Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 15/11/2021

Par : **Fengyue Zhan**

Titre : **Risque de subsidence, prédiction via méthodes de machine learning et l'Open Data**

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 : Caroline Hillairet

Membres présents du jury de l'Institut des Actuaires

Entreprise : Liberty Mutual Re

Signature :

Directeur du mémoire en entreprise :

Nom : Victor Bouton
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Ce mémoire d'actuariat réalisé au sein de l'équipe Agriculture & Parametrics de Liberty Mutual Re porte sur l'étude du risque de subsidence. Plus précisément, nous nous intéressons au phénomène de retrait-gonflement des argiles et aux conséquences de ce dernier sur les constructions. L'objectif de ce projet est de modéliser et prédire la charge IBNR totale nationale des sinistres de sécheresse d'une année dès lors que cette dernière s'achève, à l'aide des variables dérivant des données climatiques, des données géologiques et des données socio-économiques. La connaissance rapide de la charge IBNR permet notamment une meilleure gestion des provisions et une allocation plus efficace des capitaux propres pour les assureurs. Les variables climatiques, construites à partir des indices de sécheresse serviront à caractériser de manière quantitative la situation de sécheresse propre à chaque année. Les indices de sécheresse sollicités, basés sur des données de précipitation et de température, sont d'abord utilisés au cours d'une approche annuelle avec une résolution spatiale départementale. Cette approche a pour objectif d'introduire le sujet, son contexte et les données. Lors de cette dernière, nous explorons les données à l'aide des outils d'analyse en composantes principales et des modèles basés sur la théorie des modèles linéaires généralisés. Cette première approche montre rapidement ses limites en raison de la résolution de la maille utilisée qui réduit drastiquement le nombre d'observations. Pour répondre à cela, nous affinons l'approche pour passer d'une approche départementale à une approche communale. Cette dernière, moyennant une modélisation en deux étapes dite occurrence-fréquence et à l'aide des méthodes de machine learning, nous a notamment permis de développer toute une méthodologie autour de la sélection des variables explicatives et la construction des modèles. Les résultats de cette approche avec l'introduction d'une fonction déterministe à la sortie du modèle occurrence-fréquence sont très prometteurs. Néanmoins, ces résultats doivent être complétés en raison de la publication des données d'indice SWI Uniforme ou Uniforme Soil Moisture Index par Météo-France en avril 2021. Les nouvelles données nous invitent à mettre à jour le modèle d'occurrence mais surtout à développer un modèle saisonnier avec chaque modèle sous-jacent décrivant une saison particulière. Ce modèle saisonnier, réalisé à l'aide de la théorie des modèles inflatés, produit des résultats satisfaisants tout en mettant en évidence un manque de diversité dans les données annuelles et un manque de données pour certaines saisons. Ces limites, parfois inhérentes à la mise en application récente des nouveaux critères de reconnaissance, peuvent néanmoins être rapidement surmontées dans les années à venir.

Mots clefs : *Subsidence, charge IBNR, arrêtés CatNat, indice de sécheresse, SWI Uniforme, modèle zéro inflaté.*

Abstract

The internship completed within the Agriculture & Parametrics team of Liberty Mutual Re studies the risk of subsidence, a geological phenomenon. More precisely, we are interested in the shrinking and swelling of clays and its consequences over buildings. The aim of this project is to model and predict the national IBNR amount of claims at the end of a given year, with variables extracted from climate data, geological data and socioeconomic data. The quick access of IBNR amounts of claims allows a better management of provisions and a more efficient equity allocation for insurers. The climate variables, derived from classic drought indices, aim to characterize quantitatively the drought situation of each year and the geological and socioeconomic variables will help us measure the severity of drought in terms of claims. The drought indices, based on precipitation and temperature data, are first used in a regional approach which aims to introduce the subject, its context and the data. With this approach, we explored the data using principal components analysis and models based on the generalized linear model framework. This first approach shows its limits mainly due to the regional resolution which drastically reduces the number of observations. To override these limits, we used a local approach with which, thanks to a two-step modelling process and machine learning techniques, we developed a whole methodology about feature selection and model development. The results obtained from this approach, by applying a deterministic function to the outcome of the two-step model, were really promising. Nonetheless, they were also incomplete due to the data of Uniform SWI or Uniforme Soil Moisture Index published by Météo-France in April 2021. These new data invite us to update a part of the two-step model but also to develop a seasonal model which combines four underlying models, one for each season. The seasonal model was developed using the inflated model framework and its results were satisfying despite the lack of diversity in the annual data and the lack of data for some seasons. Nonetheless, these lacks, inherent to the application of newly defined recognition criteria of natural disaster in 2018, may be overridden in the near future.

Key words : *Subsidence, IBNR claims, natural disaster declaration, drought index, Uniform SWI, zero inflated model.*

Remerciements

Je voudrais ici remercier toutes les personnes qui ont contribué à la réalisation de ce mémoire.

Je tiens d'abord à remercier mon maître de stage, M. Victor Bouton, data scientist au sein de l'équipe Agriculture et Paramétrique chez Liberty Mutual Re. Au cours de ce projet, son encadrement bienveillant et le partage de ses connaissances ont été essentiels et indispensables à la réalisation de ce mémoire.

Je voudrais remercier Mme. Caroline Hillairet, chercheuse au CREST (Centre de recherche en économie et statistique) et référente pédagogique de ce projet. La mise à disposition de son expertise et ses remarques éclairantes ont grandement contribué à ce mémoire.

Je remercie vivement mes collègues de bureau, et notamment Eve Dartigues, Abdessamad Elangoudi et Yunhan Ma. Leur accueil chaleureux et leur soutien ont été une véritable force motrice au quotidien.

Je souhaite aussi remercier tous les autres membres de l'équipe Agriculture et Paramétrique et son directeur M. Jean-Christophe Garaix qui m'a fait confiance.

J'aimerai, in fine, exprimer mes remerciements pour le soutien administratif apporté par l'ENSAE tout au long de ce projet.

Table des matières

1 Introduction générale

Le sujet du mémoire porte sur l'étude du phénomène de subsidence et plus particulièrement sur le risque de gonflement et retrait des argiles, ainsi que les dommages causés par ce dernier aux constructions, dans le cadre d'un portefeuille représentant l'exposition et la sinistralité en France des risques de ce type.

En France, l'établissement public de référence pour gérer les ressources et les risques du sol et du sous-sol est le Bureau de Recherches Géologiques et Minières (BRGM). Il s'agit d'un service géologique national créé en 1959 et placé sous la tutelle des ministères chargés de la Recherche, de l'Écologie et de l'Économie. Ce bureau est chargé de réaliser une cartographie des zones françaises exposées au phénomène de retrait-gonflement des argiles ([41]) afin de permettre l'application des dispositions réglementaires introduites par l'article 68 de la loi ELAN qui porte sur la construction et l'aménagement des habitations. Cette carte, publiée en 2019 ([3]) et utilisant la base des Sinistres Indemnisés Liés aux Évènements Climatiques (SILECC) de la Mission Risques Naturels (MRN) dont les données représentent environ 70% du marché de l'assurance, fournit une vision globale du risque de subsidence à l'échelle nationale en montrant notamment que 48% du territoire français est en zone d'exposition moyenne ou forte.

En aval de la BRGM, le risque de subsidence, et plus généralement les risques de catastrophe naturelle sont couverts par la plupart des assurances et entre dans le champ d'application du régime de catastrophe naturelle porté par la Caisse Centrale de Réassurance (CCR). La CCR est une entreprise française de réassurance créée en 1946 et dont l'une des missions consiste à réassurer les assureurs couvrant le risque de catastrophe naturelle. Le déclenchement de cette couverture est conditionné à la reconnaissance d'un état de catastrophe naturelle par le ministère de l'intérieur. Cette décision du ministère est basée sur l'avis formulé par une commission interministérielle d'experts qui examine chaque dossier communal individuellement et utilise notamment différents critères de sécheresse dans leur examen.

Cette étude s'inscrit dans cette perspective et tente d'expliquer les différents critères à l'aide des indices de sécheresse pour in fine estimer la charge IBNR (*Incurred But Not Reported*) totale de sinistre en France pour une année quelconque dès lors que cette dernière se termine. Il faut noter que l'approche est loin d'être novatrice. En effet, dès le début du XXème siècle, nous avons l'émergence de cette idée de caractériser le phénomène de sécheresse à l'aide des indices avec par exemple De Martonne (1925, [18]) et son livre intitulé *Traité de Géographie Physique*. Cette idée a continué à germer au milieu du XXème siècle avec notamment les travaux de Palmer (1965, [10]). Au début des années 90, McKee et al. publient leurs résultats portant sur l'indice SPI ou Standard Precipitation Index (1993, [11] ; 1997, [12]), pionniers dans le domaine. Suite à cela, de nombreux indices de sécheresse ont émergé avec par exemple Byun et Wilhite (1993, [9]) qui ont étudié une partie des Grandes Plaines des États-Unis entre 1960 et 1996, Wu et al. (2001, [7]) et Morid et al. (2006, [8]) qui ont comparé la performance de différents indices en Chine et en Iran, Narasimhan et Srinivasan (2005, [21]) qui ont analysé plus particulièrement la gestion de la sécheresse en agriculture, Lyon et Barnston (2005, [22]) et l'étude des précipitations lors du phénomène El Niño - Oscillation Australe en région tropicale ou encore Balthas (2007, [17]) et Tsakiris et al. (2007, [23]) qui ont étudié le climat et la gestion de la sécheresse en Grèce. Après les années 2010, la problématique de l'environnement semble commencer à intéresser de plus en plus la communauté et nous commençons à voir apparaître des indices de sécheresse plus complexes qui tentent de capturer les tendances à long terme, par exemple les travaux de Begueria et al. (2010, [14] ; 2014, [15]) et de Ali et al. (2017, [6]). Dans les dernières années, nous avons de nombreux travaux qui approchent la sécheresse dans sa globalité et tentent de chiffrer les dommages causés par ce dernier, par exemple le rapport de la MRN (2018, [1]), les mémoires d'actuariat de Schulte (2016, [2]) et d'Arnaud (2016, [4]), ou encore les travaux de Météo-France avec le projet CLIMSEC ([40]) débuté en 2011.

En 2019, le gouvernement a réformé les critères de reconnaissance de catastrophe naturelle ([39]) et en 2021, Météo-France a rendu publiques les données servant de base dans le calcul des différents critères de reconnaissance ([38]). Tout ceci laisse à penser qu'il est désormais envisageable de modéliser le risque de subsidence en France et les dommages causés par ce dernier dans le cadre du régime de

catastrophe naturelle avec un niveau de confiance élevé. Le projet tente donc de relever ce défi dans le cadre d'un portefeuille France avec comme exigence principale l'obtention d'un modèle simple et facilement contractualisable qui soit capable de prédire la charge IBNR d'une année dès que cette dernière se termine. Il est intéressant de noter qu'à ce jour, des études en parallèle de la présente menées par Charpentier et al. (2021, [42]) donnent déjà des résultats encourageants. Dans le papier, les auteurs réalisent une approche annuelle et à l'échelle communale en comparant notamment des modèles basés sur des arbres et sur des régressions pour évaluer le coût économique du risque de subsidence de 2002 à 2019. Les indices de sécheresse utilisés par les auteurs sont similaires à ceux que nous sollicitons mais en plus du maximum et du minimum des indices sur une année comme variables, nous utilisons aussi des fonctions basées sur la somme pour tenir compte davantage de la cinétique lente du phénomène de subsidence. Par ailleurs, nous insistons dans ce projet sur l'aspect évolutif des critères de reconnaissance. Cet aspect évolutif, qui rend l'étude historique du phénomène de subsidence difficile, a un réel impact en termes de prédiction car les nouveaux critères de reconnaissance appliqués depuis 2018 nous invitent à une modélisation non plus annuelle mais saisonnière. Cette modélisation saisonnière, évoquée par Charpentier et al. dans leur conclusion, nous permet in fine de proposer une modélisation plus fine que ce qui existe déjà.

Pour cela, le travail s'organise autour de plusieurs axes principaux, à savoir : construction des bases de données, analyses de la cohérence de ces dernières, établissement des méthodologies et applications de ces dernières dans la construction des modèles. Pour parcourir ces axes dans leur globalité, nous divisons le rapport en quatre parties. Dans une première partie, nous commençons par présenter le contexte, le risque de subsidence, les données et les indices de sécheresse sollicités. Dans une seconde partie, nous réaliserons une première approche sommaire annuelle et à l'échelle départementale. Nous verrons que l'exploration des données à travers d'abord un modèle de base qui est le modèle linéaire généralisé (GLM), puis des modèles toujours de régression mais améliorés à partir de cette base GLM, montre que l'approche départementale est trop restrictive en raison de la résolution de la maille départementale. Pour répondre à cela, une approche toujours annuelle mais à l'échelle communale est construite dans une troisième partie. Cette approche est construite en intégrant les données du Journal Officiel portant sur l'obtention ou non des décrets de catastrophe naturelle par les communes et aboutit à un modèle dit d'occurrence-fréquence qui comprend deux modèles sous-jacents. Au cours de cette approche communale, nous développons aussi toute une méthodologie autour de la sélection des variables et construction des modèles basée sur des méthodes de machine learning. Cette dernière est notamment applicable de manière générale dans d'autres problématiques de modélisation nécessitant une étape de reconnaissance du risque. Les résultats du modèle sont très prometteurs mais se trouvent être incomplets en raison de la publication de nouvelles données par Météo-France portant sur le SWI Uniforme. Ainsi, dans la quatrième partie, nous intégrons ces nouvelles données dans l'étude car il s'agit des données utilisées par la commission interministérielle et à partir desquelles les critères de reconnaissance sont calculés exactement de manière déterministe. À l'aide des critères calculés, nous mettons à jour le modèle occurrence-fréquence mais surtout, nous construisons un modèle saisonnier, composé de quatre modèles sous-jacents et basé sur la théorie des modèles inflatés, plus adapté aux nouveaux critères de reconnaissance appliqués depuis 2018. Ce modèle, ou plutôt cette dernière approche saisonnière à l'échelle communale, conduit à des résultats satisfaisants tout en mettant en évidence un manque de diversité dans les données annuelles et un manque de données pour certaines saisons.

2 Données climatiques et indices de sécheresse

2.1 Introduction

Dans cette première partie, nous allons présenter le cadre et l'intérêt du travail ainsi que les données et les indices de sécheresse. Parmi les données, nous décrirons notamment les données climatiques car les indices de sécheresse construits à partir de ces données nous permettront de calculer par la suite les variables climatiques. Ces variables, différentes en fonction de l'approche envisagée, constituent le cœur de l'étude. Les autres données sollicitées dans l'étude sont moins importantes et seront introduites brièvement dans cette partie. Des introductions plus détaillées de ces dernières seront données au fur et à mesure de l'étude au moment approprié.

Ainsi, nous allons montrer dans une première section l'intérêt de cette étude d'un point de vue actuariel. Dans une seconde section, nous décrirons le risque de subsidence et dans une troisième section, les données non climatiques. Nous présenterons les données climatiques utilisées dans cette étude dans une quatrième section et dans la dernière section, nous définirons les types de sécheresse et les indices de sécheresse sollicités.

2.2 Objectif actuariel de l'étude

L'objectif de cette étude est de prédire la charge IBNR totale annuelle des sinistres de sécheresse pour la France dans le cadre du régime de catastrophe naturelle soutenu par la CCR. Plus précisément, nous souhaitons construire un modèle qui soit capable de prédire, pour une année n donnée, la charge totale de l'année n dès lors que cette dernière se termine, i.e au début de l'année $n+1$. L'intérêt actuariel de cette étude réside donc dans la rapidité du modèle à produire les prédictions. En effet, il faut noter qu'en moyenne, la CCR prend 6 mois pour produire une première estimation, sujette à révision, de la charge IBNR annuelle et que pour les assureurs, le décalage est plutôt de l'ordre d'un an. Cette étude vise donc à réduire ce décalage temporel afin de permettre un contrôle plus fin des provisions.

Pour répondre à cet objectif, nous sollicitons les données ouvertes ou open data. Les données ouvertes sont des données numériques dont l'accès et l'usage sont laissés libres aux usagers, qui peuvent être d'origine privée ou publique. Elles sont produites notamment par des collectivités ou des établissements publics et sont diffusées de manière structurée selon une méthode et une licence ouverte garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière. Dans le cadre de cette étude, ces dernières sont souvent satellitaires ou peu évolutives. Autrement dit, ce sont des données en temps réel ou des données qui sont quasiment indépendantes du temps, ce qui rend la construction d'un modèle minimisant le décalage temporel possible.

2.3 Risque de subsidence

La subsidence ou sécheresse géotechnique est un phénomène de diminution du volume du sol. Cette diminution est provoquée par un déséquilibre entre flux d'eau entrant et sortant dans le volume considéré avec l'extérieur. En fonction de l'épaisseur étudiée, les sources des flux entrants et sortants du volume sont différentes, mais dès lors qu'il y a des fondations construites sur la surface du volume, des gonflements ou tassements différentiels du sol issus de ce déséquilibre peuvent endommager les constructions. Et les dommages sont d'autant plus importants si les constructions ne sont pas élaborées pour tenir compte de ce phénomène.

Dans le cadre de cette étude, nous considérons uniquement les précipitations comme sources de flux d'eau entrant et le phénomène d'évapotranspiration, i.e le transfert de l'eau vers l'atmosphère depuis le sol par évaporation au niveau du sol et transpiration des plantes, comme responsable du flux d'eau sortant. Cette hypothèse est importante car par la suite, nous allons tenter de capturer ces deux phénomènes à travers des indices de sécheresse afin de modéliser le phénomène de subsidence. Par ailleurs, il faut aussi noter qu'en fonction du type de sol, à savoir argileux ou pas, i.e capable de retenir une grande quantité d'eau ou pas, les amplitudes des mouvements différentiels peuvent être plus ou moins

importantes. Il est donc aussi important de tenir compte de ce facteur dans les modèles. La figure 1 représente une illustration du phénomène de subsidence et des différents facteurs mis en jeu.

Il faut préciser aussi que les précipitations et le phénomène d'évapotranspiration dépendent grandement des saisons, et dans la mesure où la France est située dans une zone géographique où le climat est tempéré avec plutôt une abondance en eau, les sinistres liés à ce phénomène de mouvements différentiels des terrains ont tendance à avoir lieu en période de sécheresse i.e plutôt autour de l'été. Et bien entendu, il existe d'autres facteurs d'influence que nous ne prenons pas en compte comme la fonte des glaces dans les montagnes notamment si les constructions sont à proximité, l'utilisation de la nappe phréatique pour l'agriculture, la neige en hiver, mais aussi des actions anthropiques comme des travaux d'aménagement qui peuvent perturber la répartition des écoulements des eaux et les conditions d'évaporation.

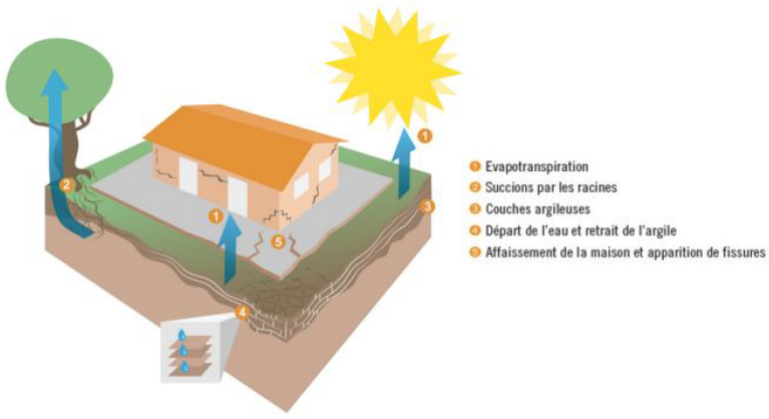

Figure 1 – Illustration du phénomène de subsidence, figure extraite de [2].

2.4 Données non climatiques

Pour étudier le phénomène, nous avons à disposition un portefeuille représentant l'exposition et la sinistralité du marché français pour le risque de subsidence et couvrant la période 2009-2018. Nous exploitons les données de ce portefeuille à la maille départementale pour les années 2009 à 2018 dans une première approche sommaire départementale. Ces dernières sont exploitées à la maille communale dans une seconde approche pour les années 2016 à 2018 pour l'apprentissage et 2019 à 2020 pour la prédiction. Ces données permettent d'avoir une estimation, par département ou par commune, du nombre de contrats sinistrés et du montant total de ces sinistres. Notez que dans cette étude, uniquement les contrats Multirisques Habitation sont considérés.

Nous avons accès ensuite aux données publiques de typologie ([3]), i.e le caractère argileux du sol et donc la capacité de retenir de l'eau de ce dernier. Ces données répartissent le territoire français en quatre niveaux de risque croissant allant du niveau 0 à 3. À partir de ces informations, il est notamment possible de construire des variables qui caractérisent la prédisposition naturelle du sol au risque de subsidence.

Nous avons aussi accès aux données de demandes de valeurs foncières (DVF) produites par la Direction générale des finances publiques qui recense les transactions immobilières en France, aux données de l'INSEE portant sur le recensement de la population et des logements, aux données d'indices d'humidités du sol (SWI) satellitaires, et aux données d'indices d'humidités du sol uniforme (SWI Uniforme) de Météo-France.

Nous reviendrons plus longuement sur toutes ces données aux moments appropriés au cours de l'étude. Pour le moment, il faut surtout retenir que toutes ces données, mis à part les données de portefeuille, sont disponibles librement et sont soit en temps réel ou mises à jour fréquemment, soit ne

dépendent pas ou très peu du temps. Cette caractéristique est importante car nous rappelons que nous souhaitons avoir un modèle qui puisse prédire la charge IBNR annuelle totale dès que l'année s'achève.

2.5 Données climatiques

Avec l'aide d'un fournisseur de données, nous avons accès aux données de climat. Plus exactement, ces données de climats sont extraites des données ERA5, une base de données de réanalyse météorologique basées sur des données satellitaires et produites par le Centre européen pour les prévisions météorologiques à moyen terme (CEPMMT/ECMWF). Cette base ERA5 couvre l'ensemble de la Terre avec une résolution spatiale de 0.25°, soit environ 30 km en France métropolitaine, et remonte jusqu'à 1950 avec une résolution horaire.

Pour notre étude, ces données de réanalyse sont récoltées à l'aide d'une grille couvrant tout le territoire français avec un pas de 0.25° en latitude et en longitude comme le montre la figure 2.

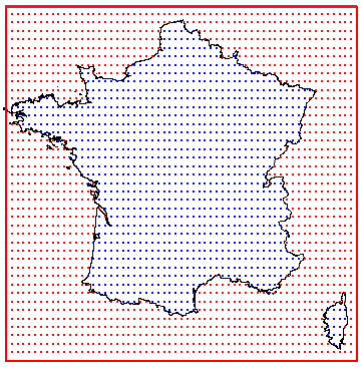

Figure 2 – Schéma montrant la grille couvrant le territoire français métropolitain. Le pas est de 0.25° en longitude et en latitude. Sur chaque point bleu, nous avons des séries temporelles de température et de précipitation. Sur les points rouges, nous n'avons pas de donnée car non nécessaire à l'étude.

Notez que pour chaque point de la grille, quatre séries temporelles horaires représentant la température maximale, la température minimale, la température moyenne comme la moyenne des deux températures extrêmes, et la précipitation accumulée sont récoltées. Par la suite, les indices de sécheresses, et donc les variables explicatives climatiques de l'étude, seront construits à l'aide de ces séries temporelles.

Avant de passer aux indices de sécheresses, nous allons analyser ces données climatiques. Plus précisément, nous allons comparer les données de réanalyse, qui sont des données modélisées et non mesurées, aux données récupérées auprès des stations météo. L'idée est que nous souhaitons savoir si les données de réanalyse sont de bonne qualité ou pas en les comparant à la vérité terrain. Pour cela, les séries horaires sont d'abord agrégées en séries journalières. Nous représentons dans la figure 3 la température moyenne mesurée par les stations météo en fonction de la température moyenne modélisée pour les quatre stations dont les distances aux points de la grille sont minimales. Autrement dit, nous avons une liste de station météo et chaque station est associée à une distance qui est la distance entre cette station et le point de la grille la plus proche. Nous choisissons ensuite les quatre stations dont les distances associées sont les plus petites pour obtenir les nuages de points de la figure.

