend donc de $\{\epsilon_t - 1, \epsilon_t - 2, \ldots\}$. Il est alors obtenu que $\tilde{\gamma}_2 = 0$.

Comme contre-exemple, l'autocovariance partielle lag-2 du processus RNN(2) est considérée :

$$\hat{y}_{t-2} = \sigma\left(\phi\sigma\left(\phi\left(\hat{y}_t + \epsilon_t\right) + \epsilon_{t-1}\right)\right)$$

qui dépend de ϵ_t. Ainsi, l'autocovariance partielle lag-2 n'est pas nulle. Il est possible de montrer que l'autocorrélation partielle $\tilde{\tau}_s = 0, s > p$ et donc que, comme le processus $AR(p)$, la fonction d'autocorrélation partielle pour un RNN(p) a une valeur nulle après p décalages. La fonction d'autocorrélation partielle est indépendante du temps. Une telle propriété peut être utilisée pour identifier l'ordre du modèle RNN à partir du PACF estimé, exactement comme dans le cas du processus autorégressif.

Annexe D

Calculs du gradient et de la Hessienne de la log-vraisemblance ajustée

Sous certains *packages*, le gradient et la Hessienne doivent être spécifiés à la main.
Il est donc vu dans cette annexe comment les calculer et les implémenter dans ces *packages* :

La log-vraisemblance de la fonction de perte ajustée avec les poids w_0 et w_1 est ici :

$$L = w_0 y_i \cdot \log(p_i) + w_1 (1 - y_i) \cdot \log(1 - p_i)$$

A partir du dessus :

$$L = w_0 y_i \cdot \log\left(\frac{1}{1 + e^{-\hat{y}_i}}\right) + w_1 (1 - y_i) \cdot \log\left(\frac{e^{-\hat{y}_i}}{1 + e^{-\hat{y}_i}}\right)$$

La dérivée peut être écrite comme :

$$L' = \frac{e^{\hat{y}_i}(y_i - 1)\omega_1 + \omega_0 y_i}{1 + e^{\hat{y}_i}}$$

L' est donc pris comme gradient dans la nouvelle fonction *objective* qui est définie pour XGBoost. Il reste maintenant à préciser la nouvelle Hessienne.

Le calcul de la Hessienne demande, lui, plus de travail.
Soient m échantillons $\{x_i, y_i\}$ tels que $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$.
Il est rappelé que, dans la régression logistique binaire, p_i est lié à $\hat{y}_i$ par une fonction $\sigma(\hat{y}_i)$:

$$\sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}},$$

et il existe $\omega \in \mathbb{R}^d$ tel que $y_i = \omega^T x_i$. La fonction de perte est alors définie comme :

$$l(\omega) = \sum_{i=1}^{m} - \left(w_0 y_i \log \sigma(\hat{y}_i) + w_1 (1 - y_i) \log\left(1 - \sigma(\hat{y}_i)\right)\right)$$

Il faut remarquer que :

$$1 - \sigma(\hat{y}_i) = e^{-\hat{y}_i} / \left(1 + e^{-\hat{y}_i}\right) = 1 / \left(1 + e^{\hat{y}_i}\right) = \sigma(-\hat{y}_i)$$

Et que :

$$\frac{\partial}{\partial \hat{y}_i}\sigma(\hat{y}_i) = e^{-\hat{y}_i}\left(1 + e^{-\hat{y}_i}\right)^{-2} = \sigma(\hat{y}_i)(1 - \sigma(\hat{y}_i))$$

Le travail est réalisé directement avec les vecteurs. Le Hessien de la fonction de perte $l(\omega)$ est donné par $\vec{\nabla}^2 l(\omega)$. Il est rappelé que $\frac{\partial \hat{y}_i}{\omega} = \frac{x^T \omega}{\partial \omega} = x^T$ et que $\frac{\partial \hat{y}_i}{\omega^T} = \frac{\partial \omega^T x}{\partial \omega^T} = x$.

Soit $l_i(\omega) = -y_i w_0 \log \sigma(\hat{y}_i) - w_1 (1 - y_i) \log (1 - \sigma(\hat{y}_i))$.

D'après ce qui précède et la règle de la chaîne :

$$\frac{\partial \log \sigma(\hat{y}_i)}{\partial \omega^T} = \frac{1}{\sigma(\hat{y}_i)} \frac{\partial \sigma(\hat{y}_i)}{\partial \omega^T} = (1 - \sigma(\hat{y}_i)) x_i$$

Et de plus :

$$\frac{\partial \log(1 - \sigma(\hat{y}_i))}{\partial \omega^\top} = -\sigma(\hat{y}_i) x_i$$

Le résultat d'intérêt est finalement atteint :

$$\nabla \overrightarrow{l_i}(\omega) = x_i [\sigma(\hat{y}_i) [\omega_1 - y_i \omega_1 - y_i \omega_0] - y_i \omega_0]$$

Et le Hessien est finalement calculé comme :

$$\vec{\nabla}^2 l_i(\omega) = \frac{\partial l_i(\omega)}{\partial \omega \partial \omega^T} = x_i x_i^\top \sigma(\hat{y}_i) (1 - \sigma(\hat{y}_i)) (y_i (\omega_0 - \omega_1) + \omega_1)$$