Notez dans cette figure que les températures moyennes journalières observées depuis les stations météo sont proches de celles modélisées, ce qui montre que les données de température moyenne sont plutôt de bonne qualité. Cette même comparaison est faite pour la température minimale journalière et la température maximale journalière et la conclusion est la même (annexe 9.4). Ainsi, les données de température sont plutôt de bonne qualité.

Cette même comparaison est ensuite faite pour les données de précipitation journalière dans la figure 4. D'après les nuages de points, les données de précipitation journalière modélisées sont clairement éloignées de celles observées depuis les stations météo. Par conséquent, les données de précipitation

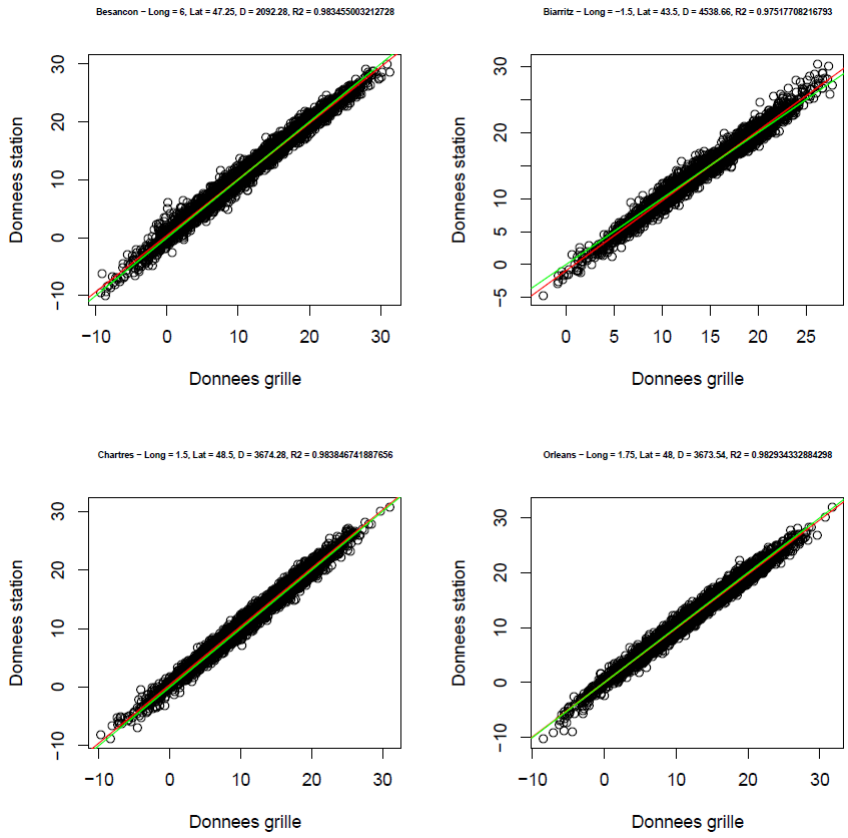

Figure 3 – La température moyenne journalière mesurée par la station en fonction de la température moyenne journalière modélisée. La droite verte correspond à la droite $y = x$ et la droite rouge correspond à la droite de régression. La distance entre la station et le point de la grille la plus proche est indiquée dans le titre en mètre accompagnée du nom de la station, de la position géographique de cette dernière et du R^2 de la régression linéaire. D'après ces graphes, les données de température moyenne sont plutôt de bonne qualité.

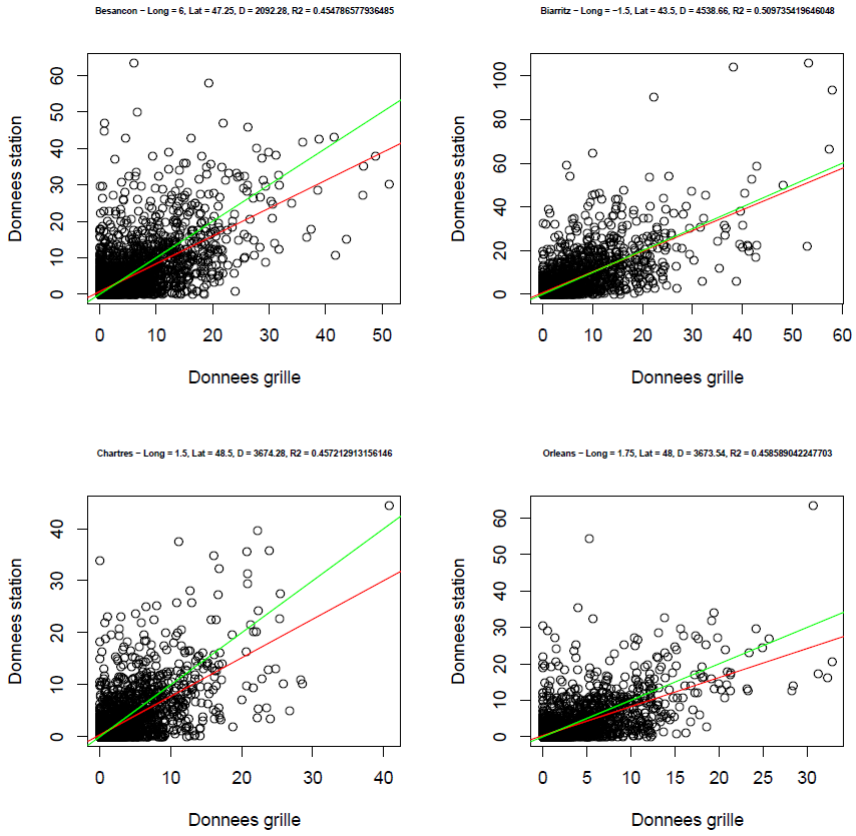

Figure 4 – La précipitation journalière mesurée depuis la station en fonction de la précipitation journalière modélisée. La droite verte correspond à la droite $y = x$ et la droite rouge correspond à la droite de régression. La distance entre la station et le point de la grille la plus proche est indiquée dans le titre en mètre accompagnée du nom de la station, de la position géographique de cette dernière et du R^2 de la régression linéaire. D'après ces graphes, les données de précipitation journalière sont plutôt de moins bonne qualité.

journalière sont plutôt de moins bonne qualité et l'utilisation de ces données modélisées introduit à priori des biais de modélisation.

À présent, au lieu d'agréger les séries horaires en séries journalières, nous les agrégeons en séries mensuelles. Il va sans dire que les données de température mensuelles modélisées restent de bonne qualité. La question est qu'en est-il des données de précipitation. Les mêmes graphes que les précédents sont représentés dans la figure 5 avec cette fois non pas des précipitations journalières mais mensuelles. Les graphes montrent clairement qu'à l'échelle mensuelle, les données de précipitation modélisées sont plutôt de bonne qualité et plus proche de la vérité terrain.

Nous reviendrons plus tard sur l'intérêt de ces comparaisons dans une prochaine partie. Pour le moment, nous allons continuer nos définitions en passant cette fois aux indices de sécheresse.

2.6 Classification et caractérisation de la sécheresse

Nous allons à présent définir les types de sécheresse et les indices de sécheresse sollicités durant cette étude. Il faut préciser d'abord que jusqu'à ce jour, il n'y a pas encore d'indice universel pour le phénomène de sécheresse. De manière générale, la gestion en eau par les autorités publiques passe souvent par l'utilisation d'indices locaux plus ou moins quantitatives combinées avec quelques indices communément recommandés par la communauté scientifique et le WMO ou Organisation météorologique mondiale ([5], [19]).

Ainsi, dans cette section, nous définissons d'abord les types de sécheresse communément acceptés

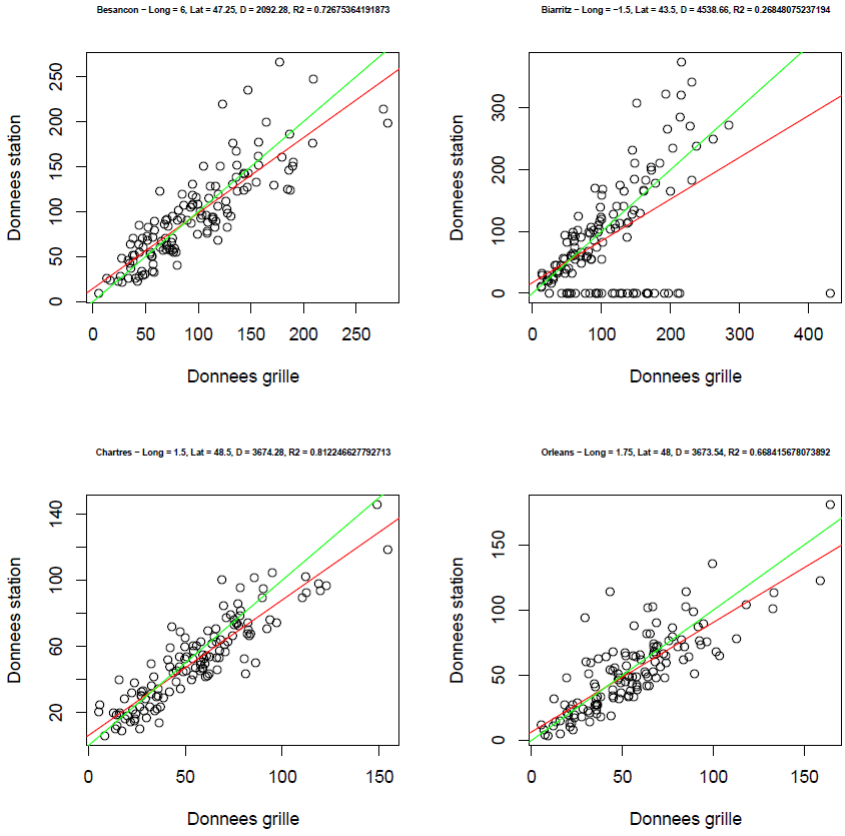

Figure 5 – La précipitation mensuelle mesurée depuis la station en fonction de la précipitation modélisée. La droite verte correspond à la droite $y = x$ et la droite rouge correspond à la droite de régression. La distance entre la station et le point de la grille la plus proche est indiquée dans le titre en mètre accompagnée du nom de la station, de la position géographique de cette dernière et du R^2 de la régression linéaire. Notez qu'il y a des données manquantes d'où certains points sur l'axe $y = 0$ pour la station de Biarritz. D'après ces graphes, les données de précipitations mensuelles sont plutôt de bonne qualité comparées aux mêmes données mais journalières. Il y a donc un réel intérêt d'utiliser l'échelle mensuelle par rapport à l'échelle journalière dans le calcul des indices de sécheresse.

dans un premier temps puis ensuite cinq indices de sécheresse communément utilisés dans un deuxième temps. Les cinq indices sont : le Standardized Precipitation Index (SPI), le Standardized Precipitation Evapotranspiration Index (SPEI), le Standardized Precipitation Temperature Index (SPTI), le China-Z Index (CZI), et l'Effective Drought Index (EDI). Notez que d'autres indices existent et pourraient être utilisés dans cette étude mais en raison des données exigées, ces derniers ne sont pas calculés. Néanmoins, des définitions rapides de ces derniers sont fournies en annexe 9.1. Les calculs des variables climatiques seront à partir de ces indices de sécheresses seront plus amplement détaillés par la suite.

2.6.1 Type de sécheresse

Nous commençons à présent par définir la classification standard des différents types de sécheresse, à savoir la sécheresse météorologique, la sécheresse agricole et la sécheresse hydrologique. Ces définitions sont intéressantes car à l'aide de ces dernières, nous pouvons faire le lien entre types de sécheresse et indices de sécheresse.

La sécheresse météorologique est surtout caractérisée par des conditions météorologiques et ne dépend pas des facteurs du sol. Elle survient donc lorsqu'il existe une période prolongée d'un taux de précipitation en dessous de la moyenne. Les signaux d'une sécheresse de ce type peuvent notamment être capturé par des indices de sécheresse calculés sur des fenêtres temporelles courtes, par exemple SPI 1 mois ou SPI 3 mois.

La sécheresse agricole correspond aux périodes où l'humidité des sols est trop faible pour les cultures. Ce type de sécheresse est causé par de nombreux facteurs divers et variés allant du manque de précipitation à une utilisation excessive des nappes phréatiques. Il se manifeste surtout par un appauvrissement de la végétation en raison du manque d'eau nécessaire à la croissance des plantes. Pour capturer les signaux d'une sécheresse agricole, il existe notamment des indices basés sur des images satellites comme le NDVI (Normalized Difference Vegetation Index, [19]) qui sont créés à cet effet.

La sécheresse hydrologique survient lorsque le débit des rivières, le niveau des réserves d'eau disponibles dans les nappes aquifères, lacs et réservoirs sont anormalement bas par rapport à une situation moyenne calculée sur le long terme. Il faut noter que ce type de sécheresse peut apparaître même en situation de précipitation normale ou au-dessus de la moyenne. En effet, il suffirait qu'il y ait une surexploitation des réservoirs d'eau ou une modification des conditions physiques d'alimentation des nappes pour engendrer une sécheresse hydrologique. Ainsi, contrairement aux sécheresses météorologiques ou agricoles, les causes des sécheresses hydrologiques peuvent être de nature anthropique de manière directe. Parmi les indices pour mesurer ce type de sécheresse, nous pouvons par exemple citer le SSFI (Standardized Streamflow Index, [19]) ou le SWSI (Surface Water Supply Index, [19])

Dans le perspective de ces différents types de sécheresse, nous estimons que pour cette étude, le risque de subsidence est probablement plus lié aux sécheresses météorologiques et agricoles qu'aux sécheresses hydrologiques. En effet, les premières affectent directement les couches surfaciques du sol tandis que les secondes concernent davantage les couches profondes. De plus, le risque de subsidence au sein d'un portefeuille d'assurance est un risque certes qui s'inscrit dans la durée, mais reste relié à un phénomène avec une cinétique de l'ordre de quelques mois voire un an au plus. Cette durée typique correspond davantage aux sécheresses météorologiques et agricoles. La sécheresse hydrologique, quant à elle, est plutôt un phénomène étalé sur plusieurs années.

2.6.2 Indice de sécheresse

2.6.2.1 SPI : Standardized Precipitation Index

Le SPI introduit par McKee et al. en 1993 ([11], [12], [13]) est l'indice de sécheresse le plus communément utilisé à nos jours et est recommandé par les institutions publiques pour la surveillance des sécheresses en complément des indices spécifiques locaux ([5], [19]).

Considérons P_{acc}^{T} la précipitation mensuelle accumulée sur la période T. Considérons aussi la distribution de P_{acc}^{T} pour une base de données de taille au moins 30 ans. Fittons alors cette distribution par une loi gamma à deux paramètres (α et β) par maximum de vraisemblance et avec la fonction

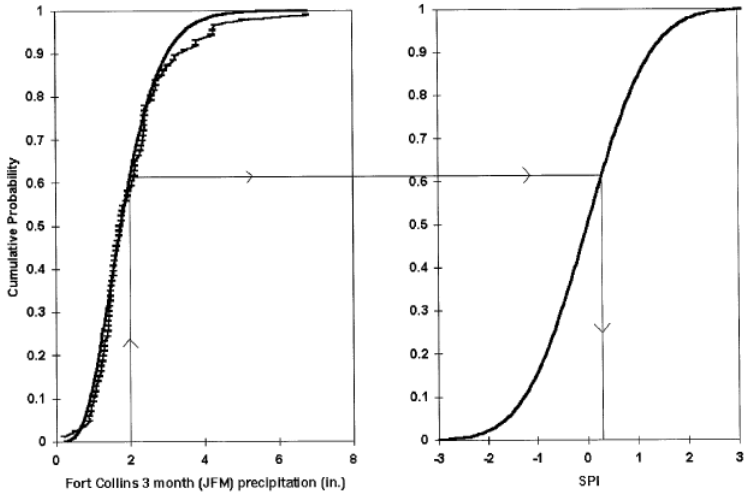

Figure 6 – Courbes illustrant le calcul de l'indice SPI (extraite de [12]). À gauche, la fonction de répartition $G^T_{\alpha,\beta}$ et à droite la fonction de répartition F. Le $SPI(x)$ se lit sur l'abscisse de la courbe à droite pour une observation de précipitation accumulée x sur l'abscisse de la courbe de gauche.

gamma résultante, calculons sa fonction de répartitions $G^T_{\alpha,\beta}$. Ainsi, pour toute nouvelle observation de précipitation accumulée x, en notant F la fonction de répartition d'une loi normale centrée réduite, l'indice SPI est défini par :

$$SPI(x) = F^{-1}(G^T_{\alpha,\beta}(x)) \approx s\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) \tag{1}$$

Avec :

$$s = \begin{cases} -1 & \text{si } 0 < G^T_{\alpha,\beta}(x) < 0.5 \\ 1 & \text{sinon} \end{cases} \quad \text{et} \quad t = \begin{cases} \sqrt{-2\ln(G^T_{\alpha,\beta}(x))} & \text{si } 0 < G^T_{\alpha,\beta}(x) < 0.5 \\ \sqrt{-2\ln(1 - G^T_{\alpha,\beta}(x))} & \text{sinon} \end{cases} \tag{2}$$

Et :

$$\begin{cases} c_0 = 2.515517 \\ c_1 = 0.802853 \\ c_2 = 0.010328 \\ d_1 = 1.432788 \\ d_2 = 0.189269 \\ d_3 = 0.001308 \end{cases} \tag{3}$$

Notez qu'ici, les coefficients c_i et d_i sont des coefficients d'approximation utilisés par les auteurs de l'indice SPI ([12]). La figure 6 extraite du papier des auteurs illustre le calcul de l'indice SPI.

Les auteurs précisent aussi que cette indice doit être utilisé avec précaution lorsqu'il y a une forte probabilité que $P^T_{acc} = 0$ dans la base de donnée soit parce que la zone d'étude est une zone aride, soit parce qu'il y a pas les 30 ans de données mensuelles minimum recommandées par les auteurs. De plus, il faut noter que pour des tailles de fenêtre T différentes (1 mois, 3 mois, etc), le type de sécheresse (météorologique, hydrologique ou agricole) capturé est aussi différent.

2.6.2.2 SPEI : Standardized Precipitation-Evapotranspiration Index

L'indice SPI défini précédemment ne prend en compte que la précipitation, ce qui peut être insuffisant pour apprécier la sécheresse dans les conditions de réchauffement climatique d'aujourd'hui. En effet, comme expliqué lors de la présentation du risque de subsidence, le phénomène d'évapotranspiration joue aussi un rôle dans la balance en eau du sol, et ceci est d'autant plus important quand la température est élevée. L'indice SPEI introduit vers 2010 par Begueria et al. ([14], [15]) tente donc de résoudre ce problème.

La méthode de calcul pour le SPEI est identique à celle du SPI, sauf qu'au lieu de faire le fit sur P_{acc}^T, ce qui est le cas pour le SPI, nous allons fitter la distribution de D_{acc}^T dont l'expression est donnée par :

$$D_{acc}^T = P_{acc}^T - PET \tag{4}$$

Avec PET le potentiel d'évapotranspiration de référence défini via l'équation de Hargreaves ([16]) :

$$PET = 0.0023 \times RA \times \left(\left(\frac{Q_{mx} + Q_{mi}}{2} \right) + 17.8 \right) \times (Q_{mx} - Q_{mi})^{0.5} \tag{5}$$

Où Q_{mx} est la température maximale moyenne (en $°C$), Q_{mi} la température minimale moyenne (en $°C$), et RA la radiation extraterrestre qui dépend de la latitude (cf. [16] pour les valeurs). Ainsi, par rapport au calcul de SPI, en plus des données de précipitation, les séries temporelles de température journalière maximale et minimale ainsi que les latitudes des points seront aussi nécessaires dans le calcul du SPEI.

Notez que pour estimer le potentiel d'évapotranspiration, il existe bien entendu d'autres équations. En somme, les auteurs du SPEI recommandent d'utiliser l'équation de Penman–Monteith en premier lieu, mais ce dernier nécessitant des données plus complexes (radiation solaire, humidité, vitesse du vent, etc), l'équation d'Hargreaves est utilisée en second lieu. Et l'équation de Thornthwaite en dernier lieu si seules les données de température moyenne sont disponibles.

Notez aussi que comme pour le SPI, le choix de la fenêtre T est libre et le type de sécheresse capturé est différent en fonction. Par ailleurs, les auteurs recommandent d'utiliser une distribution log-logistique à trois paramètres comme fonction de fit au lieu de la loi gamma, notamment parce que la première donne des résultats plus cohérents pour les événements situés dans la queue de distribution. Les auteurs pensent aussi que dans certains cas, pour calculer P_{acc}^T, il est nécessaire de considérer non pas une fenêtre rectangulaire mais triangulaire ou gaussien.

2.6.2.3 SPTI : Standardized Precipitation temperature Index

L'indice SPTI est introduit par Ali et al en 2017 ([6]). La définition de cet indice est basée sur celle de l'indice de De Martonne et la méthode de calcul est la même que pour le SPI. L'indice de De Martonne est défini par :

$$I_{DM} = \frac{12P}{Q + 10} \tag{6}$$

Avec P la précipitation moyenne mensuelle (en mm) et Q la température moyenne mensuelle (en °C). Notez que cette expression considère des données mensuelles et dans le cas où ce sont des données annuelles, il n'y a pas le facteur 12 et P et Q seront des moyennes annuelles.

Ainsi, l'indice SPTI consiste à fitter une distribution gamma à la distribution de I_{DM} au lieu de P_{acc}^T pour l'indice SPI. De même que pour l'indice SPI, le choix de la fenêtre T est libre en fonction du phénomène étudié. Par ailleurs, d'après les auteurs, l'indice SPTI semble être plus adapté pour les basses températures et donne des résultats semblables à l'indice SPEI. L'avantage néanmoins par rapport à l'indice SPEI est de ne pas avoir à estimer le potentiel d'évapotranspiration qui est problématique pour des climats extrêmes.

2.6.2.4 CZI : China Z-Score

L'indice CZI ([7]) est développé vers 1996 en Chine avec pour objectif d'être un indice plus simple à calculer que l'indice SPI tout en étant aussi performant que ce dernier. Il utilise les données mensuelles de précipitation comme l'indice SPI et, sous l'hypothèse que ces données suivent une distribution de Pearson Type III, est défini par :

$$CZI_{Tj} = \frac{6}{C_{sT}} \left(\frac{C_{sT}}{2} \phi_{Tj} + 1 \right)^{1/3} - \frac{6}{C_{sT}} + \frac{C_{sT}}{6} \tag{7}$$

Avec :

$$C_{sT} = \frac{1}{n\sigma_T^3} \sum_{j=1}^{n} (x_{Tj} - \overline{x_T})^3 \quad \text{et} \quad \phi_{Tj} = \frac{x_j - \overline{x_T}}{\sigma_T} \tag{8}$$

Où n est le nombre total de mois, j le mois présent, T la période (entre 1 à 72 mois), C_{sT} le coefficient de skewness, x_{Tj} la précipitation du mois j accumulée sur les i mois précédents, $\overline{x_T} = \sum_{j=1}^{n} x_{Tj}/n$ la moyenne et $\sigma_T = \sqrt{\sum_{J=1}^{n} (x_{Tj} - \overline{x_T})^2/n}$ l'écart-type.

Nous précisons qu'il existe une variante de l'indice CZI, noté MCZI et est défini en remplaçant la moyenne par la médiane dans l'expression de ϕ_{Tj} et de C_{sT}. Par ailleurs, les auteurs montrent que l'indice CZI et l'indice SPI sont très corrélés mais avec le CZI qui indique des valeurs négatives beaucoup plus importantes lorsque les conditions de sécheresse sont sévères.

2.6.2.5 EDI : Effective Drought Index

L'indice EDI est développé sur la base des données quotidiennes de précipitation et utilise la notion de précipitation effective, noté EP, qui représente l'eau de précipitation stocké dans le sol, et est défini par :

$$\forall d \in \mathbf{N} \setminus \{0, 1, 2, ..., T\}, \quad EDI_T(d) = \frac{PRN_T(d)}{\sigma(PRN_T)} = \frac{DEP_T(d)}{\sigma(DEP_T)} \tag{9}$$

Où nous avons :

$$\forall d \in \mathbf{N} \setminus \{0, 1, 2, ..., T\}, \quad \begin{cases} EP_T(d) = \sum_{k=1}^{T} \left(\frac{1}{k} \sum_{l=1}^{k} P_l(d) \right) \\ DEP_T(d) = EP_T(d) - MEP \\ PRN_T(d) = \dfrac{DEP_T(d)}{\sum_{i=1}^{T} 1/i} \end{cases} \tag{10}$$

Avec $P_l(d)$ la précipitation du jour $d-l$, T la durée de sommation, MEP la moyenne de la série EP, PRN la précipitation nécessaire pour revenir à une situation normale et σ la fonction écart-type. Des études semblent montrer que l'indice EDI capte mieux l'émergence des sécheresses comparé à l'indice SPI.

Notez que d'autres définitions de précipitation effective sont proposées par les auteurs en partant des équations d'évolution de la réserve en eau différentes. Les auteurs proposent aussi des indices plus anecdotiques qui se sont révélés lors de leur étude.

2.7 Conclusion

Dans cette partie, nous avons vu d'abord que l'intérêt actuariel de l'étude réside dans le fait que le modèle que nous allons tenter de construire sera capable de prédire la charge IBNR totale pour une année donnée dès lors que l'année sera écoulée. Cette connaissance rapide de la charge IBNR permet

notamment une gestion plus contrôlée des provisions au sein des entreprises d'assurance et de réassurance.

Nous avons présenté ensuite le risque de subsidence et nous notons que la modélisation de ce dernier nécessite des données climatiques comme la température ou la précipitations, mais aussi des données non climatiques comme les propriétés du sol. Ces données libres d'accès sont globalement de bonne qualité sauf pour les données de réanalyse de précipitation journalière.

À partir des données climatiques, nous définissons in fine cinq indices de sécheresse, à savoir le Standardized Precipitation Index (SPI), le Standardized Precipitation Evapotranspiration Index (SPEI), le Standardized Precipitation Temperature Index (SPTI), le China-Z Index (CZI), et l'Effective Drought Index (EDI). La suite de l'étude consistera donc à exploiter ces indices de sécheresse et les données non climatiques à travers différentes approches.

Notez que l'idée générale de l'étude réside dans le principe de parcimonie. En effet, nous partons de l'idée que nous voulons un modèle simple et efficace. Au fur et à mesure des analyses, pour répondre aux différentes limites, nous verrons que le modèle est amené à se complexifier de plus en plus. Cette progression étape par étape est présente tout au long de l'étude et nous permettra, in fine, d'apporter une réponse claire à l'objectif actuariel.

3 Approche maille départementale

3.1 Introduction

Dans cette première partie, nous allons réaliser une approche annuelle et à l'échelle départementale. Dans une démarche parcimonieuse, cette dernière correspond plutôt à une approche exploratoire car nous verrons qu'elle atteint rapidement ses limites notamment en raison de la résolution départementale.

Ainsi, nous présentons dans un premier temps la base de données départementale composée de la variable d'intérêt et des variables explicatives. Dans un second temps, nous introduisons le modèle GLM et une amélioration possible de ce dernier en vue de la problématique de prédiction du projet. Cette amélioration étant insuffisante en raison des variables explicatives utilisées qui ne sont pas suffisamment pertinentes, nous utilisons l'outil d'analyse en composantes principales dans un dernier temps afin d'examiner la qualité de la base de données construite.

3.2 Base de données départementale

Nous commençons dans cette première section par définir et construire la base de données départementale. Cette dernière comporte, dans l'ordre, la variable d'intérêt ou variable réponse, les variables non climatiques et les variables climatiques.

3.2.1 Variable d'intérêt

Nous définissons ici la variable d'intérêt de l'approche départementale, i.e la variable que nous cherchons à expliquer à l'aide des variables explicatives. Il s'agit du taux de destruction par département r_x, à savoir le montant des sinistres observé au sein du département divisé par le montant total de la somme assurée de ce département. Ce taux, exprimé en pourcentage, est défini par :

$$\forall d \in \mathcal{D},\ \forall t \in \mathcal{A}, \quad r_x(d,t) = \frac{A_{dep}(d,t)}{SA(d,t)} \tag{11}$$

Avec $\mathcal{D}$ qui est l'ensemble des départements, $\mathcal{A} = 2009, ..., 2018$ l'ensemble des années, SA la somme totale assurée des bâtiments sans le contenu du département d pour l'année t, et A_{dep} le montant total des sinistres du département d pour l'année t en prenant en compte les provisions IBNR (Incurred but not reported). Notez que cette dernière correction est importante car le phénomène de sécheresse est un phénomène étalé dans le temps et le délai pour la reconnaissance d'un état de catastrophe naturelle est souvent long. Par conséquent, les montants des sinistres observés ne correspondent souvent qu'à une partie des montants réels des sinistres. En pratique, depuis 2009, le délai de décret d'un arrêté est de l'ordre d'un an et dépasse rarement deux ans. Pour ce type de sinistre, il faut considérer généralement un temps moyen de développement de trois ans.

3.2.2 Variables explicatives non climatiques

Nous allons à présent donner une définition pour les variables non climatiques dans le tableau de la figure 7. Ces variables explicatives non climatiques sont construites à partir des données portefeuille et des données du Bureau de Recherches Géologiques et Minières (BRGM, [3]).

Ces variables explicatives non climatiques utilisent des données comme le nombre de contrats ou le nombre de sinistres par zones issues du portefeuille. Elles utilisent aussi des cartes géologiques de la BRGM et notamment celles qui tentent de capturer les conditions argileuses du sol. La carte de la figure 8, réalisée par la BRGM, représente le niveau d'aléa retrait-gonflement des argiles des départements français métropolitains. Le niveau d'aléa va de 0 à 3 avec 0 le niveau le plus faible correspondant à une absence d'exposition et 3 le niveau le plus élevé correspondant à une exposition forte.

Notez que dans la construction des variables non climatiques, le niveau d'aléa de la surface d'une commune est déterminé par celui de ses coordonnées géographiques et que le niveau d'aléa d'un contrat

Nom	Description
$SA(d,t)$	La somme assurée totale en comptant uniquement les bâtiments pour le département d et l'année t.
$n_{risk}(d,t)$	Le nombre total de contrats au sein du département d pour l'année t.
$Surface(d,t)$	La surface totale des bâtiments assurés.
$lng(d)$	La longitude du département d comme moyenne des longitudes de ses communes.
$lat(d)$	La latitude du département d comme moyenne des latitudes de ses communes.
$rga_0(d)$	Le pourcentage de surface du département d classé en zone non argileuse par le BRGM.
$rga_1(d)$	Le pourcentage de surface du département d classé en zone de niveau d'aléa retrait-gonflement 1 par le BRGM.
$rga_2(d)$	Le pourcentage de surface du département d classé en zone de niveau d'aléa retrait-gonflement 2 par le BRGM.
$rga_3(d)$	Le pourcentage de surface du département d classé en zone de niveau d'aléa retrait-gonflement 3 par le BRGM.
$Prop_0(d)$	La proportion des contrats du département d situés en zone classée non argileuse par le BRGM.
$Prop_1(d)$	La proportion des contrats du département d situés en zone classée niveau d'aléa retrait-gonflement 1 par le BRGM.
$Prop_2(d)$	La proportion des contrats du département d situés en zone classée niveau d'aléa retrait-gonflement 2 par le BRGM.
$Prop_3(d)$	La proportion des contrats du département d situés en zone classée niveau d'aléa retrait-gonflement 3 par le BRGM.
$N_{pop}(d)$	La population du département d.
$TS(d)$	Le niveau d'aléa moyen du département d défini dans l'équation **??**.
$TC(d)$	Le niveau d'aléa moyen des contrats du département d défini dans l'équation **??**.
$r_{x,z}(d)$	Le taux zonier de destruction du département d défini dans l'équation **??**.

Figure 7 – Tableau descriptif des variables non climatiques construites à partir des données portefeuille et des données de la BRGM (figure 8, [3]) pour l'approche départementale.

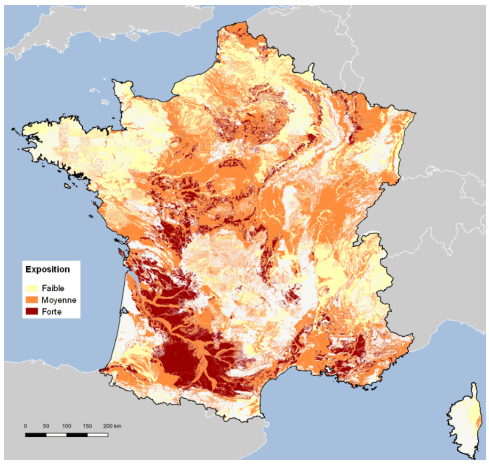

Figure 8 – Cartographie de l'aléa sécheresse réalisée par la BRGM. Le niveau d'aléa va de 0 à 3 avec 0 le niveau le plus faible correspondant à une absence d'exposition (blanc) et 3 le niveau le plus élevé correspondant à une exposition forte (rouge).

est le même que celui associé aux coordonnées géographiques de la commune où le bien réside. Notons que certains indices seront définis et présentés plus amplement par la suite.

3.2.3 Variable climatiques

À partir des indices de sécheresse de la partie précédente, il est possible de construire les variables explicatives climatiques de l'approche départementale. D'abord, notez qu'en considérant des données journalières et des fenêtres temporelles T en jour, il est possible d'adapter les indices afin d'avoir une résolution journalière. À ce stade de l'étude, nous avons 10 ans de données climatiques journalières, ce qui est plutôt restreint si nous restons à l'échelle mensuelle. Ainsi, dans la perspective d'avoir suffisamment de points pour calculer les indices de sécheresse qui nécessite d'approximer une distribution, nous travaillons à l'échelle journalière i.e avec des données de précipitation et de température journalières.

Dans cette étude, les cinq indices de sécheresse définis précédemment sont considérés avec chacun trois fenêtre T différentes ($T = 30$ jours, $T = 90$ jours, et $T = 180$ jours), ce qui nous fait un total de $3 \times 5 = 15$ séries temporelles. Le choix de ces indices et du T a pour objectif de construire un ensemble de variables suffisamment large pour capturer toutes les informations disponibles dans les données, quitte à supprimer certaines variables par la suite si ces dernières se révèlent peu significatives.

Notez que nous souhaitons construire ici une base de données annuelle et à résolution départementale. Cependant, les indices de sécheresse sont des données journalières et sur une grille couvrant le territoire français métropolitain comme montre la figure 2. Ainsi la question est de savoir comment nous pouvons convertir les données journalières en données annuelles et les données de grille en données de département. En effet, dans cette approche départementale, nous souhaitons avoir une base de données annuelle et à résolution départementale.

Pour répondre au problème soulevé précédemment, nous allons procéder en deux étapes. D'abord, pour passer des données de grille à des données par département, nous allons simplement utiliser une fonction moyenne. Soit $f_T^{lat,lng}$ une série temporelle journalière calculée sur la fenêtre temporelle T, par exemple SPI 30 jours, et localisée par sa latitude lat et longitude lng. Alors ce dernier est défini par :

$$\forall t \in \mathbf{N},\ \forall d \in \mathcal{D}, \quad f_T^d(t) = \frac{1}{|d|} \sum_{(lat,lng)\in d} f_T^{lat,lng}(t) \tag{12}$$

Avec t l'indice temporel, (lat, lng) le couple qui indique la localisation de la série, i.e les coordonnées du point bleu de la grille (figure 2) sur lequel est calculé la série $f_T^{lat,lng}$, et $|d|$ le nombre de points bleus de la grille situés dans le département d. Cette définition permet donc d'agréger géographiquement les points et de n'avoir qu'une seule série f_T^d par département. Et ceci pour les 15 indices f_T^d possibles.

Une fois que les données sont agrégées géographiquement, pour agréger l'axe temporelle, nous définissons pour toute date $dd/mm/yyyy$, $\overline{f_T^d}$ est défini par :

$$\overline{f_T^d}(dd/mm/yyyy) = \overline{f_T^d}(dd/mm) = \sum_{aaaa=2009}^{2018} f_T^d(dd/mm/aaaa) \tag{13}$$

Ainsi, par exemple pour le 1er janvier, $\overline{f_T^d}(01/01)$ est la moyenne sur les années des valeurs de f_T^d observées le 1er janvier. Notez que les 29 février des années bissextiles sont supprimés car il n'y a que deux années bissextiles sur la période 2009-2018 (2012 et 2016).

La figure 9 représente la courbe de $SPTI_T^d$ et la courbe de $\overline{SPTI_T^d}$ de l'Ardèche (département 07) avec une fenêtre temporelle $T = 180$ jours. Ainsi pour chaque année $k \in \mathcal{A}$, trois dates t_1^k, t_2^k et t_3^k qui correspondent respectivement à fin février, mi-juillet et mi-septembre sont définies.

À l'aide de ces dates, en notant $\mathbf{H}_k = \{(t_3^{k-1}, t_1^k), (t_1^k, t_2^k), (t_2^k, t_3^k)\}$ l'ensemble des trois couples de l'année k, pour tous les f_T^d disponibles, les quantités m_1, m_2, M_1, M_2, A_1, A_2, μ_1, μ_2 et S sont définies par :

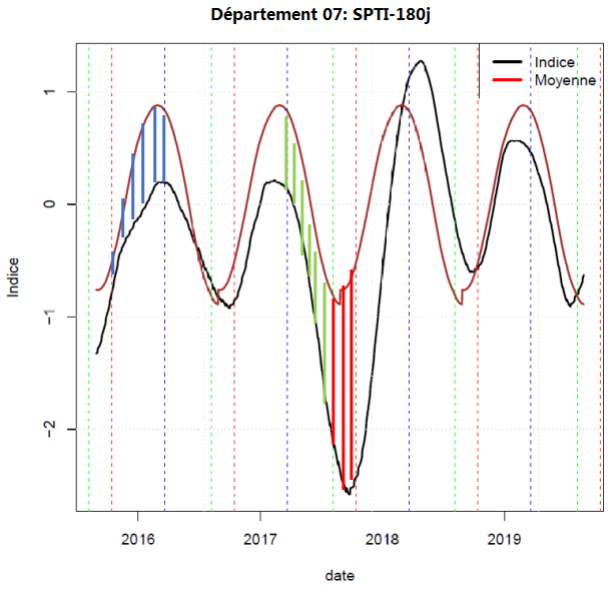

Figure 9 – Schéma illustrant le processus d'agrégation de l'axe temporel avec l'indice SPTI 180 jours de l'Ardèche (département 07) comme exemple. La courbe en noire correspond à f_T^d et la courbe en rouge à $\overline{f_T^d}$. Les droites verticales en pointillées bleue, verte et rouge correspondent respectivement aux trois dates t_1^k, t_2^k et t_3^k.

$$\forall k \in \mathcal{A},\ \forall (a,b) \in \mathbf{H}_k, \quad \begin{cases} m_1(k,a,b,f_T^d) = \min\limits_{a \le i < b} (f_T^d(i) - \overline{f_T^d}(i)) \\ m_2(k,a,b,f_T^d) = \min\limits_{a \le i < b} f_T^d(i) \\ M_1(k,a,b,f_T^d) = \max\limits_{a \le i < b} (f_T^d(i) - \overline{f_T^d}(i)) \\ M_2(k,a,b,f_T^d) = \max\limits_{a \le i < b} f_T^d(i) \\ A_1(k,a,b,f_T^d) = M_1(k,a,b) - m_1(k,a,b) \\ A_2(k,a,b,f_T^d) = M_2(k,a,b) - m_2(k,a,b) \\ \mu_1(k,a,b,f_T^d) = \dfrac{1}{b-a} \sum_{i=a}^{b} (f_T^d(i) - \overline{f_T^d}(i)) \\ \mu_2(k,a,b,f_T^d) = \dfrac{1}{b-a} \sum_{i=a}^{b} f_T^d(i) \\ S(k,a,b,f_T^d) = \dfrac{1}{b-a} \sum_{i=a}^{b} |f_T^d(i) - \overline{f_T^d}(i)|(|f_T^d(i) - \overline{f_T^d}(i)| + 1) \end{cases} \tag{14}$$

Notez pour chaque fenêtre T, chaque indice f_T^d et chaque couple (a,b), 9 indices sont calculés. Pour chaque indice, il y a autant d'observations que de nombre de départements multiplié par le nombre d'années. Autrement dit, dans la base de données de l'approche départementale, les variables climatiques sont par exemple $m_1(k, t_3^{k-1}, t_1^k, SPI_{30}^d)$ où le couple (k,d) décrit l'ensemble des observations. Sur le plan de la notation dans les différents graphes et sorties **R**, pour une meilleure lecture, nous suivons la même logique définie dans la partie 4.2.2.1.

Il faut noter que les quantités ci-dessus ont été construites dans la perspective de refléter les critères de reconnaissance d'état de catastrophe naturelle utilisés par la commission interministérielle ([1]). Par ailleurs, toutes ces quantités ainsi que les dates t_i^k ont été obtenues en examinant les évolutions temporelles des indices et des taux de destruction r_x de telle sorte que les variables climatiques résultante soient soient le plus corrélées possibles aux taux de destruction. En prime, cet examen a permis la suppression de certaines variables peu significatives. Les variables supprimées sont notamment ceux

issus des indices CZI et EDI qui sont trop bruités et instables pour être utilisables. In fine, la base de données de l'approche départementale comprend 890 observations et 243 variables en comptant celles qui ne sont pas issues des données climatiques.

3.3 Modèles linéaires

Pour explorer la base de données construite précédemment, des analyses basées sur des modèles de régression sont conduites dans cette section. Plus précisément, un modèle GLM est construit dans un premier temps en définissant au préalable une métrique d'erreur. Cependant ce dernier surestime le montant de la charge en raison des taux de destruction nuls non pris en compte. Pour remédier à cela, une amélioration est proposée en ajoutant un modèle logistique en parallèle.

3.3.1 Métrique d'erreur

D'abord, nous allons définir la métrique d'intérêt. En effet, la problématique de cette étude est de prédire le montant total annuel des sinistres et pour cela, il faut définir une métrique de décision. Ainsi, le taux de destruction annuel R_x est défini par :

$$\forall t \in \mathcal{A}, \quad R_x(t) = \frac{\sum_{d\in\mathcal{D}} A_{dep}(d,t)}{\sum_{d\in\mathcal{D}} SA(d,t)} = \frac{\sum_{d\in\mathcal{D}} r_x(d,t)SA(d,t)}{\sum_{d\in\mathcal{D}} SA(d,t)} \tag{15}$$

À partir de la définition de R_x, la métrique d'erreur ϵ_a, exprimée en pourcentage, est définie par :

$$\epsilon_a = \max_{t\in\mathcal{A}} |R_x(t) - \widetilde{R_x}(t)| \quad \text{avec} \quad \widetilde{R_x}(t) = \frac{\sum_{d\in\mathcal{D}} \widetilde{r_x}(d,t)SA(d,t)}{\sum_{d\in\mathcal{D}} SA(d,t)} \tag{16}$$

Avec $\widetilde{r_x}$ le résultat estimé à la sortie du modèle. Ou de manière équivalente, la métrique d'erreur ϵ_b, exprimée en euros, est définie par :

$$\epsilon_b = \max_{t\in\mathcal{A}} \left(|R_x(t) - \widetilde{R_x}(t)| \sum_{d\in\mathcal{D}} SA(d,t) \right) \tag{17}$$

Ainsi, si de manière individuelle, les prédictions de r_x par le modèle sont insatisfaisantes mais qu'en somme sur l'ensemble des départements, l'erreur sur R_x est acceptable, alors la modélisation peut être acceptée.

3.3.2 Modèle linéaire généralisé

3.3.2.1 Principe

Le modèle linéaire généralisé est une généralisation de la régression linéaire simple et permet notamment de traiter les problèmes dont la distribution des résidus ne suivent plus une loi gaussienne et dont la moyenne et la variance de la variable expliquée dépend d'une fonction de lien et de la modélisation de la variable expliquée.

Plus précisément, si $(1, X_1, X_2, ..., X_n)$ est l'ensemble des variables explicatives et Y la variable expliquée, alors le modèle peut s'écrire :

$$Y = \widehat{Y} + \epsilon \tag{18}$$

Où $\widehat{Y} = \mathbf{E}(Y|X)$ est l'estimateur linéaire, ϵ le terme d'erreur et Y supposé généré à partir d'une distribution de la famille des exponentielles. Il s'agit donc d'un modèle plus général car en notant g la fonction de lien, nous avons :

$$\mathbf{E}(Y|X) = g^{-1}(X\theta) \quad \text{et} \quad \mathrm{Var}(\widehat{Y}) = \mathrm{Var}(g^{-1}(X\theta)) \tag{19}$$

Avec θ le vecteur des paramètres associé à X le vecteur des variables explicatives. Notons que la constante de régression est inclue dans θ compte tenu de l'expression de X qui contient la constante 1. Et puisque Y suit une loi de la famille des exponentielles par hypothèse, θ est souvent estimé via l'optimisation de la log-vraisemblance.

3.3.2.2 Application

Dans cette approche départementale, la variable expliquée est r_x et les variables explicatives sont les 243 variables définies précédemment. La figure 10 représente la distribution du taux de destruction avec et sans les taux de destruction nuls. De manière générale, il s'agit plutôt d'une distribution à queue épaisse.

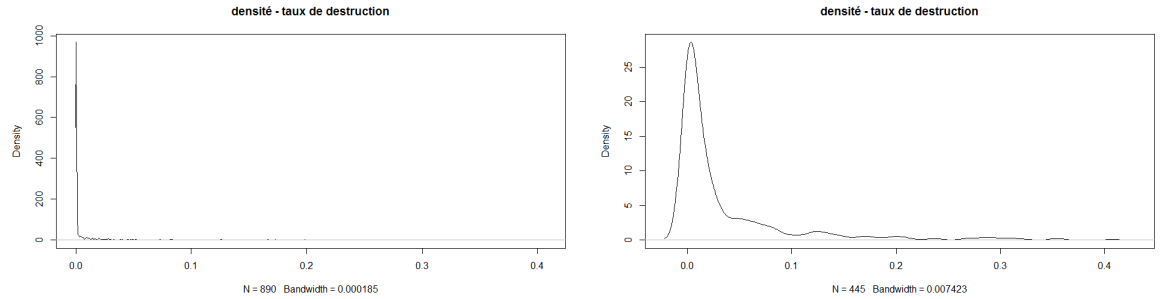

Figure 10 – La distribution du taux de destruction au sein de la base de données. La distribution de gauche contient les 890 observations et la distribution de droite contient uniquement les r_x non nuls. De manière générale, il s'agit plutôt d'une distribution à queue épaisse.

Ainsi, dans la modélisation GLM, une distribution gamma est utilisée. Dans la mesure où les taux de destruction sont strictement positifs, le choix de la fonction de lien est porté sur la fonction logarithme.

Par la suite, dans la mesure où il y a beaucoup plus de variables par rapport au nombre d'observations, il est nécessaire de faire un tri et sélectionner uniquement les variables significatives. Notez qu'un premier tri des variables a déjà été réalisé lors de la procédure de construction de ces dernières. Pour cela, l'algorithme de sélection des variables utilisé lors de la construction du modèle GLM est la suivante :

Algorithme

1. Soit un modèle initial qui comporte une seule variable explicative, en l'occurrence le taux zonier de destruction $r_{x,z}$.
2. Ajouter une variable explicative au modèle initial. Si le nouveau modèle construit a une erreur ϵ_b et un critère AIC plus petits que ceux du modèle initial, tout en ayant les p-values inférieur au seuil de 5%, alors le nouveau modèle avec l'ajout de la variable est considéré comme nouveau modèle initial.
3. Réitérer l'étape précédente tant qu'il est encore possible d'ajouter des variables explicatives.
4. Réitérer les deux précédentes étapes en considérant l'ensemble des termes d'interaction du type $a \times b$ avec a ou b déjà présent dans le modèle issus de l'étape précédente.

L'algorithme est appliqué et dans la figure 11, les montants des sinistres observés entre 2009 et 2018 et les montants des sinistres prédits par le modèle sur la même période sont représentés. Notez qu'il n 'y a pas d'ensemble test dans cette première modélisation, donc tout est mesuré sur l'ensemble d'entraînement. Le modèle GLM gamma ainsi construit contient 10 variables explicatives avec un terme d'interaction et les détails de la sortie sont fournis en annexe 9.2. Notez qu'il n'y a pas d'intérêt à s'attarder sur les coefficients à ce stade puisqu'il s'agit encore d'une approche exploratoire.

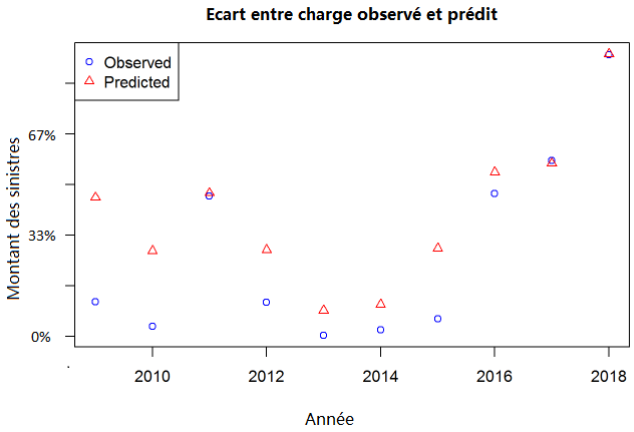

Figure 11 – Écart entre les montants des sinistres observés entre 2009 et 2018 (rouge) et les montants des sinistres prédits par le modèle sur la même période (bleu) en pourcentage par rapport au maximum des montants. De manière générale, il y a une surestimation quasi systématique des montants. La modélisation est donc loin d'être satisfaisante.

A travers la figure 11, il semble que le modèle surestime les montants quasi systématiquement. De plus, les écarts entre les montants des sinistres observés et les montants des sinistres prédits sont très importants. En effet, pour avoir une précision de l'ordre de 90%, il faut que ϵ_b soit de l'ordre de quelques pourcents (par rapport au maximum des montants), ce qui n'est pas le cas par exemple pour 2009 ou 2010.

3.3.3 Modèle composé

Dans l'objectif de répondre au problème de surestimation du modèle linéaire généralisé précédent, nous allons envisager un modèle binomial en parallèle du modèle linéaire. En effet, dans le modèle précédent, afin de pouvoir utiliser la distribution gamma et la fonction de lien logarithme, toutes les observations dont le taux de destruction r_x est nul ont été enlevées, ce qui semble aboutir à une surestimation des r_x.

L'idée est donc de prendre en compte ces observations où $r_x = 0$ via un modèle binomial ou modèle logistique en prédisant pour chacune des 890 observations, la probabilité p_1 que cette observation est associée à un taux de destruction non nul. Ce type de modèle composé est plus communément connu dans la littérature sous le nom de modèle zéro modifié ou zéro inflaté en fonction l'hypothèse sur la provenance des zéros excessifs faite au départ. La construction du modèle logistique est effectuée de nouveau en suivant l'algorithme défini précédemment. Le modèle résultant contient 28 variables explicatives avec 14 termes d'interaction et les détails de la sortie sont fournis en annexe 9.3. De même, il n'est pas nécessaire de s'attarder sur les valeurs des coefficients car l'intérêt est limité à ce stade.

Ainsi la nouvelle prédiction est définie à présent comme $p_1\widetilde{r_x}$, i.e le produit entre la sortie du modèle logistique, p_1, et la sortie du modèle linéaire généralisé, $\widetilde{r_x}$. Les résultats agrégés par année de cette modélisation sont fournis dans la figure 12.

Notez d'abord qu'il n'y a plus cette surestimation générale des montants observée précédemment. Néanmoins, la modélisation reste insatisfaisante, par exemple pour les années 2016, 2017 et 2018. Ceci pose problème car dans cette étude, il est crucial de bien modéliser les montants élevés. Une raison possible pouvant expliquer cela est que les variables explicatives utilisées ne sont peut-être pas suffisamment pertinentes. Pour essayer de distinguer s'il s'agit d'un problème lié à une mauvaise sélection des variables ou s'il s'agit plutôt d'un problème lié à la qualité des variables construites, nous allons faire une analyse en composantes principales.

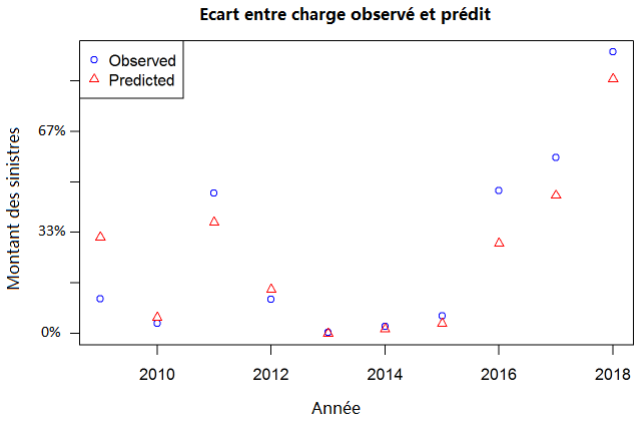

Figure 12 – Écart entre les montants des sinistres observés entre 2009 et 2018 (rouge) et les montants des sinistres prédits par le modèle composé sur la même période (bleu). Notez qu'il n'y a plus cette surestimation générale des montants observée précédemment. Néanmoins, la modélisation reste insatisfaisante, notamment pour les années 2016, 2017 et 2018.

3.4 Modélisation via l'analyse en composantes principales

L'idée de cette section est que puisque les variables explicatives semblent ne pas être suffisamment pertinentes d'après nos premières tentatives de modélisations, nous allons construire explicitement des variables pertinentes. Cette construction sera réalisée à l'aide de l'analyse en composantes principales (ACP), une méthode d'algèbre linéaire. Ainsi, nous commençons par définir la méthode d'ACP puis nous analyserons la pertinence de l'approche départementale à l'aide des régressions sur composantes principales.

3.4.1 Définition

Nous commençons par décrire rapidement le principe de l'analyse en composantes principales. L'idée de cette méthode consiste à décomposer l'espace des données en des directions appelées composantes principales. À chaque direction est associée une valeur propre qui correspond à la variance des données dans cette direction. Il s'agira ensuite de projeter les données dans les quelques directions où les variances sont les plus grandes afin de perdre le moins d'information possible. L'analyse est ensuite réalisée dans cette nouvelle base réduite.

D'un point de vu mathématique, si $(X_1, X_2, ..., X_N)$ sont les variables explicatives avec N le nombre de variables et $K = 890$ le nombre de réalisations, en regroupant les données sous forme matricielle $\mathbf{X} = (X_{i,j})_{\substack{1\leq i\leq K\\ 1\leq j\leq N}}$, alors l'analyse en composantes principales consiste à réaliser une décomposition en valeurs singulières, i.e trouver des matrices de passages $\mathbf{U}$ et $\mathbf{V}$ tel que :

$$\overline{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}} \tag{20}$$

Avec $\overline{\mathbf{X}} = (X_{i,j} - K^{-1}\sum_{l=1}^{K} X_{i,l})_{\substack{1\leq i\leq N\\ 1\leq j\leq K}}$, $\mathbf{U}$ et $\mathbf{V}$ des matrices de passage orthogonales respectivement de tailles $K \times K$ et $N \times N$, et $\mathbf{D}$ la matrice rectangulaire diagonale des variances de taille $K \times N$. Dans la figure 13, le principe de l'analyse en composantes principales est illustré à l'aide d'une image extraite du web. L'idée de la méthode est donc d'analyser les données non pas dans la base cartésienne initiale mais dans la base des vecteurs propres, i.e les vecteurs qui composent les matrices de passages $\mathbf{U}$ et $\mathbf{V}$. Ainsi, sur la figure 13, cela revient à considérer les deux vecteurs centrés sur le centre de gravité du nuage des points comme nouvelle base d'étude.

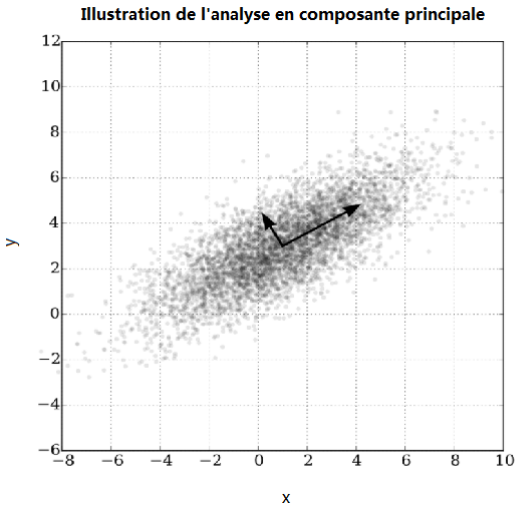

Figure 13 – Image extraite du web illustrant le principe de l'analyse en composantes principales. L'idée de la méthode est donc d'analyser les données non pas dans la base cartésienne mais dans la base des vecteurs propres indiquée par les deux vecteurs centrés sur le centre de gravité du nuage des points.

3.4.2 Modèle de régression sur composantes principales

3.4.2.1 Principe

Nous allons à présent réaliser un modèle de régression linéaire sur composante principale ou PCR. L'objectif majeur étant de réduire le nombre de régresseurs ou variables explicatives tout en perdant le moins d'information possible.

Cette méthode, basée sur l'analyse en composantes principales, utilise non pas les variables explicatives directement comme régresseurs mais projette au préalable les variables explicatives dans la base des vecteurs propres issue de l'ACP. L'avantage est donc de choisir uniquement les quelques directions qui captent la majorité de la variance pour construire la base des vecteurs propres. L'histogramme de la figure 14 montre la contribution de chaque vecteur propre ou composante principale à l'explication de la variance totale au sein des données.

Notez qu'avec les cinq premières composantes principales, environ 60% de la variance totale est expliquée. Et avec trente composantes principales, ce pourcentage de variance expliquée monte à environ 85%. L'idée est donc de régresser sur les trente composantes principales, ce qui reste un nombre important mais néanmoins à relativiser par rapport aux 243 variables initiales.

3.4.2.2 Modélisation

Dans la construction du modèle PCR, pour évaluer le problème de généralisation, une année servant d'année de test est mise de côté. Le modèle PCR est ensuite calibré sur toutes les années sauf l'année de test. La figure 15 montre ainsi l'écart entre les montants des sinistres observés et les montants des sinistres prédits avec soit l'année 2014, soit l'année 2018 qui sert d'année de test.

Notez qu'en fonction de l'année de test, il y a une grande variabilité dans les montants des sinistres prédits. Lorsque l'année 2014 sert d'année de test, les erreurs sont moindres en général comparées au modèle composé de la figure 12. De l'autre côté, lorsque l'année 2018 sert d'année de test, les résultats sont meilleurs sur les années d'entraînement mais la généralisation est loin d'être acceptable puisque l'erreur pour 2018 est très grande. Nous allons donc analyser plus en détail ce problème de généralisation par la suite.

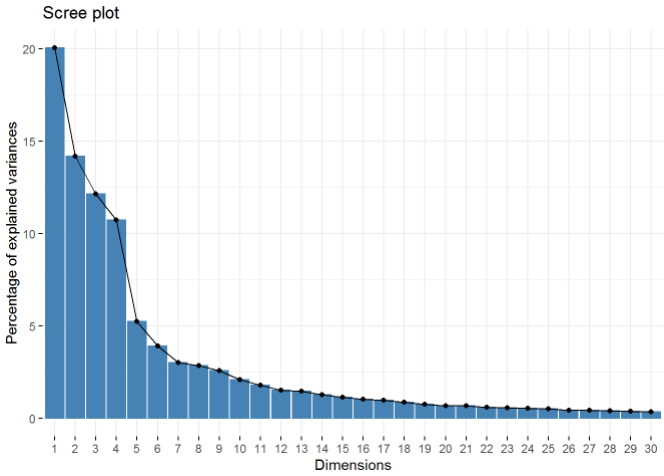

Figure 14 – Scree plot représentant le pourcentage de variance expliqué par chaque composante principale dans l'ordre décroissant. Avec les cinq premières composantes principales, environ 60% de la variance totale est expliquée. Et avec trente composantes principales, ce pourcentage de variance expliquée monte à environ 85%.

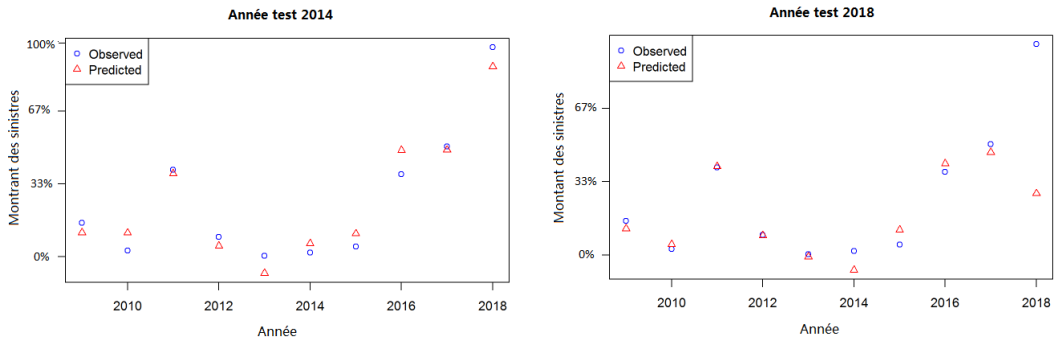

Figure 15 – Montants des sinistres observés (bleu) et prédits (rouge) avec 2014 (à gauche) ou 2018 (à droite) comme année de test. Lorsque l'année 2014 sert d'année de test, les erreurs sont moindres en général comparées au modèle composé de la figure 12. Néanmoins les prédictions des années 2011 et 2016 restent discutables. Lorsque l'année 2018 sert d'année de test, les résultats sont certes meilleurs sur les années d'entraînement mais la généralisation est loin d'être acceptable compte tenu de l'erreur en 2018.

Figure 16 – Nuage des points des données projetées dans les deux directions principales de l'ACP où les variances expliquées sont les plus grandes avec en échelle de couleurs les années. Les observations de l'année 2018 correspondent aux points violets et les observations de l'année 2014 correspondent aux points roses foncés. Notez que les observations de l'année 2018 se distinguent des autres, ce qui n'est pas le cas pour les observations de l'année 2014.

3.4.2.3 Analyse

Afin de comprendre le problème de généralisation lorsque l'année 2018 sert d'année de test, il faut essayer de visualiser les points de la base de données dans la base des vecteurs propres issus de l'ACP. Notez en passant que ce problème s'apparente à un problème de sur-apprentissage avec une très bonne prédiction sur les données d'entraînement mais une très mauvaise prédiction sur les données de test. Ainsi, dans la figure 16, les données sont projetées dans les deux directions où les variances expliquées sont les plus grandes.

D'après la figure 16, les observations de l'année 2018 sont particulières car se distinguent de la masse de points. Par conséquent, dès lors que ces observations sont enlevées, il n'existe plus d'observations dans la région occupée par les points violets et donc il est assez logique de ne pas pouvoir prédire les points de cette région à l'aide des autres années, d'où la mauvaise généralisation. À l'opposé, dans la région où se trouvent les observations de l'année 2014, d'autres années comme 2012 ou 2016 sont présentes. Ainsi, même si les points 2014 sont retirés, il est toujours possible d'avoir une bonne prédiction pour les points situés dans cette région, d'où la bonne généralisation. Notez donc qu'il ne s'agit pas réellement de la sur-apprentissage mais plutôt d'apparition de caractéristiques nouvelles et il est par conséquent normal que les modèles simples ne soient pas capables de bien prédire.

Remarquez que pour avoir la certitude qu'il y a d'autres points dans la région occupée par les observations 2014, il aurait fallu visualiser les données dans tout l'espace et non pas uniquement dans la base des vecteurs projetés. En effet, il se peut très bien qu'il existe une direction, perdue lors de la projection, qui sépare les observations 2014 des autres. Cette visualisation multidimensionnelle étant impossible, plusieurs visualisations similaires à celle de la figure 16 mais avec les deux vecteurs propres tirés aléatoirement parmi les 30 vecteurs propres les plus importants sont faites. Et à chaque tirage, il existe d'autres points dans la région occupée par les observations 2014 contrairement aux observations 2018.

3.4.3 Modèle de régression sur composantes principales avec données ajustées

Avant d'émettre des conclusions sur la pertinence des variables ou de manière générale sur la base de données à l'échelle départementale, nous allons analyser davantage les limites du modèle PCR. En effet, dans ce modèle, l'ensemble des 890 observations disponibles sont utilisées. Parmi ces observations, environ la moitié ont des taux de destruction nuls et parmi l'autre moitié restante, plus de deux tiers correspondent plutôt à des risques attritionnels. La question naturelle ici est donc de savoir si en restreignant aux gros risques, le modèle PCR ne se généralisent pas mieux.

Il s'agit d'une question intéressante car schématiquement à partir de la figure 16, il faut se dire que le

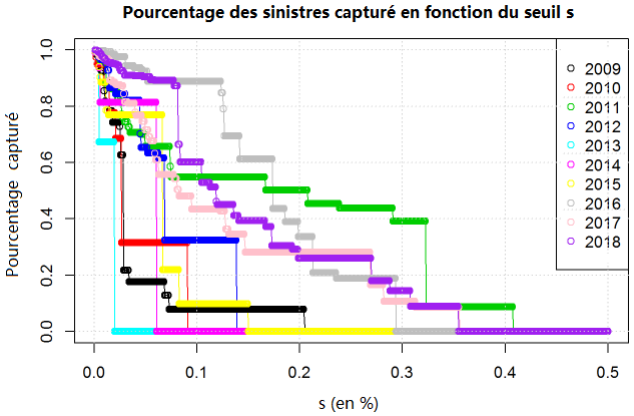

Figure 17 – Évolution de p_{cap} en fonction du seuil s pour tout $t \in \mathcal{A}$. Les risques attritionnels correspondent plutôt aux p_{cap} proche de 1 et s proche de 0. Nous fixons arbitrairement le seuil à 0.1% d'après ce plot.

modèle fonctionne lorsque les nouvelles observations tombent dans le voisinage, au sens mathématique du terme, de l'un des points présent dans les données d'entraînement. Ce voisinage, où nous avons une bonne prédiction, peut être vu comme une boule multidimensionnelle d'un certain rayon r qui est déterminé par les données d'entraînement, le modèle utilisé et la méthode de calibration. Le fait de restreindre aux gros sinistres revient, pour notre base de données départementale, à diminuer le nombre d'observations dans les régions de l'espace où il y a déjà beaucoup de points et ainsi forcer le modèle à étendre son voisinage de bonne prédiction. En d'autres termes, cela facilite le modèle à se calibrer sur un rayon r plus grand et la question est de savoir si c'est suffisant pour englober la région qui contient les observations de l'année 2018.

Ainsi, un seuil noté s est fixé afin de ne tenir compte que des sinistres importants. Pour fixer ce seuil s, la quantité p_{cap}, appelé pourcentage capturé, est définie par :

$$\forall t \in \mathcal{A},\ \forall s \in [0,1]\,, \quad p_{cap}(t,s) = \frac{\sum\limits_{\substack{d \in \mathcal{D} \\ r_x(d,t) > s}} r_x(d,t).SA(d,t)}{\sum\limits_{\substack{d \in \mathcal{D} \\ r_x(d,t) > 0}} r_x(d,t).SA(d,t)} = 1 - \frac{\sum\limits_{\substack{d \in \mathcal{D} \\ r_x(d,t) \leq s}} r_x(d,t).SA(d,t)}{\sum\limits_{\substack{d \in \mathcal{D} \\ r_x(d,t) > 0}} r_x(d,t).SA(d,t)} \tag{21}$$

La figure 17 représente p_{cap} pour toutes les années disponibles. Notez que la définition de p_{cap} est semblable à la définition d'une fonction de survie avec comme variable aléatoire le montant total A_{dep} des sinistres.

À l'aide de cette figure, les risques attritionnels sont facilement visualisés. En effet, un risque attritionnel correspond à une faible diminution de p_{cap} et est souvent associé à un taux de destruction faible. Ces risques sont donc sur la figure au niveau de p_{cap} proche de 1 et s proche de 0. Les risques importants quant à eux correspondent plutôt à des diminutions importantes de p_{cap}. Ainsi, le seuil s est fixé arbitrairement à 0.1%.

Après avoir fixé le seuil, le modèle PCR est recalibré avec uniquement les observations où le taux de destruction est supérieur à s. De même, une année est mise de côté pour servir d'ensemble de test et le modèle est entraîné sur les données restantes. Dans la figure 18, les montants des sinistres observés et les montants des sinistres prédits avec soit l'année 2011, soit l'année 2018 qui sert d'année de test sont représentés.

Le modèle construit est un modèle PCR avec une seule composante principale pour 69 observations et il y a clairement une meilleure généralisation comme voulu. Néanmoins l'erreur associée à ce modèle est de l'ordre de 10%, ce qui n'est pas acceptable. De plus, même si l'erreur est nulle, la modélisation tient compte uniquement des risques importants et néglige les risques attritionnels, qui, en somme,

Figure 18 – Montants des sinistres observés (bleu) et prédits (rouge) avec 2011 (à gauche) ou 2018 (à droite) comme année de test pour un seuil $s = 0.1\%$. De manière générale, il y a clairement une meilleure généralisation comparé au cas où les sinistres attritionnels et les sinistres importants sont modélisés en même temps. Néanmoins, notez que le modèle affiche une erreur de l'ordre de 10%, ce qui n'est pas négligeable.

peuvent devenir non négligeables. Par exemple, pour le seuil s choisi, $p_{cap}(2018, 0.1) = 60$ %, ce qui signifie 40% des charges, chiffre qui correspond à la somme des risques attritionnels, sont négligées.

Bien entendu, la modélisation PCR peut être améliorée par exemple en traitant séparément les gros sinistres et les sinistres attritionnels. Cependant cette approche n'est pas souhaitable. En effet, de manière sous-jacente, le modèle PCR utilise l'ensemble des 243 variables car chaque composante principale est construite en utilisant l'ensemble de la base de données. Et ceci est contraire à l'idée de départ qui est d'avoir un modèle simple avec peu de variables.

Néanmoins, cette modélisation à l'aide de l'ACP confirme que d'une part, il y a un manque de données dans les années qui rend les modèles difficiles à généraliser. D'autre part, les variables explicatives construites ne sont pas suffisamment pertinentes. Ces dernières semblent être adaptées pour décrire les sinistres importants mais pas les sinistres attritionnels. Ainsi, il est probablement nécessaire de revoir l'approche plus en amont et notamment la construction de la base de données. Par conséquent, une approche toujours annuelle mais à résolution communale est envisagée par la suite.

3.5 Conclusion

Dans cette première approche annuelle à résolution départementale, nous avons d'abord construit une base de données départementale à l'aide des données de climat et des données du sol. Un modèle GLM est ensuite proposé dans un premier temps en définissant au préalable une méthode de sélection des variables et une amélioration du modèle en ajoutant un modèle logistique en parallèle de ce dernier est proposée dans un second temps afin de répondre aux problèmes de surestimation. Les résultats étant toujours insatisfaisants, nous nous demandons s'il ne s'agit pas d'un problème de qualité des variables explicatives construites. Ce problème de qualité peut provenir d'une mauvaise sélection des variables mais aussi de la mauvaise qualité des données de précipitations journalière utilisées. Pour répondre à cela, nous utilisons l'analyse en composantes principales et des régressions sur composantes principales. L'analyse des résultats montrent que d'une part, il y a un réel manque de données dans les années, et que d'autre part, les variables explicatives construites ne sont pas suffisamment pertinentes. Ces dernières semblent être adaptées pour décrire les sinistres importants mais pas les sinistres attritionnels. Ainsi, toujours dans la perspective d'une démarche parcimonieuse, afin d'apporter une réponse à ces limites soulevées par l'approche départementale, il est probablement nécessaire d'adopter une approche plus fine et plus poussée.

4 Affinage des données, approche communale et méthodologies

4.1 Introduction

Pour répondre au problème de manque de données et de la pertinence des variables, dans cette partie, une nouvelle approche toujours annuelle mais avec une résolution spatiale communale est envisagée. Notez que pour cette approche, afin d'être le plus proche de la réalité et donc du régime CatNat français, il est nécessaire de faire une modélisation en deux étapes. En effet, il faut d'abord modéliser le reconnaissance ou non d'une commune en état de catastrophe naturelle par l'autorité, puis modéliser les charges des sinistres pour les communes reconnues.

Pour cela, dans un premier temps, une nouvelle base de données à résolution communale couvrant la période 2016-2018 avec environ 35 000 observations par année et en intégrant les données de reconnaissance issues du Journal Officiel est construite. Notez que le choix de la période 2016-2018 est motivé par les nouveaux critères de reconnaissance qui seront présentés en détails dans une prochaine partie. Dans un second temps, des méthodes de sélection des variables pour le modèle de reconnaissance ou d'occurrence vont être définies, puis la construction de ce dernier à l'aide des méthodes de machine learning sera présentée dans un troisième temps. Dans un quatrième temps, le modèle de charge, ou plutôt de fréquence, sera construit à l'aide de la théorie des modèles GLM, en détaillant de nouveau les méthodes de sélection des variables. Enfin, dans un dernier temps, les résultats du modèle complet dit occurrence-fréquence combinant le modèle d'occurrence et le modèle de fréquence seront analysés et il se trouve que ces derniers sont plutôt encourageants.

Notez que pour obtenir des variables explicatives pertinentes, l'idée générale de cette partie est d'utiliser les données de température et de précipitation de bonne qualité, i.e mensuelles, et de construire la base de données la plus complète possible afin de s'assurer que tous les informations nécessaires à la modélisation sont présentés. Bien entendu, cela introduit de nombreux problèmes de corrélation et de sur-apprentissage et c'est pourquoi le cœur du travail par la suite sera notamment de réduire au fur et à mesure, pour chacun des deux modèles, le nombre de variables à l'aide des méthodes de sélection. Le fait d'établir une méthodologie claire et bien fondée permet notamment de contrôler chaque étape de sélection et d'être en mesure d'expliquer le choix fait par les algorithmes. Cette démarche minutieuse, en partie survolée lors de l'approche départementale, donne des garanties sur la pertinence des variables finalement utilisées. À titre indicatif, la base comporte plus de 800 variables explicatives mais après sélection, moins d'une centaine sont effectivement utilisées dans le modèle occurrence-fréquence.

4.2 Base de données de catastrophe naturelle

Dans cette section, nous allons d'abord rappeler brièvement la procédure de reconnaissance de l'état de catastrophe naturelle et ses fondements. Puis nous définirons les variables climatiques comme pour l'approche départementale avec néanmoins les données de précipitations et de température qui ne sont plus journalières mais mensuelles. Nous définirons ensuite les variables géologiques à l'échelle communale et nous ajouterons en plus de ces variables, des variables socio-économiques dérivées des données d'INSEE et de la Direction générale des finances publiques. La base de données construite dans cette section est appelée base de données de catastrophe naturelle.

4.2.1 Procédure reconnaissance

Dans la préambule de la Constitution du 27 octobre 1946 et repris par la Constitution de 1958, il est stipulé que "La Nation proclame la solidarité et l'égalité de tous les Français devant les charges qui résultent des calamités nationales". Ainsi, un dispositif, instauré par la loi du 13 juillet 1982 et codifié par les articles L.125-1 et suivants du Code des Assurances, organise l'indemnisation des sinistrés dont les biens assurés ont été endommagés par un phénomène naturel intense : il s'agit de la garantie catastrophe naturelle. Notez que l'article L.125-1 du Code des Assurances précise notamment que "sont considérés comme les effets des catastrophes naturelles, les dommages matériels directs ayant eu pour cause déterminante l'intensité anormale d'un agent naturel, lorsque les mesures habituelles à prendre

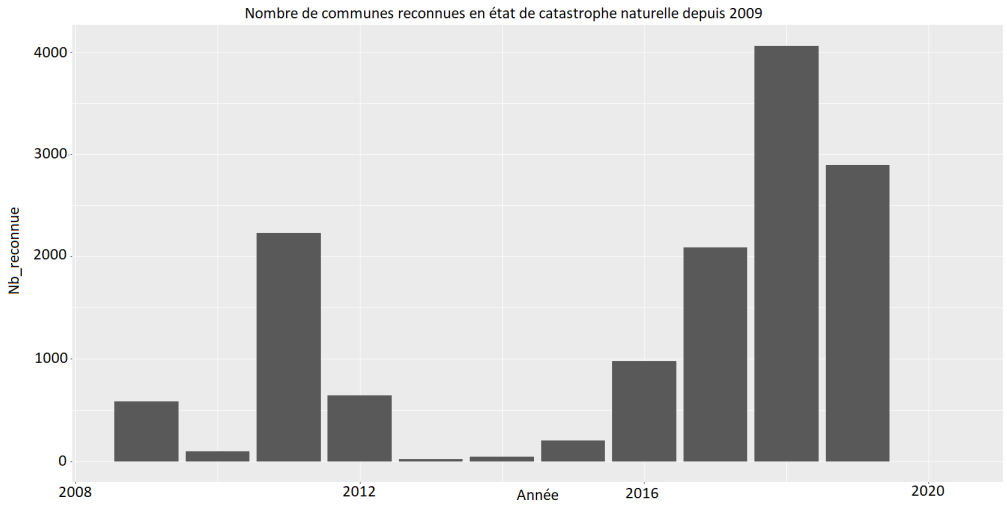

Figure 19 – Évolution du nombre de reconnaissances d'état de catastrophe naturelle depuis 2009. Les chiffres de l'année 2019 sont encore susceptibles d'évoluer et pour l'année 2020, les décisions d'arrêté ne sont pas encore tombées. Notez que l'année 2011 et les années 2016, 2017, 2018 et 2019 sont des années très sinistrées, ce qui est cohérent avec la littérature et les différentes dates de modification des critères de reconnaissance (partie 5.2.1)

pour éviter ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises", ce qui inclut le risque de subsidence bien entendu.

En pratique, mis à part les démarches personnelles des personnes sinistrées auprès de leurs propres assureurs selon les conditions prévues par leurs contrats, le maire de la commune ayant subi une catastrophe naturelle peut formuler une demande de reconnaissance auprès du préfet de département. Les services compétents de ce dernier contrôlent ensuite le contenu de la demande et réunissent les rapports d'expertise permettant de caractériser l'intensité du phénomène naturel à l'origine des dégâts recensés par la mairie. Par la suite, une commission interministérielle, présidée par le ministère de l'intérieur, est chargée de donner un avis sur le caractère naturel et l'intensité anormale du phénomène en se basant sur les expertises techniques, pour chacune des dossiers communaux transmis par les préfets des départements. Sur la base de cet avis, les ministres compétents décident ensuite de la reconnaissance ou non des communes en état de catastrophe naturelle. Les décisions finales sont in fine formalisées par un arrêté interministériel publié au Journal Officiel et c'est seulement en cas de décision positive que les différentes assurances peuvent se déclencher.

Les critères exacts utilisés par la commission interministérielle pour formuler leur avis seront plus amplement abordés dans la section 5.2.1. Pour le moment, il faut surtout noter que les données historiques des arrêtés des catastrophes naturelles sont disponibles publiquement et remontent jusqu'aux années 90. Ainsi, pour les communes françaises et pour le risque de subsidence, nous savons quelles sont les communes qui ont fait une demande d'arrêté, la période de l'année où le sinistre a eu lieu et la décision finale de la commission interministérielle. Dans l'approche communale, pour le modèle d'occurrence, l'obtention ou non d'un arrêté de catastrophe naturelle est la variable d'intérêt et sera expliquée à l'aide des variables construites par la suite.

La figure 19 représente le nombre de communes reconnues en état de catastrophe naturelle pour les années 2009 jusqu'à 2019. Néanmoins, il faut préciser que pour l'année 2019, les chiffres sont encore susceptibles d'évoluer et pour l'année 2020, les décisions d'arrêté ne sont pas encore tombées.

Par ailleurs, notez que l'année 2011 et les années 2016, 2017, 2018 et 2019 sont des années très sinistrées, ce qui est cohérent avec la littérature et les différentes dates de modification des critères de reconnaissance (partie 5.2.1).

4.2.2 Variables climatiques

Nous allons à présent définir les variables climatiques de cette approche communale. Notez que ces dernières sont différentes de celles de l'approche départementale car nous utilisons désormais des séries temporelles mensuelles et non plus journalières. Ce passage des données journalières aux données mensuelles est motivé par le fait que les données mensuelles de précipitation sont de meilleure qualité comme le montre la comparaison de la section 2.5. Bien entendu, cela nécessite de récupérer plus de données depuis la base ERA5 sachant que dans l'approche départementale, dans la perspective d'une démarche parcimonieuse, seules 10 ans de données ont été récupérées. D'après les recommandations des auteurs des indices de sécheresses, il faut récupérer au moins 30 ans de données mensuelles. Ensuite, comme dans l'approche départementale, pour construire les variables climatiques, il faut agréger les données temporellement et spatialement.

4.2.2.1 Agrégation temporelle

Dans la construction de la nouvelle base de données, les indices de sécheresse utilisés sont les mêmes que précédemment, à savoir le SPI, le SPEI, le SPTI, le CZI, et l'EDI. Les définitions des variables explicatives sont semblables à celles dans l'équation 14. Dans la mesure les données sont mensuelles et non plus journalières, pour chaque année t, l'ensemble $\mathcal{P}_t$ composé de cinq périodes est défini par :

$$\forall t \in \mathcal{A}, \quad \begin{cases} P^t_{hiv} = 01-03 \\ P^t_{pri} = 04-06 \\ P^t_{ete} = 07-09 \\ P^t_{aut} = 10-12 \\ P^t_{all} = 01-12 \end{cases} \tag{22}$$

Avec les mois (janvier, février, ..., décembre) en chiffres (01, 02, ..., 12). Notez que ces périodes correspondent à celles utilisées pour calculer les critères de catastrophe naturelle par la commission interministérielle([1]).

Ainsi, comme pour l'approche départementale, il faut agréger les données de grille de telle sorte à obtenir des données par commune et par année. Pour cela, pour agréger l'axe temporelle, semblablement à l'approche départementale, pour chaque indice f^c_T où T est la taille de la fenêtre glissante et c la commune, par exemple le SPI 1 mois, les quantités m, M, S_{+-} et S_- sont définies par :

$$\forall t \in \mathcal{A},\ \forall P \in \mathcal{P}_t, \quad \begin{cases} m(t,P,f^c_T) = \min\limits_{i\in P} f^c_T(i) \\ M(t,P,f^c_T) = \max\limits_{i\in P} f^c_T(i) \\ S_-(t,P,f^c_T) = -\sum_{i\in P} f^c_T(i).\mathbf{1}_{f^c_T(i)<0}(i) \\ S_{+-}(t,P,f^c_T) = -\sum_{i\in P} f^c_T(i) \end{cases} \tag{23}$$

En d'autres termes, pour chaque fenêtre T, pour chaque indice f^c_T et pour chaque période P, 4 indices sont calculés. Le nombre d'observations est donc égale au de nombre de communes multiplié par le nombre d'années, i.e le couple (c,t) décrit l'ensemble des observations. Pour le choix de la taille de fenêtre, afin d'être le plus exhaustif possible, T vaut 1 mois, 3 mois, 6 mois, 12 mois ou 24 mois. Pour une meilleure lecture des résultats, dans la notation de ces quantités dans les figures, les quantités $m(t,P,f^c_T)$ sont accompagnées du suffixe min, les quantités $M(t,P,f^c_T)$ sont accompagnées du suffixe max, les quantités $S_-(t,P,f^c_T)$ sont accompagnées du suffixe $sum\_negative$ et les quantités $S_{+-}(t,P,f^c_T)$ sont accompagnées du suffixe sum.

Ce point est illustré dans la figure 20 avec l'indice SPEI. Par exemple la quantité $S_-(t,P^t_{hiv},SPEI^c_{12})$, i.e l'indice SPEI calculé avec les données mensuelles sur une fenêtre glissante de 12 mois pour la commune c et l'année t, est notée par $SPEI\_12m\_critere\_hiver\_sum\_negative(c,t)$. Ainsi, notez qu'en fixant c mais pas t, nous avons une série temporelle et qu'en fixant t mais pas c, nous avons une carte

de la France avec l'indice en échelle de couleur.

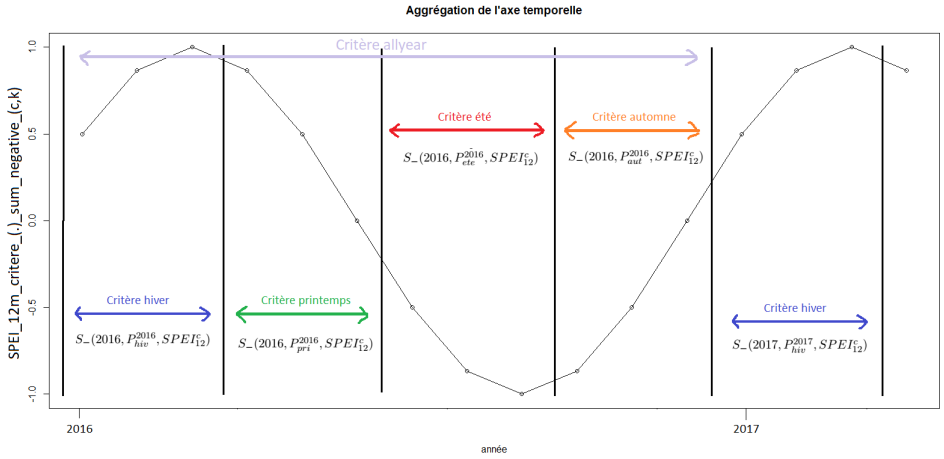

Figure 20 – Illustration des quantités calculées dans l'équation 23. Pour l'indice SPEI de la commune c et l'année t, calculé avec des données mensuelles à l'aide d'une fenêtre glissante de 12 mois, le critère hiver (respectivement printemps, été, automne et allyear) avec $(.) = hiver$ (respectivement $printemps$, ete, $automne$ et $allyear$) est calculé sur les intervalles bleus (respectivement vert, rouge, orange et gris).

En plus des quantités de l'équation 23, une autre quantité est définie par :

$$\forall t \in \mathcal{A},\ \forall P \in \mathcal{P}_t,\ Q(t, P, f_T^c) = \sum_{i \in P} I(f_T^c(i) < q_{0.05}(f_T^c, P_{all}^t)) \tag{24}$$

Avec I la fonction indicatrice et q la fonction quantile. L'idée de cette quantité est notamment de diminuer l'influence des valeurs extrêmes et de n'être sensible qu'au rang des valeurs. Ainsi, pour une année t donnée, Q est le nombre de points parmi les points de la période P qui sont inférieurs à la valeur correspondant au quantile 5% des $(f_T^c(i))_{i \in P_{all}^t}$. En notation dans les graphes, cette quantité est accompagnée par le suffixe $quan005$.

Par ailleurs les données de température minimale et maximale et les données de précipitation sont aussi présentes dans la base de données de catastrophe naturelle comme variables explicatives. Pour ces données, l'agrégation de l'axe temporel est réalisée respectivement à l'aide des fonctions minimum, maximum et somme. Les indices issus sont notés par exemple $Tmax\_critere\_automne\_max$ ou $P\_critere\_ete\_sum$.

En plus de ces variables, les données de satellite décrivant l'humidité du sol (SWI ou Soil Moisture Index) sur une épaisseur du sol de 20 cm, 40 cm, 60 cm et 100 cm ont pu être récupérées. Cet indice d'humidité est un indice modélisé et varie entre 0 et 1 avec 0 qui indique un sol aride et 1 un sol abondant en eau. Notez qu'il s'agit d'un indice qui tente d'approcher l'indice d'humidité du sol utilisé par Météo-France. Les notations liées à ces derniers sont par exemple $SWI\_040\_critere\_allyear\_quan005$.

De manière générale, les variables explicatives sont divisées en variables saisonnières, i.e calculées par rapport à une saison, et variables annuelles. À titre indicatif, la base de données comporte in fine plus de 700 indices de climat en tout sans compter les indices géologiques et autres. Notez que ce nombre est à comparer avec le nombre d'observations qui est d'environ $3 \times 35\ 000$ et que l'idée de départ est d'avoir la base la plus exhaustive possible. Le cœur de cette approche communale par la suite sera notamment de sélectionner parmi ce grand nombre de variables, les variables explicatives les plus pertinentes.

4.2.2.2 Krigeage ordinaire

Contrairement au cas précédent où chaque département est couvert par plusieurs points bleus (figure 2), dans le cas présent, une bonne partie des communes ne sont couvertes par aucun des points. Autrement dit, les données de climat pour ces communes ne sont pas disponibles, et par conséquent, pour garder une résolution spatiale communale, il faut estimer les données manquantes. Pour cela, la méthode de krigeage sera utilisée.

L'idée de la méthode de krigeage est de faire une interpolation géographique en partant des données climatiques de la grille (figure 2). D'un point de vue mathématique, si Z est l'indice et x_0 la position du point estimé, alors une estimation de $Z(x_0)$, notée $\widetilde{Z}(x_0)$, est donnée par :

$$\widetilde{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{25}$$

Avec n le nombre de sites où les valeurs de l'indice Z sont connues et λ_i les poids affectés à la valeur de Z en x_i par le krigeage. Notez que $Z(x_0)$ et les $Z(x_i)$ sont des séries temporelles et qu'en fonction des hypothèses de travail sur Z (stationnarité, moyenne, variance, ...), différents types de krigeage peuvent être utilisés. Ce qui est décrit ici et appliqué dans l'étude correspond à la méthode de krigeage dite ordinaire. Il s'agit de la méthode la plus communément utilisée.

Dans cette méthode, pour obtenir la valeur de l'indice au point x_0, Z il faut calculer les poids λ_i sous contrainte que $\sum_{i=1}^{n} \lambda_i = 1$ tout en minimisant la variance de l'écart σ_e définie par :

$$\begin{aligned}
\sigma_e^2 &= \mathrm{Var}(Z(x_0) - \widetilde{Z}(x_0)) \\
&= \mathrm{Var}(Z(x_0)) + \mathrm{Var}(\widetilde{Z}(x_0)) - 2\mathrm{Cov}(Z(x_0), \widetilde{Z}(x_0)) \\
&= \mathrm{Var}(Z(x_0)) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}(Z(x_i), Z(x_j)) - 2\sum_{i=1}^{n} \lambda_i \mathrm{Cov}(Z(x_0), Z(x_i))
\end{aligned} \tag{26}$$

Puis, pour résoudre ce problème mathématique, la méthode de Lagrange est utilisée en posant $\mathcal{L}$ qui est défini par :

$$\mathcal{L}(\lambda_1, \lambda_2, ..., \lambda_n, \mu) = \sigma_e^2 + \mu \left(\sum_{i=1}^{n} \lambda_i = 1 \right) \tag{27}$$

Et le système d'équation du krigeage ordinaire est donné par :

$$\begin{cases} \forall i = 1, 2, ..., n, \ 2\sum\limits_{i=1}^{n} \lambda_i \mathrm{Cov}(Z(x_i), Z(x_j)) - 2\mathrm{Cov}(Z(x_0), Z(x_i)) + \dfrac{\mu}{2} = 0 \\ \sum\limits_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{28}$$

Avec $\mathrm{Cov}(Z(x_0), Z(x_i))$ ou variogramme empirique estimé à partir des autres points connus ([44]). La résolution de ce système d'équation permet d'obtenir les poids λ_i et donc l'estimateur $\widetilde{Z}(x_0)$. Il suffit alors de réitérer ces calculs pour toutes les localisations x_0 pour obtenir toutes les données manquantes. Notez qu'une variante appelée krigeage simple existe et est tout aussi communément utilisée que le krigeage ordinaire. Ce dernier a notamment l'avantage de ne plus avoir de contrainte sur les poids. En contrepartie, il est nécessaire de connaître à l'avance les moyennes $\mathbf{E}(Z(x_i))$. Bien entendu, il y a aussi d'autres méthodes de krigeage qui utilisent d'autres hypothèses de départ comme le krigeage universel ou le krigeage IRFk.

Ainsi, à l'aide de cette méthode de krigeage ordinaire, les données de climat sont ramenées à la maille communale.

4.2.3 Variables géologiques

Comme pour la base de données de maille départementale, nous allons aussi construire des variables géologiques pour la base de données de maille communale. De même, ces variables sont issues des données de la BRGM ([3]) et sont définies dans le tableau de la figure 21.

Nom	Description
$lng(c)$	La longitude de la commune c.
$lat(c)$	La latitude de la commune c.
$rga_0(c)$	Le pourcentage de surface de la commune c classé en zone non argileuse par le BRGM.
$rga_1(c)$	Le pourcentage de surface de la commune c classé en zone de niveau d'aléa retrait-gonflement 1 par le BRGM.
$rga_2(c)$	Le pourcentage de surface de la commune c classé en zone de niveau d'aléa retrait-gonflement 2 par le BRGM.
$rga_3(c)$	Le pourcentage de surface de la commune c classé en zone de niveau d'aléa retrait-gonflement 3 par le BRGM.
$aleasol(c)$	Le niveau de risque selon la BRGM du point $(lng(c), lat(c))$.
$aleacirc5(c)$	Le niveau de risque pondéré dans un cercle de rayon $r = 5$ km centré en $(lng(c), lat(c))$ selon la BRGM.
$aleacirc10(c)$	Le niveau de risque pondéré dans un cercle de rayon $r = 10$ km centré en $(lng(c), lat(c))$ selon la BRGM.
$aleacirc20(c)$	Le niveau de risque pondéré dans un cercle de rayon $r = 20$ km centré en $(lng(c), lat(c))$ selon la BRGM.
$aleacirc50(c)$	Le niveau de risque pondéré dans un cercle de rayon $r = 50$ km centré en $(lng(c), lat(c))$ selon la BRGM.

Figure 21 – Tableau descriptif des variables géologiques construites à partir des données de la BRGM (figure 8, [3]) pour l'approche communale.

Dans ce tableau, les variables de niveau de risque pondéré sont définies tous de la même manière. Par exemple la variable $aleacirc5(c)$ est définie par :

$$aleacirc5(c) = \sum_{i=0}^{3} \frac{i.surf(i, c, r = 5)}{\pi.5^2} \tag{29}$$

Où $surf(i, c, r)$ est la surface en mètre carré classée au niveau de risque i (allant de 0 à 3) selon le BRGM (figure 8) dans le cercle de rayon $r = 5$ km centré en $(lng(c), lat(c))$. Les autres variables du même type sont donc définies avec comme rayon $r = 10$ km, $r = 20$ km et $r = 50$ km. Notez que ces variables géologiques servent notamment à caractériser la prédisposition naturelle des communes et de leurs sols au risque de subsidence.

4.2.4 Variables socio-économiques

Lorsque nous regardons l'évolution temporelle des critères pour qu'une commune soit décrétée en état de catastrophe naturelle dans la partie 5.2.1, nous constatons rapidement que ces critères sont loin d'être immuables. La commission interministérielle en charge du décret adapte fréquemment les critères en fonction de la situation comme montre par exemple l'introduction de l'indice SWI (Soil Wetness Index) de Météo France dans les critères ou encore la suppression en 2016 de la condition selon laquelle il faut que 10% de la commune soit touché par la sécheresse. De plus, des exceptions à la règle ont souvent eu lieu en période de sécheresse aggravée comme par exemple l'épisode de printemps 2011.

Nom	Description
nb_log	Nombre de logements au sein d'une commune donnée.
nb_rp	Nombre de résidences principales au sein d'une commune donnée.
nb_sec	Nombre de résidences secondaires au sein d'une commune donnée.
nb_vac	Nombre de logements vacants au sein d'une commune donnée.
nb_type	Nombre de logements (principaux, secondaires et vacants) de type maison ($type = maison$) ou appartement ($type = appart$) au sein d'une commune donnée.
rp_iP	Nombre de résidences principales avec i pièce où i vaut 1, 2, 3, 4 ou 5 pièces et plus (respectivement $iP = 1P, 2P, 3P, 4P, 5PP$).
np_piece_type	Nombre de pièces totales des logements du type maison ($type = maison$), appartement ($type = appart$) ou les deux ($type = both$).
nb_rp_type	Nombre de résidences principales de type maison ($type = maison$) ou appartement ($type = appart$).
nb_rp_achtot	Nombre de résidences principales d'âge d'au moins 2 ans.
$nb\_rp\_type\_periode$	ge des résidences principales de type maison ($type = maison$), appartement ($type = appart$) ou les deux ($type = both$) et construites durant les périodes 1919 à 1945, 1946 à 1970, 1971 à 1990, 1991 à 2005 ou 2006 à 2014 (respectivement $periode = 1945, 4670, 7190, 9105, 0614$).
rp_type_agemoy	ge moyennes des logements de type maison ($type = maison$), appartement ($type = appart$) ou les deux ($type = both$).
$prixM2$	Prix du marché du mètre carré moyen estimé au sein de la commune.
pop	Population de la commune.
pop_agemoy	ge moyenne de la population de la commune.
$superficie$	Superficie de la commune.
$densite$	Densité de population de la commune.

Figure 22 – Tableau descriptif des variables socio-économiques construites à partir des données de recensement de l'INSEE et des données de DVF de la Direction générale des Finances publiques.

Tout ceci témoigne que l'accord d'un décret de catastrophe naturelle n'est pas seulement une question de condition climatique mais aussi une question humaine et politique qui nécessite souvent la prise en compte des situations particulières de chaque commune. Ainsi, afin de tenir compte de cette dimension sociale et politique dans les modèles, il est nécessaire de compléter les variables climatiques et géologiques avec des variables relatives à la composition de la population et au paysage immobilier. Pour ce faire, les données de recensement de la population produites par l'INSEE en 2017 ([26]) sont utilisées. En effet, le recensement dans sa forme actuelle est en place depuis 2008 et les données récoltées concernent la composition de la population mais aussi les conditions des logements et notamment l'âge des constructions. Par ailleurs, les données de demandes des valeurs foncières (DVF, [25]) produites par la Direction générale des Finances publiques sous la direction du Ministère de l'économie, des finances et de la relance sont aussi sollicitées. Cette base DVF comprend les données issues des actes notariés de transaction immobilière intervenus depuis 2016 et contient notamment la localisation des biens échangés et les prix des transactions.

En utilisant la base DVF et la base de recensement, les 43 variables du tableau de la figure 22 sont définies afin d'essayer de capturer cette notion de pression politique. En effet, il se peut très bien que les habitants d'une commune ayant subi des dégâts de sécheresse importants, en raison de l'âge de la construction ou de la valorisation des biens par le marché, exercent davantage de pression sur leurs maires qui, par effet domino, font remonter cette pression jusqu'à la commission interministérielle en passant par les préfets des départements. Bien entendu, cette pression dépend à priori de la composition de population et du poids de la commune.

Notez ensuite que les données de l'INSEE correspondent à une photo prise en 2017, et qu'afin de

prolonger les données dans le temps dans les deux sens, des taux de croissance sont appliqués. D'après les données de l'INSEE ([28]), le taux de croissance du parc de logements métropolitain s'accroît de 1,1% par an en moyenne depuis trente ans. Ce taux est donc utilisé pour actualiser les données des logements. Par exemple pour une commune c, si la valeur de la variable $nb\_log(c,t)$ est connue pour $t = 2017$, alors pour t compris entre 2009 et 2020, la variable $nb\_log(c,t)$ est définie par :

$$nb\_log(c,t) = \frac{nb\_log(c,2017)}{(1+1.1\%)^{2017-t}} \tag{30}$$

De même, la même opération est faite pour la population avec des taux de croissance par département qui sont aussi estimés par l'INSEE (sur 5 ans de données, [27]). Il aurait fallu faire la même chose pour les prix du mètre carré, i.e les actualiser par l'inflation afin d'avoir une vision à terme dans les prix d'aujourd'hui mais cela ne sera pas utile. En effet, les données de DVF concernent les transactions réalisées et par conséquent, le nombre de transactions pour certaines communes est trop peu pour avoir ne serait-ce qu'un prix de marché moyen. En d'autres termes, avant de penser à l'actualisation, il faut d'abord estimer les prix pour les communes où il n'y a pas eu de transactions puis faire un lissage géographique. Et dans la mesure où les erreurs liées à ces estimations sont à priori plus grandes que celles liées à l'ajustement par l'inflation, l'opération d'ajustement peut simplement être négligée.

Ainsi, un zonier pour le prix du mètre carré est construit comme précédemment dans l'approche départementale pour le taux de destruction. La variable notée $prixM2$ est donc le résultat d'un zonier et non pas une moyenne calculée à partir des données extraites des DVF. En d'autres termes, pour une commune c et $\mathcal{V}_d$ ses d plus proches communes voisines, la variable $prixM2$ est définie par :

$$\begin{aligned} \forall t = 2009, ..., 2020,\ prixM2(c,t) &= prixM2(c,2007) \\ &= \frac{1}{|\mathcal{V}_d|} \sum_{k \in \mathcal{V}_d} prixM2_{raw}(k,2007) \end{aligned} \tag{31}$$

Avec $prixM2_{raw}$ le prix moyen du mètre carré au sein d'une commune calculé directement à partir des prix de transaction issus de la base DVF. Il faut noter que pour certaines communes, notamment les communes peu attractives, ce prix est nul, d'où l'intérêt de faire un zonier pour avoir tout de même une estimation du prix malgré l'absence des transactions.

Par ailleurs, pour obtenir la variable d'âge moyenne des logements, nous utilisons celles sur l'âge des constructions (maison et/ou appartement) par période en prenant l'année moyenne pondérée par le nombre de constructions. En d'autres termes, en notant $s_i = 1917, 1946, 1971, 1991, 2006$ l'année du début de la période, $f_i = 1945, 1970, 1990, 2005, 2014$ l'année de fin de la période et c la commune, nous avons par exemple pour rp_both_agemoy :

$$\forall t = 2009, ..., 2020, \quad rp\_both\_agemoy(c,t) = \frac{1}{2} \frac{\sum\limits_{(s_i,f_i)} (f_i - s_i) nb\_rp\_both\_s_if_i(c,t)}{\sum\limits_{(s_i,f_i)} nb\_rp\_both\_s_if_i(c,t)} \tag{32}$$

Avec $(f_i - s_i)/2$ l'âge moyenne associée à la période (s_i, f_i) et les poids $nb\_rp\_both\_s_if_i$ définis dans le tableau de la figure 22.

4.3 Méthodes de sélection des variables pour le modèle d'occurrence

Notez d'abord qu'avec les variables définies précédemment, la base de données de catastrophe naturelle possède plus de 800 variables explicatives dont plus de 700 sont des variables climatiques. L'objectif est de construite à partir de cette base, un modèle qui permet de prédire pour une nouvelle année, en observant les conditions de climat subies par les communes et compte tenu de leurs situations géologiques et socio-économiques, si une commune sera décrétée en situation de catastrophe naturelle par la commission interministérielle.

Il est clair que dans le perspective de construction d'un produit paramétrique d'assurance-réassurance,

le modèle final doit être simple et comporter peu de variables explicatives. De plus, il faut noter que les variables climatiques, par construction, sont très corrélées entre eux et par conséquent, il n'est pas nécessaire de toutes les garder. Ainsi, avant de construire le modèle de prédiction d'arrêté de catastrophe naturelle, ou modèle d'occurrence, le nombre de variables et notamment les variables climatiques doit être réduit.

Par la suite, les variables explicatives sont d'abord filtrées par une méthode ensembliste, puis par une méthode basée sur des distances entre deux distributions. L'idée est que la méthode ensembliste est plus robuste mais moins spécifique et qu'il faut donc affiner la sélection à l'aide de la méthode basée sur des distances qui elle sélectionne davantage les variables pertinentes pour le problème considéré.

4.3.1 Sélection ensembliste

Dans un premier temps, pour réduire le nombre de variables climatiques, une méthode de sélection des variables ensemblistes est utilisée. Le principe d'une telle méthode est de combiner plusieurs métriques, construites à partir des hypothèses initiales différentes, pour obtenir un score pour chaque variable explicative permettant ainsi de classer ces dernières. L'idée de la méthode est qu'une variable jugée comme intéressante par tous les métriques est plutôt à considérer dans la construction d'un modèle tandis qu'une variable évaluée comme intéressante pour certaines métriques seulement aura un score final plus faible. En d'autres termes, chaque méthode de classification a à priori ses propres biais de sélection et une méthode ensembliste consiste à utiliser cette variance du biais en fonction de la méthode de classification de telle sorte que les biais se contre-balancent en somme. Ainsi, dans une méthode ensembliste, il est nécessaire que les méthodes de classifications sous-jacentes soient fondamentalement différentes afin de pouvoir tirer profit de la diversification.

La méthode ensembliste utilisée dans cette étude est développée par Genze N., Neumann U. et al. ([29]) et met en jeu huit métriques qui ont déjà montré leurs intérêts individuellement dans la littérature et qui se regroupent en quatre catégories. Par la suite, nous allons décrire brièvement ces métriques. Notez aussi que la variable réponse est binaire (1 pour la reconnaissance de catastrophe naturelle et 0 sinon) et que les variables climatiques sont réelles et continues.

4.3.1.1 Médiane

La première métrique dite sélection des variables par la médiane consiste à calculer d'abord, pour une variable climatique X donnée, sa médiane conditionnellement à $Y = 1$, notée med_1, et sa médiane conditionnellement à $Y = 0$, notée med_0 avec Y la variable réponse. Un test de Wilcoxon-Mann-Whitney ou test U de Mann-Whitney est ensuite réalisé. Notez qu'il s'agit d'un test statistique non paramétrique proposé par Frank Wilcoxon en 1945 et par Henry Mann et Donald Ransom Whitney en 1947. À supposer que les observations sont indépendantes et qu'il existe une relation d'ordre, les hypothèses de test sont :

$$\begin{cases} H_0 : & med_0 = med_1 \\ H_1 : & med_0 \neq med_1 \end{cases} \tag{33}$$

La statistique de test notée U, d'où le nom test U, est définie ici par :

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} Q(x_i(0), x_j(1)) \tag{34}$$

Avec :

$$Q(x_i(0), x_j(1)) = \begin{cases} 1 \text{ si } x_j(1) < x_i(0) \\ \frac{1}{2} \text{ si } x_j(1) = x_i(0) \\ 0 \qquad \text{sinon} \end{cases} \tag{35}$$

Où n est le nombre d'observations de $X|Y=0$, m le nombre d'observations de $X|Y=1$, $x_i(0)$ une observation de $X|Y=0$ et $x_i(1)$ une observation de $X|Y=1$. Pour des échantillons de taille supérieure à 20, ce qui est clairement le cas ici, la statistique U, sous l'hypothèse nulle, suit une distribution connue qui peut être approchée par une loi normale $\mathcal{N}(\mu, \sigma^2)$ avec μ la moyenne et σ l'écart-type de la loi normale définis par :

$$\mu = \frac{nm}{2} \quad \text{et} \quad \sigma^2 = \frac{nm(n+m+1)}{12} \tag{36}$$

Ainsi, le p-value du test U est ensuite utilisé comme valeur issue de cette méthode de sélection par la médiane. Remarquez que le p-value est compris entre 0 et 1, ce qui est important à noter lorsqu'il faudra combiner différents scores.

4.3.1.2 Coefficient de corrélation

Les deux prochaines métriques sont basées sur le calcul d'un coefficient de corrélation et résulte d'un papier de Yu et Liu ([30]). L'idée est de sélectionner une variable X_i si elle est corrélé à la variable réponse Y et peu corrélé aux autres variables X_j où $j \neq i$. Les auteurs de la méthode ensembliste fixe le seuil de corrélation entre variable X_i et X_j à 0.7 et les variables ne satisfaisant pas ces conditions auront une importance nulle. Les autres auront une importance égale à leurs corrélations avec Y en valeur absolue.

Dans cette démarche, les coefficients de corrélation de pearson et de spearman sont utilisés. Notez que le coefficient de corrélation de pearson entre X et Y, noté r_{XY}, est défini par :

$$r_{XY} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}} \tag{37}$$

Avec les x_j les réalisations de X, les y_j les réalisations de Y et n le nombre de réalisations.

De manière similaire, le coefficient de spearman, ρ_{XY}, avec les même notations, est défini par :

$$\rho_{XY} = 1 - 6\sum_{j=1}^{n} \frac{(\mathrm{rg}(x_i) - \mathrm{rg}(y_i))^2}{n(n^2-1)} \tag{38}$$

Où $\mathrm{rg}(x_i)$ et $\mathrm{rg}(y_i)$ sont respectivement le rang de x_i et y_i. Ainsi, r_{XY} et ρ_{XY} sont les deux scores issus de la sélection par coefficient de corrélation. Remarquez aussi que ces coefficients de corrélation sont compris entre -1 et 1.

4.3.1.3 Régression logistique

La quatrième métrique s'obtient en utilisant une régression logistique. Il s'agit d'abord de normaliser les variables explicatives X_i et réaliser ensuite la régression de Y sur l'ensemble des variables normalisées X_i^N où :

$$X_i^N = \frac{X_i - \overline{X_i}}{\sigma_i} \tag{39}$$

Avec $\overline{X_i}$ et σ_i respectivement la moyenne et l'écart-type de X_i. Les coefficients de régression β_i des variables X_i^N en valeur absolue correspondent aux mesures d'importance recherchées. Notez que le fait de normaliser les X_i rend les coefficients de régression comparables aux autres métriques d'importance.

4.3.1.4 Forêt aléatoire

Les quatre métriques restantes sont définies à partir du modèle de forêt aléatoire dont la construction est basée sur des arbres de décision. Nous n'allons pas reprendre l'ensemble de la théorie sur les arbres

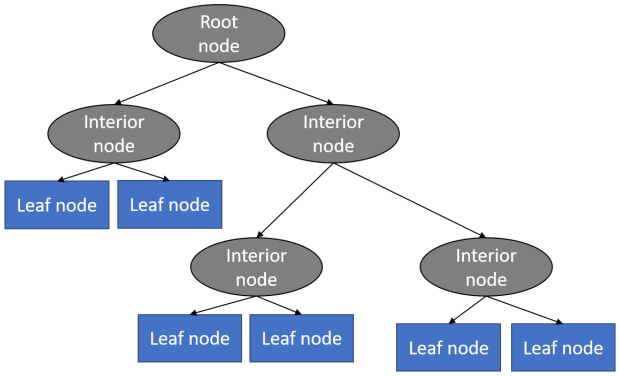

Figure 23 – Illustration d'un arbre de décision avec le vocabulaire associé (extraite du web).

de décision mais simplement rappeler le principe général. Dans la figure 23, une illustration d'un arbre de décision avec le vocabulaire associé est présentée.

Un arbre de décision est construit nœud par nœud du haut en bas à partir du nœud racine jusqu'aux feuilles. Au début du processus, une métrique de construction, qui sera maximisée ou minimisée lors du processus de construction, est fixée et un ensemble de données d'entraînement, qui seront classés au fur et à mesure de la construction, est donné. Les données sont initialement présentes dans le nœud racine et à chaque nœud est associé une valeur de la métrique de construction. Pour construire les deux nœud fils suivants à partir du nœud racine ou nœud père, toutes les variables disponibles sont testées et la variable qui optimise la métrique de construction est sélectionnée. Si les variables sont numériques, il faut en plus tester différents seuils qui séparent les données présentes dans le nœud père en deux groupes de manière optimale. La variable sélectionnée sera ensuite utilisée pour calculer les valeurs de métrique associées aux nœud fils, répartissant à la même occasion les données du nœud père dans ces deux derniers. Et ce processus continue tant qu'il reste des variables explicatives non utilisées et qu'il est possible de continuer de minimiser ou maximiser la métrique de construction. En pratique, la métrique de construction est souvent l'indice de Gini (qui n'a rien à voir avec l'indice de Gini en économie qui mesure l'inégalité entre pays) et des critères d'arrêt sont précisés au début pour interrompre la construction des arbres.

Il faut noter que pour construire un arbre de décisions, il est possible d'utiliser seulement une partie des données et un sous-ensemble de variables explicatives. Et c'est ce que fait une forêt aléatoire qui est un ensemble de T arbres de décision. Chaque arbre de décision est construit individuellement en choisissant aléatoirement un sous-ensemble de données et un sous-ensemble de variables explicatives. Pour chaque arbre t, les données non utilisées lors de sa construction sont appelées données OOB (out-of-bag) et notées $B(t)$. Les mesures d'importance sont définies à l'échelle d'un arbre de décision et l'importance finale correspond à une moyenne sur les arbres. Si $VI_{X_i}(t)$ est l'importance de X_i calculée sur l'arbre t, alors l'importance finale de X_i, notée $\widehat{VI_{X_i}}$, est :

$$\widehat{VI_{X_i}} = \frac{1}{T}\sum_{t=1}^{T} VI_{X_i}(t) \tag{40}$$

Une première mesure d'importance est celle associée à l'indice de Gini. Cette dernière est définie par :

$$VI_{X_i}(t) = \sum_{j \in J} d_{ij} I(X_i, j) \tag{41}$$

Avec les d_{ij} définis par :

$$d_{ij} = G_j - \left(\frac{N_L(j)}{N_j} G_L(j) + \frac{N_R(j)}{N_j} G_R(j)\right) \tag{42}$$

Où :

$$G_j = 1 - \sum_{k=1}^{2} \frac{N_k(j)}{N_j} = 2\frac{N_1(j)}{N_j}\left(1 - \frac{N_1(j)}{N_j}\right) \tag{43}$$

Dans ces expressions, J est l'ensemble des nœuds de l'arbre t, N_j le nombre de données dans le nœud j, $N_1(j)$ le nombre de données de classe 1 parmi les N_j données, $N_L(j)$ et $N_R(j)$ les nombres de données distribuées dans les deux nœuds fils depuis le nœud père j, $G_L(j)$ et $G_R(j)$ les indices de Gini des deux nœuds fils, et $I(X_i, j)$ un indicatrice qui vaut 1 si X_i est la variable sélectionnée au nœud j et 0 sinon. Notez que cette mesure d'importance correspond globalement à mesurer à quel point la variable X_i contribue dans la réduction de l'indice de Gini.

Une deuxième mesure d'importance est basée sur l'AUC (l'aire sous la courbe ROC) de l'arbre t sur les données OOB, i.e $B(t)$, et est définie par :

$$VI_{X_i}(t) = AUC(t) - AUC_{\pi_i}(t) \tag{44}$$

Avec $AUC_{\pi_i}(t)$ calculé sur les données OOB mais dont la colonne X_i a subi une permutation π_i. L'idée est qu'en permutant les valeurs de X_i, si ce dernier est crucial dans la prédiction de la variable réponse Y par l'arbre t, alors une chute d'AUC conséquente sera observée.

Les deux dernières métriques d'importance sont quasiment les mêmes et sont basées sur l'exactitude de la prédiction réalisée par l'arbre t sur les données OOB, i.e $B(t)$. Elle sont définies par :

$$VI_{X_i}(t) = \frac{1}{|B(t)|} \sum_{j \in B(t)} (I(y_j = p_j) - I(yi = p_{j,\pi_i})) \tag{45}$$

Et :

$$VI_{X_i}(t) = \frac{1}{|B(t)|} \sum_{j \in B(t)} (I(y_j = p_j) - I(yi = p_{j,\pi_i|Z})) \tag{46}$$

Avec $I(y_j = p_j)$ l'indicatrice valant 1 si la prédiction de l'arbre t, p_j, est égale à la valeur réel y_j et $I(yi = p_{j,\pi_i})$ la même indicatrice mais où les valeurs de X_i dans les données OOB ont été permutées par π_i. L'indicatrice $I(yi = p_{j,\pi_i|Z})$ est semblable à $I(yi = p_{j,\pi_i})$ sauf qu'en plus de la permutation, il y a un conditionnement par Z où $Z = X_j$ avec $j \neq i$.

4.3.1.5 Score final

Ainsi, nous avons défini les huit métriques utilisées dans la méthode de sélection de variable ensembliste. Par la suite, afin de rendre tous ces métriques comparables, les auteurs proposent de normaliser les valeurs, i.e si $\|X_i\|_K$ est l'importance de la variable X_i selon la métrique K, alors l'importance normalisée $Imp(X_i, K)$ est :

$$Imp(X_i, K) = \frac{\|X_i\|_K}{\max_j \|X_j\|_K} \tag{47}$$

Sauf dans le cas de la métrique de médiane où :

$$Imp(X_i, K) = 1 - \|X_i\|_K + \min_j \|X_j\|_K \tag{48}$$

Ainsi, le score final S de la méthode ensembliste pour la variable explicative X_i est donné par :

$$S(X_i) = \sum_K Imp(X_i, K) \tag{49}$$

Bien entendu, ici, les différentes métriques ainsi que les idées générales des auteurs sont décrites que très brièvement. Les lecteurs intéressés pourront aller voir plus en détails l'article associé ([29]).

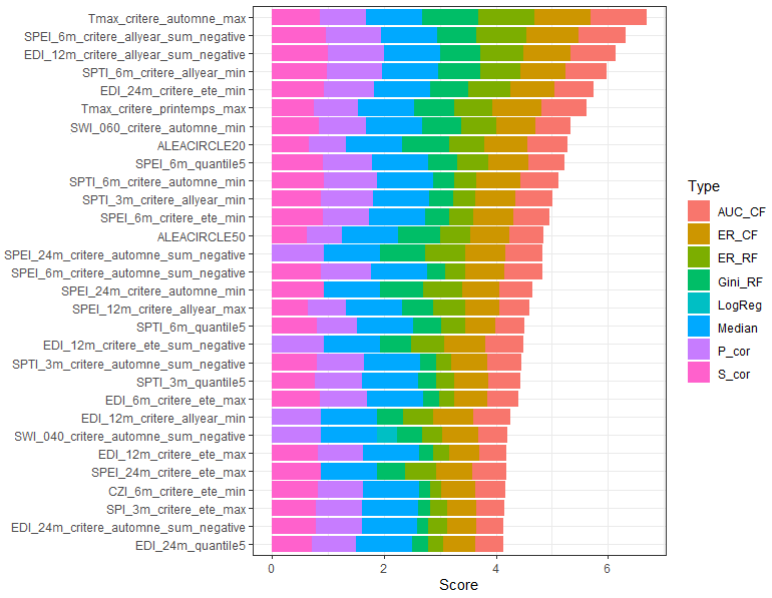

Figure 24 – Résultat de la méthode de sélection de variable ensembliste ([29]). Seules les 30 variables les plus importantes selon l'algorithme ensembliste sont représentées. Pour ces variables, l'importance selon la régression logistique est négligeable. Notez qu'en général, c'est surtout les variables climatiques avec une large fenêtre de calcul qui se distinguent, ce qui semble logique puisque la sécheresse est un phénomène étalé dans le temps.

Dans la figure 24, les 30 variables les plus importantes selon l'algorithme ensembliste sont représentées. Pour ces variables, l'importance selon la régression logistique est négligeable. Il faut noter qu'en général, c'est surtout les variables climatiques avec une large fenêtre de calcul qui se placent en première, ce qui semble logique puisque la sécheresse est un phénomène étalé dans le temps. Ainsi, une première sélection des variables consiste à ne retenir que les 150 variables climatiques les plus importantes parmi les plus de 700 variables climatiques initiales. Notez que cette méthode est disponible sur **R** via la fonction $ensemble\_fs()$ de la librairie EFS.

4.3.2 Sélection par distance statistique

La sélection ensembliste des variables permet de passer de plus de 800 variables à un peu plus de 200 variables (150 variables climatiques, les autres sont des variables géologiques, socio-économiques, ou issues des contrats). Cependant, ce nombre reste conséquent et il faut pousser davantage la sélection en intégrant cette fois les variables non climatiques. Idéalement, il faut garder uniquement une trentaine de variables tout en ayant un maximum de diversité dans les variables retenues. En effet, cela garantit notamment que le modèle est à la fois simple par le peu de variables utilisées mais aussi riche car tous les phénomènes identifiés précédemment sont pris en compte.

4.3.2.1 Statistique de Kolmogorov-Smirnov et statistique de Kuiper

Pour sélectionner davantage les variables, des critères de distance sont définis. L'idée que adoptée est que si y_{occ} est la variable réponse binaire et X une variable explicative, alors plus la distribution de $X|(y_{occ} = 0)$ est différente de la distribution de $X|(y_{occ} = 1)$, plus la variable X est intéressante. Plus précisément, nous allons utiliser des statistiques de test déjà présentent dans la littérature comme métrique. Le principe statistique derrière est que nous souhaitons tester l'hypothèse nulle qui est que la distribution de $X|(y_{occ} = 0)$ est la même que la distribution de $X|(y_{occ} = 1)$.

Soient E la distribution de $X|(y_{occ} = 0)$, F la distribution de $X|(y_{occ} = 1)$ et D_X l'ensemble des réalisations de X. Alors la statistique de test de Kolmogorov-Smirnov, notée KS, et la statistique de

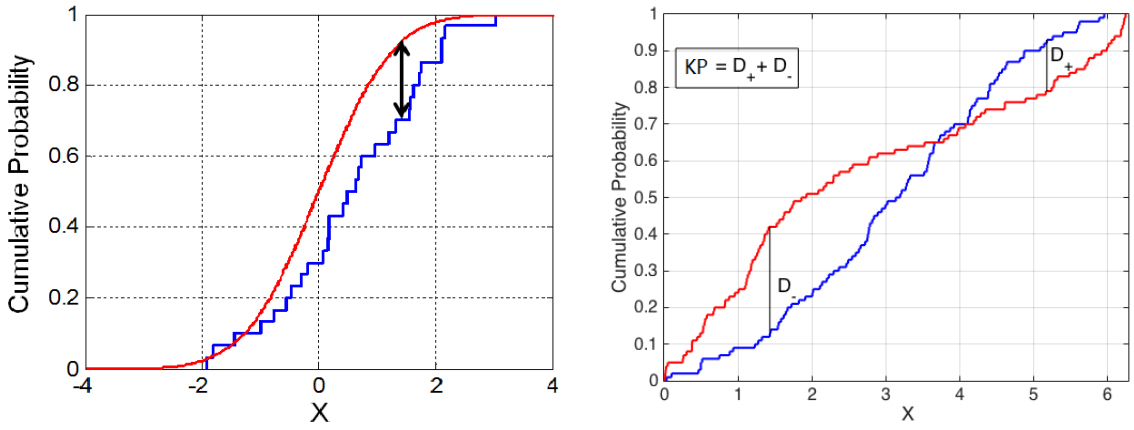

Figure 25 – À gauche, il s'agit d'une illustration extraite du web représentant deux distributions (rouge et bleue) et la statistique de Kolmogorov-Smirnov (noir) comme distance maximale entre les deux distribution. À droite, il s'agit d'une illustration extraite du web représentant deux distributions (rouge et bleue) et la statistique de Kuiper (noir) qui correspond à une légère modification de la première de la statistique de test de Kolmogorov-Smirnov.

test de Kuiper, notée KP, avec la seconde qui correspond à une légère modification de la première, sont définies par ([35]) :

$$\begin{cases} KS(X) = \max\limits_{x \in D_X} |F(x) - E(x)| \\ KP(X) = |\max\limits_{x \in D_X} E(x) - F(x)| + |\max\limits_{x \in D_X} F(x) - E(x)| \end{cases} \tag{50}$$

L'idée de la modification apportée par Kuiper est d'introduire une certaine sensibilité à la queue de distribution qui n'est pas présente dans la statistique de Kolmogorov-Smirnov. Une illustration de ces deux statistiques est représentée dans la figure 25.

4.3.2.2 Statistique de Cramér-von Mises et distance de Wasserstein

Dans le même registre des statistiques de test sensibles à la queue de distribution, avec les même notations que précédemment, la statistique de test de Cramér-von Mises, notée CVM, et la distance de Wasserstein, notée WS, sont définies par ([35]) :

$$\begin{cases} CVM(X) = \sum\limits_{x \in D_X} |F(x) - E(x)|^2 \\ WS(X) = \left(\int_{x \in \mathbb{R}} |F(x) - E(x)|^p dx\right)^{1/p} \end{cases} \tag{51}$$

Dans l'étude, p est fixé à 1 pour la distance de Wasserstein. Remarquez que la différence majeure entre la distance de Wasserstein et la statistique de test de Cramér-von Mises, mis à part l'élévation au carré dans la formule, est que la première permet aux observations extrêmes d'avoir plus de poids tandis que la seconde traite l'ensemble des observations avec la même pondération. La figure 26 illustre la statistique de test de Cramér-von Mises et la distance de Wasserstein. La première est représentée par la somme des hauteurs des barres noires au carré sur l'illustration de gauche et la seconde correspond à l'aire en noire sur l'illustration de droite.

4.3.2.3 Statistique d'Anderson–Darling

Une autre extension de CVM est la statistique de test d'Anderson–Darling, notée AD, et est définie par :

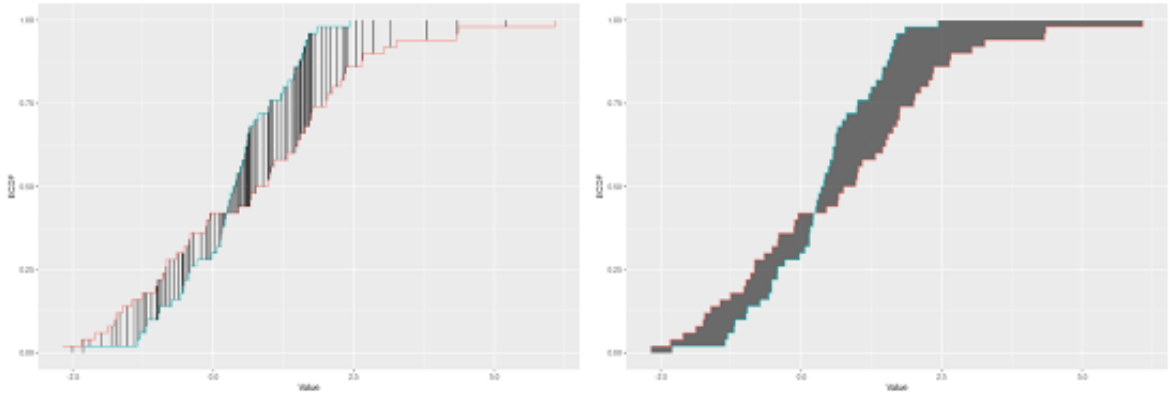

FIGURE 26 – À gauche, il s'agit d'une illustration (extraite de [35]) représentant deux distributions $\mathcal{N}(0,1)$ et $\mathcal{N}(0.5,4)$ et la statistique de Cramér-von Mises qui correspond à la somme des hauteurs des barres noires au carré. À droite, il s'agit d'une illustration (extraite de [35]) représentant les mêmes distributions et la distance de Wasserstein qui correspond à l'aire en noire.

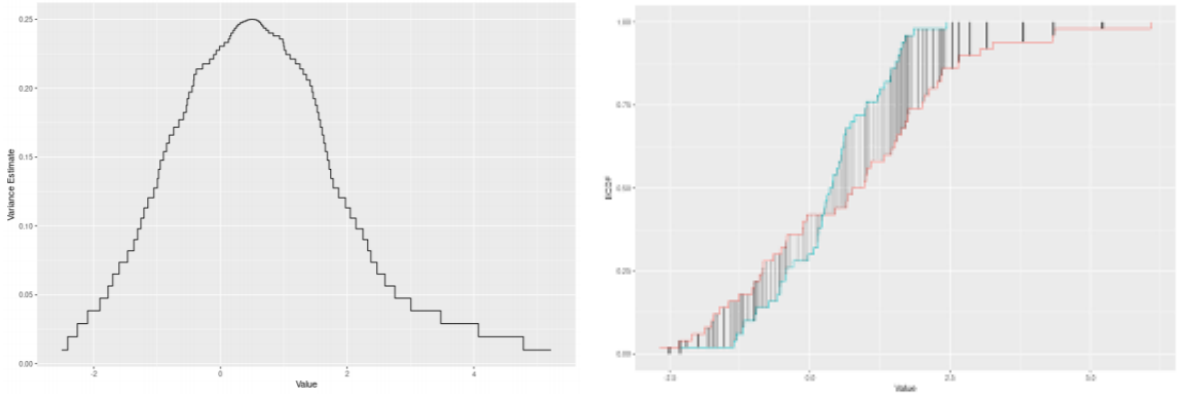

FIGURE 27 – Toujours pour les deux distributions $\mathcal{N}(0,1)$ et $\mathcal{N}(0.5,4)$, la variance de $F(x)-E(x)$ en fonction de x est représentée à gauche et la statistique de test d'Anderson–Darling qui correspond à la somme des hauteurs des barres noires au carré avec la pondération en niveau de gris est donnée à droite (figures extraites de [35]).

$$AD(X) = \sum_{x \in D_X} \frac{|F(x) - E(x)|^2}{G(x).(1-G(x))} \tag{52}$$

Avec G la distribution de X sans conditionnement. L'idée de cette extension est que la variance de $F(x)-E(x)$ dépend de x, et que pour corriger cela, il faut ajouter une fonction de pondération.

Dans la figure 27 extraite de [35], nous représentons l'évolution de la variance de $F(x)-E(x)$ en fonction de x pour les deux distributions $\mathcal{N}(0,1)$ et $\mathcal{N}(0.5,4)$ ainsi que la statistique de test d'Anderson–Darling qui correspond à la somme des hauteurs des barres noires au carré avec la pondération en niveau de gris.

4.3.2.4 Statistique DTS

La dernière statistique présentée est la statistique de test DTS, notée DTS. Il s'agit d'une statistique combinant la distance de Wasserstein pour prendre en compte la notion de distance, et la statistique de test d'Anderson–Darling pour prendre en compte le problème de la variabilité de la variance. La statistique de DTS est donc définie par ([35]) :

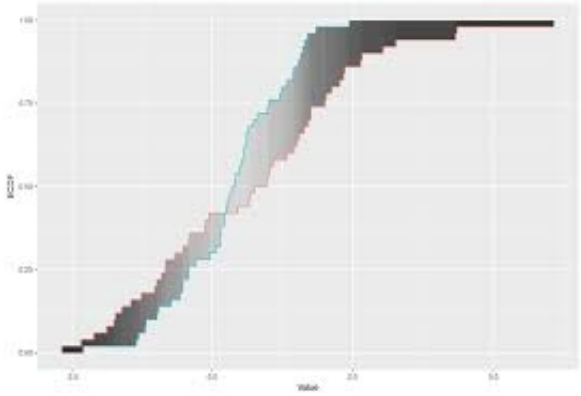

Figure 28 – Toujours pour les deux distributions $\mathcal{N}(0,1)$ et $\mathcal{N}(0.5,4)$, nous représentons la statistique de test de DTS qui correspond à l'aire en noire entre les deux distributions avec la pondération en niveau de gris (extraite de [35]).

$$DTS(X) = \int_{x\in\mathbb{R}} \frac{|F(x) - E(x)|}{G(x).(1-G(x))} dx \tag{53}$$

Dans la figure 28, une illustration (extraite de [35]) de cette statistique avec statistique DTS, qui correspond à l'aire en noire entre les deux distributions avec la pondération en niveau de gris, est présentée. Il est intéressant de noter que d'après l'auteur, les tests statistiques basés sur le DTS tendent à produire des résultats semblables voire meilleurs comparés aux autres statistiques quel que soit le type de différence entre distributions (dans la queue, au niveau de la moyenne, ...). Ce qui est très intéressant puisque cela montre que le DTS semble être très robuste alors que les autres statistiques tendent à fonctionner seulement pour certains types de différence uniquement.

4.3.2.5 Classement

Pour sélectionner les variables explicatives, les six statistiques définies précédemment sont utilisées. Plus exactement, nous allons nous inspirer de la méthode ensembliste et classer les variables en fonction d'un score total. Pour cela, les six statistiques sont d'abord normalisées, i.e :

$$\begin{cases} \overline{KS}(X) = \dfrac{KS(X)}{\max\limits_i KS(X_i)} \\ \overline{KP}(X) = \dfrac{KP(X)}{\max\limits_i KP(X_i)} \\ \overline{CVM}(X) = \dfrac{CVM(X)}{\max\limits_i CVM(X_i)} \\ \overline{WS}(X) = \dfrac{WS(X)}{\max\limits_i WS(X_i)} \\ \overline{AD}(X) = \dfrac{AD(X)}{\max\limits_i AD(X_i)} \\ \overline{DTS}(X) = \dfrac{DTS(X)}{\max\limits_i DTS(X_i)} \end{cases} \tag{54}$$

Avec les X_i qui sont les variables explicatives. Le score S est ensuite défini pour tout variable explicative X_i par :

$$S(X_i) = \overline{KS}(X_i) + \overline{KP}(X_i) + \overline{CVM}(X_i) + \overline{WS}(X_i) + \overline{AD}(X_i) + \overline{DTS}(X_i) \tag{55}$$

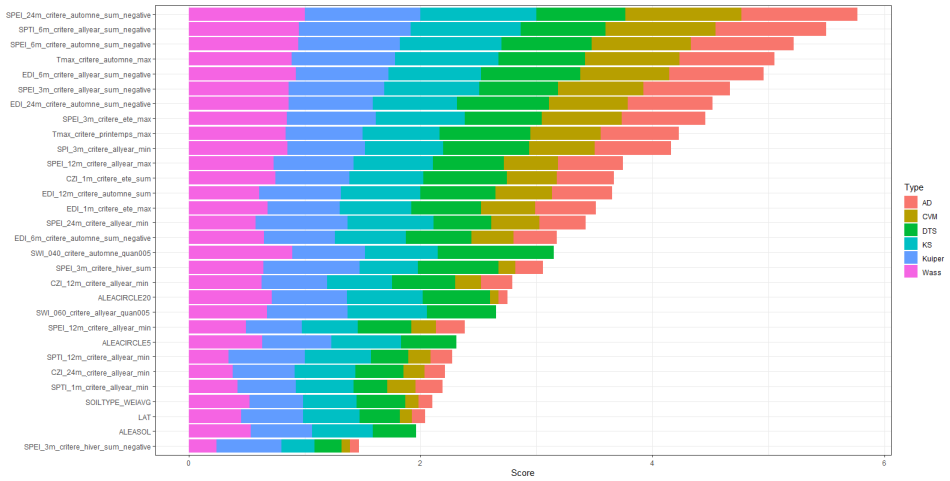

Figure 29 – Les 30 variables explicatives les plus importantes selon le score S. Ces variables composeront le modèle d'occurrence.

Ainsi, en s'inspirant de la méthode de sélection ensembliste, les variables explicatives sont classées à l'aide du score S définie dans l'équation 55. La figure 29 représente les 30 variables sélectionnées par le score et donc utilisées dans cette étude pour le modèle d'occurrence.

Notez que de manière générale, les six statistiques sont décroissantes lorsque nous passons de la variable explicative avec le plus haut score S à la variable avec le plus bas score. Ceci semble être cohérent puisque les statistiques mesurent plus ou moins la même chose, à savoir la différence entre deux distributions.

Parmi les variables explicatives sélectionnées, notez qu'il y a surtout des variables climatiques et quelques variables géologiques. Et parmi les variables climatiques, notez qu'il y a surtout des variables avec une fenêtre de calcul de 3 mois, 6 mois et 12 mois. Les variables climatiques avec une fenêtre de calcul de 1 mois et de 24 mois sont minoritaires malgré le fait qu'il s'agit d'une de ces variables qui a le meilleur score. Les variables socio-économiques, quant à elles, sont absentes de la liste des variables sélectionnées. Il est aussi intéressant de noter que le maximum de la statistique DTS, qui est à priori la plus robuste parmi les six statistiques, est atteint pour l'indice SWI. Nous reviendrons plus tard sur ce dernier point dans la partie 5.2.2 mais au final, ceci témoigne en faveur de la puissance de la statistique DTS.

Par la suite, nous allons construire le modèle d'occurrence et nous reviendrons plus tard sur les remarques et notamment la pertinence des variables sélectionnées au regard des résultats que nous allons obtenir (section 4.6.3).

4.4 Construction du modèle d'occurrence

À présent que les 30 variables explicatives ont été sélectionnées, il est possible de construire le modèle d'occurrence, i.e pour une commune donnée, modéliser la probabilité que cette commune soit décrétée en situation de catastrophe naturelle par la commission interministérielle. Il faut noter que parmi les variables sélectionnées, il y a surtout des variables climatiques et quelques les variables géologiques. Les variables socio-économiques ont toutes été filtrées. Cette liste de variables sélectionnées sera discutée plus loin lors de la section 4.6.3.

Par la suite, nous allons construire un ensemble de modèles candidats pour le modèle d'occurrence et notamment choisir les hyperparamètres de ces derniers. Pour les modèles candidats, six types de modèles différents sont considérés. Bien entendu, la liste est loin d'être exhaustive mais contient néanmoins les modèles les plus prometteurs. En effet, ce sont soit des modèles qui ont déjà fait leurs preuves comme le GLM, soit des modèles en plein développement comme les réseaux de neurones. Les six types de modèles sont :

- Generalised Linear Model (GLM) : modèle linéaire généralisé avec une distribution binomiale, un lien *logit*, et la possibilité d'ajouter des termes de régularisation.
- Gradient Boosting Machine (GBM) : modèle basé sur la descente du gradient et qui est composé d'un ensemble de modèles secondaires.
- Extreme Gradient Boosting (XGBOOST ou XGB) : modèle correspondant à une version régularisée du modèle de GBM.
- Random Forest (RF) : modèle de forêt aléatoire composé d'un ensemble d'arbres de décision.
- Extreme Random Forest (XRF) : modèle de forêt aléatoire mais utilisant des arbres de décision construits avec un degré supplémentaire d'aléa.
- Deep learning (DL) : modèle de réseau de neurone avec plusieurs couches cachées et des fonctions d'activation plus ou moins complexes.
- Ensemble Model (SE) : modèle GLM utilisant les résultats de tous les autres modèles comme variables explicatives.

Ces modèles seront présentés plus en détails dans cet ordre mais d'ores et déjà, il faut préciser que ces modèles possèdent de nombreuses déclinaisons possibles et que les définitions données par la suite correspondent à des définitions adaptées au projet et non des définitions générales.

4.4.1 Méthode de recherche sur grille

Bien entendu, tous les types modèles possèdent des hyperparamètres qu'il est nécessaire d'optimiser. Pour faire cela, pour chaque type de modèle, nous utilisons la méthode de recherche sur grille (Grid Search en anglais) qui consiste à rechercher de manière aléatoire l'espace des hyperparamètres jusqu'à trouver les hyperparamètres qui optimisent une métrique de sélection prédéfinie. La figure 30 montre une illustration de cette méthode dans le cas où il n'y a que deux hyperparamètres à optimiser. Bien entendu, les limites et la résolution de la grille, donc par conséquent le nombre de modèles possibles, doivent être définies au préalable.

Sur la figure 30, sur chaque point de la grille, un modèle est calculé en utilisant le couple d'hyperparamètres (x_1, x_2) du point sélectionné. Le modèle optimal est alors obtenu en choisissant le point, i.e le couple (x_1, x_2), qui optimise la métrique de sélection. Néanmoins, il est important de noter que la méthode de recherche sur grille ne calcul pas l'ensemble des modèles possibles car cela nécessite souvent beaucoup de ressources informatiques. L'algorithme de recherche calcule plutôt un petit nombre, spécifié en amont, de modèles sélectionnés aléatoirement dans l'espace des hyperparamètres. Le modèle final est celui qui optimise le critère de sélection parmi ce petit nombre de modèles.

Notez que dans l'étude, en fonction du type de modèle, le nombre d'hyperparamètres à optimiser dépasse souvent deux. Par ailleurs l'AUC, i.e l'aire sous la courbe ROC, est considérée comme métrique de sélection des hyperparamètres. Ainsi, pour chaque type de modèle, la méthode de recherche sur grille cherche le modèle qui maximise l'AUC parmi les modèles calculés. Compte tenu des ressources informatiques disponibles, le nombre de modèles calculés dans cette étude est fixé à au plus 300 par type de modèle. En d'autres termes, si la grille des hyperparamètres a une résolution ou des limites définies de telle sorte que le nombre de modèles possibles est plus grand que 300, alors seuls 300 modèles, sélectionnés aléatoirement, seront calculés.

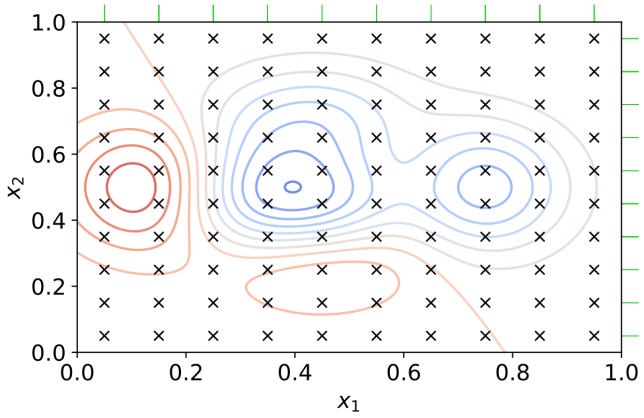

Figure 30 – Illustration extraite du web représentant la méthode de recherche sur grille avec deux hyperparamètres x_1 et x_2. Sur chaque point de la grille, un modèle est calculé en utilisant le couple (x_1, x_2) du point sélectionné. Les grandes valeurs de la métrique de sélection sont indiquées par les lignes de niveaux rouge et les petites valeurs par les lignes bleues. Le modèle optimal est obtenu en choisissant le point, i.e le couple (x_1, x_2), qui optimise la métrique de sélection.

4.4.2 Régression linéaire nette élastique

À présent que nous avons introduit la méthode d'optimisation des hyperparamètres, nous allons présenter plus en détails les différents types de modèles.

Dans cette approche, la variable réponse, notée $y_{occ}(c,t)$, est binaire et vaut 1 si la commune c est décrétée en situation de catastrophe naturelle à l'année t par la commission interministérielle et 0 sinon. Nous considérons donc un modèle GLM binomial avec lien *logit* et deux paramètres de régularisation. La méthode est plus connue sous le nom de régression linéaire nette élastique qui combine la régression lasso et la régression ridge. L'écriture du modèle reste celui d'un modèle GLM avec une fonction de lien et la différence provient uniquement du problème d'optimisation qui possède des termes de régularisation supplémentaires.

En notant y_i les réalisations de la variable réponse, x_i les vecteurs de longueur p des réalisations des variables explicatives, N le nombre d'observations, β le vecteur des coefficients de régression, β_0 la constante de régression, λ et α deux hyperparamètres de régularisation, alors le problème d'optimisation, i.e la maximisation de la vraisemblance ou, de manière équivalente, la minimisation de la moins log vraisemblance, est :

$$\min_{(\beta,\beta_0)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i.(\beta_0 + x_i^T\beta) - \ln(1+\mathrm{e}^{\beta_0+x_i^T\beta})\right] + \lambda\left[\alpha\|\beta\|_1 + \frac{(1-\alpha)}{2}\|\beta\|_2^2\right] \tag{56}$$

Avec $\|\beta\|_1 = \sum_{i=1}^{p}|\beta_i|$ et $\|\beta\|_2 = \sum_{i=1}^{p}\beta_i^2$ les normes 1 et 2. Dans le cas d'un modèle GLM non régularisé, il s'agit du même problème d'optimisation mais avec $\lambda = 0$. La méthode de recherche sur grille s'applique donc sur les deux hyperparamètres λ et α. Le modèle optimal trouvé correspond à $\alpha = 0$, ce qui revient en fait à une régression ridge. Pour l'hyperparamètre λ, dans la mesure où la librairie *h2o* de **R** est utilisée, il est possible de spécifier une recherche automatique pour trouver la valeur optimale de λ, et c'est ce qui a été fait. Le λ optimal correspond au plus petit λ tel que les erreurs des prédictions sont minimales.

Notez que l'intérêt d'ajouter un terme de régularisation pour pénaliser la log vraisemblance est de limiter le problème de sur-apprentissage. En effet, l'idée est que la solution optimale qui limite le sur-apprentissage se trouve non pas au point exact où la moins log vraisemblance atteint son minimum, mais plutôt dans le voisinage de ce dernier, ce qui semble logique. Ainsi, avec cette régularisation, les coefficients de régression, en se rapprochant de la solution optimale, va certes faire diminuer la moins

log vraisemblance mais va aussi faire augmenter le terme de régularisation de telle sorte que la solution optimale devient un point du voisinage du point minimum.

4.4.3 Gradient boosting

Les deux prochains modèles décrits sont le GBM et le XGB avec le second qui est, d'un point de vue modélisation, un GBM régularisé comme pour le GLM et le GLM régularisé. Notez que n'allons pas rentrer dans les détails de ces modèles car ces derniers, en plus d'être des modèles, correspondent en fait plutôt à des algorithmes, voire des librairies, avec toutes les problématiques d'optimisation des ressources informatiques derrières comme le contrôle de la complexité ou les calculs parallèles. Dans cette partie, nous nous contenterons de présenter les problèmes mathématiques que ces modèles tentent de résoudre.

Le GBM est un modèle d'apprentissage supervisé utilisé pour des problèmes de classification ou de régression. L'idée est d'utiliser un ensemble de petits modèles pour construire le modèle GBM qui minimise une certaine fonction objective. En pratique, les petits modèles en question sont souvent des arbres de décision.

Soit y_{occ} la variable réponse, X_i les variables explicatives avec x_i les réalisations et N le nombre de réalisations. Le problème général que tout modèle souhaite résoudre est :

$$y_{occ} = F(X) \tag{57}$$

Avec F la fonction que nous souhaitons approximer. Dans le cas du GBM, la fonction F est approximée de manière itérative, i.e une suite $(F_m)_{m\leq M}$ est construite tel que F_m converge vers F quand m tend vers M. Bien entendu, M doit valoir $+\infty$ idéalement mais ce n'est pas matériellement possible. En pratique, il est nécessaire de spécifier des conditions d'arrêt. Pour raison de lisibilité d'écriture et sans perdre de généralité, nous allons supposer qu'il n'y a qu'une seule variable explicative X de réalisations x_i. Ainsi, dans le cadre du modèle GBM, la suite $(F_m)_{m\leq M}$ est définie par :

$$\begin{cases} F_0(X) = \underset{\gamma}{\operatorname{argmin}}\, \mathcal{L}(\gamma) = \underset{\gamma}{\operatorname{argmin}} \sum\limits_{c,t} L(y_{occ}(c,t), \gamma) \\ F_m(X) = F_{m-1}(X) + h_m(X) \end{cases} \tag{58}$$

Les $y_{occ}(c,t)$ sont les réalisations de la variable réponse avec c qui indique la commune et t l'année, L est la métrique ou la fonction loss à optimiser, et les h_m sont des petits modèles qui sont souvent des arbres de décision. Il faut préciser que le problème consiste à optimiser non pas individuellement chaque $L(y_{occ}(c,t), \gamma)$ pour chacune des observations mais plutôt la somme des loss $\mathcal{L}$. En effet, cela permet notamment de garantir une cohérence globale et de limiter par exemple des perturbations liées à des observations particulières. En pratique, minimiser une somme et minimiser chaque terme de la somme donne des résultats souvent similaires.

Par ailleurs, les modèles qui composent le modèle GBM, i.e h_m, vérifient :

$$h_m(X) = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_{occ,i}, F_{m-1}(x_i) + h(x_i)) \tag{59}$$

Où $\mathcal{H}$ est l'ensemble des fonctions possibles pour h_m. Néanmoins, trouver h_m est un problème complexe car $\mathcal{H}$ est infini. Une simplification possible est de s'inspirer de la méthode de descente de gradient. Ainsi, l'expression simplifiée, donc approximative, est :

$$\begin{cases} h_m(X) \approx -\gamma_m \sum\limits_{i=1}^{N} \nabla_{F_{m-1}} L(y_{occ,i}, F_{m-1}(x_i)) \\ \gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum\limits_{i=1}^{N} L(y_{occ,i}, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_{occ,i}, F_{m-1}(x_i))) \end{cases} \tag{60}$$

Avec ∇ l'opérateur gradient mais dans un espace fonctionnel ([33]). L'idée derrière l'expression de h_m est qu'à chaque itération, F_m doit être une fonction qui est plus approche de la fonction où le minimum

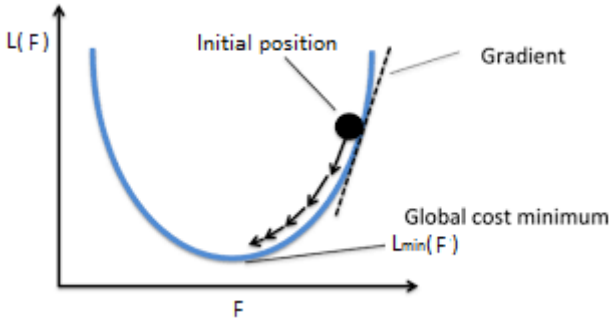

Figure 31 – Schéma illustrant la méthode de descente de gradient. Pour atteindre le minimum de $\mathcal{L}$, la fonction F doit évoluer dans la direction indiquée par moins le gradient de $\mathcal{L}$ par rapport à F.

de $\mathcal{L}$ est atteint comparée à la fonction F_{m-1}. En d'autres termes, h_m doit être une fonction qui fait bouger F_{m-1} dans la direction de diminution de $\mathcal{L}$.

Pour mieux comprendre ces expressions, il faut considérer le schéma de la figure 31. Pour une fonction initiale F quelconque, afin d'atteindre le minimum de $\mathcal{L}$, F doit évoluer dans la direction indiquée par moins le gradient (dans un espace de fonctionnelle) de $\mathcal{L}$ par rapport à F, i.e $-\partial\mathcal{L}/\partial F$. Pour obtenir ensuite l'expression de h_m complète, il faut ajouter un terme γ_m en plus afin de contrôler la taille du pas fait dans cette direction de diminution de $\mathcal{L}$. Et ce γ_m est choisie de telle sorte que la diminution de $\mathcal{L}$ est maximale.

Le modèle XGB, mis à part toutes les optimisations des ressources informatiques en plus, se distingue du GBM notamment dans la fonction objective. En effet, comme pour la différence entre une régression linéaire classique et une régression linéaire nette élastique, le XGB introduit un terme de régularisation dans l'expression de la métrique à optimiser. Ainsi, en notant $\widehat{y}^m_{occ}$ l'estimation de y_{occ} à l'étape m calculée à partir des F_j où $j \leq m$, et $\mathcal{L}_m$ la fonction objective à l'étape m, les expressions des fonctions objectives pour le GBM et le XGB sont :

$$\begin{cases} \mathcal{L}_m^{GBM} = \sum\limits_{c,t} L(y_{occ}(c,t), \widehat{y}^m_{occ}(c,t)) \\ \mathcal{L}_m^{XGB} = \Omega(h_m) + \sum\limits_{c,t} L(y_{occ}(c,t), \widehat{y}^m_{occ}(c,t)) \end{cases} \tag{61}$$

Notez donc que pour le XGB, à chaque choix de h_m, il faut que ce dernier respecte une certaine régularité représentée par Ω tandis que dans le cas du GBM, il n'y a pas cette contrainte. Dans la littérature, il semble que le modèle XGB tend souvent à avoir de meilleures performances que le GBM. Les lecteurs intéressés par ces deux modèles peuvent se référer aux [31], [32] et [33].

En conclusion, les modèles GBM et XGB optimaux sont obtenus via la recherche sur grille à l'aide de la librairie *h2o* de **R** et possèdent de nombreux hyperparamètres dont les plus importants portent sur h_m. Dans le cas où les h_m sont des arbres de décisions, les hyperparamètres les plus importants sont la profondeur maximale des arbres de décision, le ratio du nombre de variables sélectionnées sur le nombre de variables total, et le ratio du nombre d'observations utilisées sur le nombre d'observations total. Les valeurs optimales de ces hyperparamètres utilisées par la suite sont dans la figure 32. Les autres hyperparamètres ne sont pas détaillés car les impacts de ces derniers sont très limités.

4.4.4 Forêt aléatoire

Le modèle de forêt aléatoire est déjà introduit lors de la description de la méthode ensembliste de sélection des variables. Nous n'allons donc pas insister davantage et dans cette section, nous nous contenterons de faire quelques rappels mais surtout mettre en évidence la différence entre forêt aléatoire (RF) et forêt aléatoire extrême (XRF).

Nous nous souvenons que le modèle de forêt aléatoire est construit à partir des arbres de décision, et que chaque arbre est construit via l'optimisation d'un critère en utilisant un sous-ensemble de variables

	GBM	XGB
Profondeur maximale	20	14
Ratio observation	0.94	0.77
Ratio variable	0.22	0.25

Figure 32 – Tableau résumant les trois hyperparamètres les plus importants des modèles GBM et XGB ainsi que les choix pour ces derniers obtenus via la recherche sur grille. Ces hyparamètres portent sur les arbres de décision h_m.

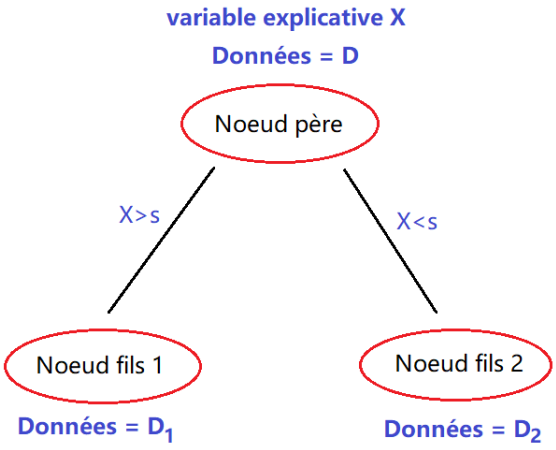

Figure 33 – Schéma illustrant la division d'un nœud. La différence entre une forêt aléatoire et une forêt aléatoire extrême réside dans le choix du seuil s.

et un sous-ensemble d'observations. Dans la construction d'un arbre de décision, le plus important est de savoir comment sont divisés les nœuds. Dans le cas d'une variable réponse catégorielle, cette division est souvent basée sur l'impureté de Gini.

Le modèle XRF trouve son intérêt notamment lorsque les variables explicatives sont numériques et non catégorielles, ce qui est le cas ici. La différence entre RF et XRF réside notamment dans la construction des arbres de décision, et plus précisément dans le choix du seuil lors de la division des nœuds. Pour une meilleure compréhension du problème, il faut considérer le schéma de la figure 33.

Soient X la variable explicative sélectionnée pour la division du nœud père, X_{min} et X_{max} respectivement le minimum et le maximum de X, D les données présentes dans le nœud père, et D_1 et D_2 les données qui vont transiter du nœud père respectivement aux nœuds fils 1 et 2 avec $D = D_1 \cup D_2$. Soit s_{RF} le seuil dans le cas du modèle RF. Alors ce seuil est défini par :

$$s_{RF} = \underset{s\in\{ih,\ i=1,2,\dots,N-1\}}{\operatorname{argmin}} \frac{\operatorname{card}(D_1(s))}{\operatorname{card}(D)} G(D_1, s) + \frac{\operatorname{card}(D_2(s))}{\operatorname{card}(D)} G(D_2, s) \tag{62}$$

Avec card la fonction cardinal pour un ensemble et :

$$\begin{cases} h = \dfrac{X_{max} - X_{min}}{N} \\ G(D_k, s) = 1 - \displaystyle\sum_{i=1}^{C} p_i(k, s) \end{cases} \tag{63}$$

Ainsi, h est le pas avec N classiquement de l'ordre de 10^3, C le nombre de classes dans la variable réponse, i.e $C = 2$ pour le cas présent, et $p_i(k, s)$ la fraction de données de classe i dans l'ensemble des données D_k. Il faut préciser que les D_1 et D_2 dépendent du seuil car en fonction de ce dernier, les données qui sont attribuées à D_1 et D_2 depuis le nœud père sont différentes.

Notez que l'importance ici est que le seuil s_{RF} est choisi parmi un ensemble de seuils possibles uniformément distribués. Ce qui n'est pas le cas pour le modèle XRF. En effet, dans le XRF, les seuils possibles ne sont plus uniformément distribués mais sont purement aléatoires. Ainsi, en notant s_{XRF} ce seuil, ce dernier est défini par :

$$s_{XRF} = \underset{s=s_1,s_2,...,s_{N-1}}{\operatorname{argmin}} \frac{\operatorname{card}(D_1(s))}{\operatorname{card}(D)} G(D_1, s) + \frac{\operatorname{card}(D_2(s))}{\operatorname{card}(D)} G(D_2, s) \tag{64}$$

Où les s_i sont des valeurs de seuils aléatoires prises dans l'intervalle $[X_{min}, X_{max}]$. Selon les auteurs ([34]), l'intérêt du XRF réside surtout dans le fait que ce dernier est plus efficace que le RF sur le plan d'optimisation des ressources informatiques. En pratique, il semble que le XRF permet d'avoir un modèle de moindre variance dans les prédictions au prix d'une légère augmentation du biais.

De même que pour les modèles GBM et XGB, les modèles RF et XRF optimaux sont obtenus via la recherche sur grille à l'aide de la librairie *h2o* de **R**. Les hyperparamètres optimaux sont nombreux mais les plus importants, qui portent aussi sur les arbres de décisions, sont les mêmes que pour le GBM et le XGB. Les valeurs des hyperparametres sont présentées dans la figure 34.

	RF	**XRF**
Profondeur maximale	20	28
Ratio observation	0.86	0.59
Ratio variable	0.67	0.27

Figure 34 – Tableau résumant les trois hyperparamètres les plus importants des modèles RF et XRF ainsi que les choix pour ces derniers obtenus via la recherche sur grille. De même que pour les modèles GBM et XGB, ces hyparamètres portent sur les arbres de décision.

4.4.5 Réseau de neurones artificiels

L'avant dernier type de modèle testé est le réseau de neurones artificiel. La littérature autour des réseaux de neurones est très riche et en constante progression passant du simple perceptron aux réseaux complexes comme les réseaux de convolution (CNN ou Convolutionnal Neural Network). L'architecture des réseaux de neurones est un domaine en pleine innovation et constitue le cœur de nombreux métiers technologiques de nos jours. Face à toute cette complexité, nous nous contenterons de donner des définitions simples et adaptées à ce qui est testé au cours du projet.

De manière naïve, un réseau de neurones est un modèle d'apprentissage supervisé ou plus simplement une fonction qui à une entrée fait correspondre une sortie. Ce modèle peut être représenté à l'aide d'un réseau de plusieurs couches de neurones à l'image d'un cerveau humain. La figure 35 représente un réseau de neurones avec une couche d'entrée de 3 neurones, deux couches cachées respectivement de 4 et 3 neurones, et une couche de sortie de 4 neurones. Les fonctions f_i sont des opérations entre couches de neurones et les fonctions a_i sont des fonctions dites d'activation. Lorsque chaque neurone d'une couche est connectée à tous les neurones de la couche suivante, et ceci pour toutes les couches, alors le réseau de neurones est dit complètement connecté.

En formulation mathématique, si x est une observation, y la réponse et F la fonction tel que $y = F(x)$, alors un réseau de neurones tente d'approximer F par $\widehat{F}$ défini par :

$$\widehat{F} = f_{N-1} \circ a_{N-2} \circ ... \circ f_2 \circ a_2 \circ f_1 \tag{65}$$

Avec $\circ$ l'opérateur composition pour les fonctions. De manière générale, les fonctions a_i sont souvent des fonctions déterministes prédéfinies dont la plus connue est le redresseur (ReLU ou rectified Linear Unit) avec $a_i(x) = \max(x, 0)$ et les f_i sont des fonctions d'une famille fixée au départ. Toute la procédure d'apprentissage du réseau de neurones consiste alors à utiliser les données d'entraînement afin de trouver les paramètres des f_i via l'optimisation d'une fonction objective ou fonction loss. Par

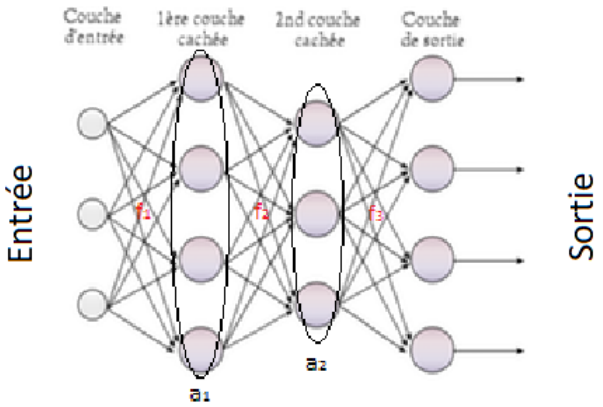

Figure 35 – Schéma représentant un réseau de neurones complètement connecté avec, de gauche à droite, une couche d'entrée de 3 neurones, deux couches cachées respectivement de 4 et 3 neurones, et une couche de sortie de 4 neurones. Les fonctions f_i sont des opérations entre couches de neurones et les fonctions a_i sont des fonctions dites d'activation.

exemple si f_i appartient à la famille des fonctions linéaires, i.e f_i est une matrice de taille définie par le nombre de neurones dans les deux couches que f_i relie, alors l'objectif de l'apprentissage sera de trouver les différents coefficients de la matrice qui minimise la fonction objective. Durant cette phase, il faut noter qu'une même donnée peut être utilisée plusieurs fois, et ce nombre de fois d'utilisation est appelé époque. De plus, à chaque pas d'apprentissage, seul un sous-ensemble de données d'entraînement est utilisé. La taille de cet sous-ensemble est souvent fixée au départ et appelée batch size.

Par ailleurs, il est fréquent d'introduire une couche de décrochage (ou Dropout) entre les a_i et les f_i et/ou avant la couche d'entrée de manière à simuler des neurones morts et éviter le sur-apprentissage. Dans le même registre, il est aussi fréquent d'utiliser un pas d'apprentissage variable qui joue le même rôle que le γ_m dans le modèle GBM ou d'utiliser des méthodes diverses pour trouver les paramètres des f_i. Bien entendu, la méthode la plus célèbre est la méthode dite de rétropropagation du gradient ou backpropagation, qui est basée sur la descente de gradient stochastique. Le mot stochastique provient du fait que seul un seul un petit nombre d'observations est utilisé à chaque pas d'entrainement. Notez donc que si le batch size vaut exactement le nombres de données, alors cela revient à l'algorithme de descente de gradient classique.

Dans le cas présent avec l'algorithme de recherche sur grille, les réseaux de neurones sont complètement connectés avec plus ou moins de couches cachées et plus ou moins de neurones par couche. Nous autorisons aussi l'utilisation des activations autre que ReLU comme la tangente hyperbolique. Ainsi, le réseau optimal obtenu via l'aide de la librairie *h2o* de **R** est un réseau avec 3 couches cachées de 500 neurones chacune, une activation tangente hyperbolique, un ratio de dropout de 0.3 à l'entrée du réseau (i.e 30% des signaux d'entrés sont mis à zéro), une taille des sous-ensembles ou un batch size de 16 dans la descente du gradient stochastique et avec l'entropie croisée comme fonction de loss.

Dans le cas où le modèle d'occurrence serait un réseau de neurone, avec l'objectif de prédire si une commune sera décrétée en situation de catastrophe naturelle ou pas par la commission interministérielle, si $\mathcal{L}$ est la moins log-vraisemblance, alors $\mathcal{L}$ est défini par :

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N_b} p_i \ln(q_i(\theta)) + (1-p_i)\ln(1-q_i(\theta)) \tag{66}$$

Avec N_b le batch size ou le taille du sous-ensemble dans la méthode de rétropropagation du gradient, q_i la probabilité prédite que la commune i sera décrétée en situation de catastrophe naturelle, p_i la probabilité non plus prédite mais observée, et θ l'ensemble des paramètres du modèle.

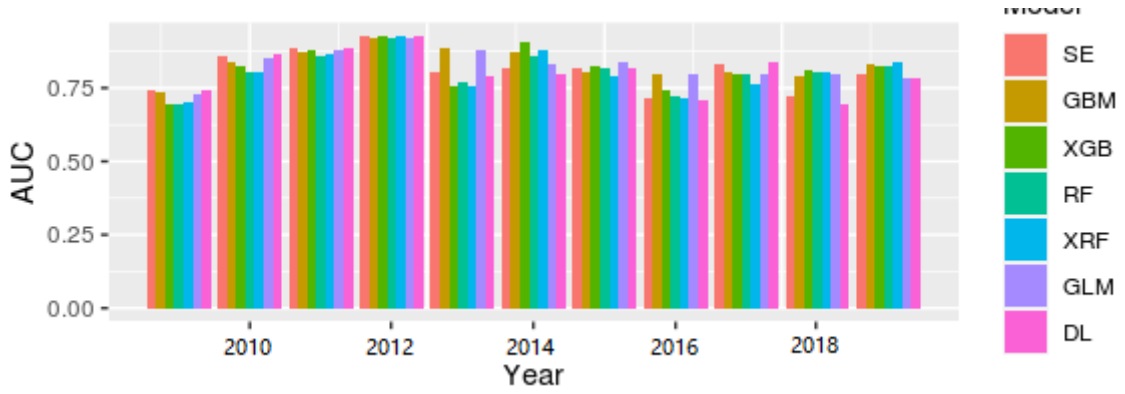

Figure 36 – Plot représentant l'AUC des 7 modèles décrits précédemment pour les années 2009, 2010, ..., 2019. Notez qu'il s'agit bien de l'AUC sur l'ensemble test. Notez aussi que pour 2013 et 2016, le GBM et le GLM se distinguent des autres et que pour les autres années, leurs AUC sont relativement similaires aux AUC des autres modèles.

4.4.6 Modèle ensembliste

Le dernier modèle considéré s'inspire de la méthode ensembliste de sélection de variable. En effet, naïvement, chaque modèle a à priori ses propres biais mais s'il y a une diversification importante du biais parmi les modèles, alors en somme, les biais peuvent s'annuler entre eux. L'idée est donc que pour une observation donnée, la réponse prédite est plus ou moins correcte en fonction du modèle utilisé mais la moyenne des prédictions, sur les différents modèles distincts, peut être à priori mieux que les résultats individuels.

Un modèle ensembliste est donc construit en utilisant les prédictions des 6 précédents modèles (GBM, XGB, RF, XRF, DL et GLM) comme variable explicative. Le modèle ensembliste construit est un GLM classique non régularisé et avec comme fonction de lien la fonction identité.

4.4.7 Résultat et analyse

Pour évaluer la performance des différents modèles, le critère d'AUC est utilisé. Plus exactement, pour chaque modèle, nous suivons les étapes suivantes :

- Étape 1 : Choisir une année et mettre de côté les données de cette année choisie.
- Étape 2 : Calibrer le modèle sur les données des années non choisies
- Étape 3 : Calculer l'AUC du modèle avec les données de l'année mise de côté
- Étape 4 : Réitérer les étapes 1 à 3 tant qu'il reste encore des années non choisies.

Notez que ces étapes permettent notamment d'avoir un AUC sur un ensemble test et qu'en itérant sur toutes les années, nous avons une idée de la performance globale du modèle. Et en faisant cela pour tous les modèles décrits précédemment, nous obtenons la figure 36.

Il faut remarquer dans cette figure que pour 2013 et 2016, le GBM et le GLM se distinguent des autres et que pour les autres années, leurs AUC sont relativement similaires aux AUC des autres modèles. Pour confirmer cela, en notant $\mathcal{M} = \{GLM, GBM, ..., DL\}$ l'ensemble des modèles, le score Z_{model} est défini par :

$$\forall model \in \mathcal{M},\ Z_{model} = \sum_{k=2009}^{2019} AUC(model, k) - \max_{i \in \mathcal{M}} AUC(i, k) \tag{67}$$

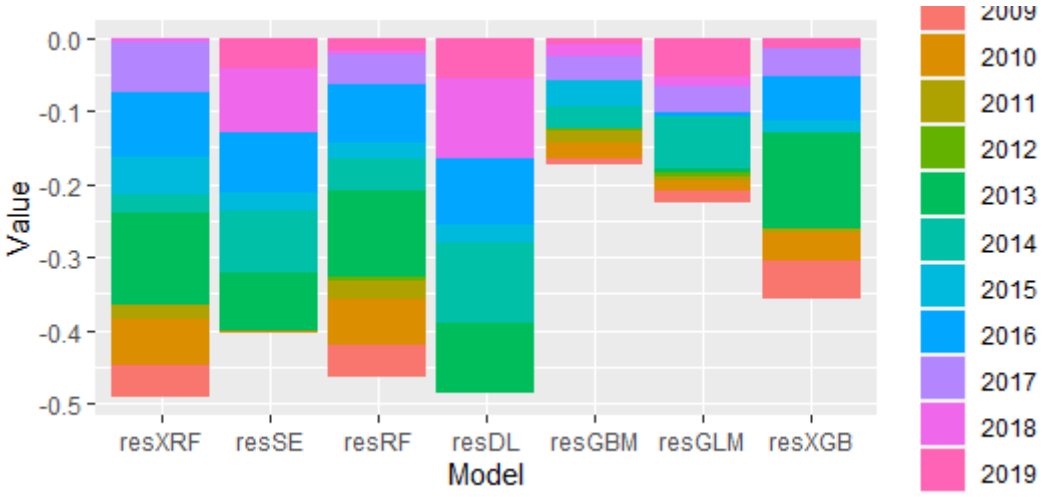

Figure 37 – Plot représentant les scores Z_{model} pour les 7 modèles décrits précédemment avec les années 2009, 2010, ..., 2019 en couleur. Notez qu'effectivement, avec les 30 variables explicatives sélectionnées en 4.3, les modèles GBM et GLM sont les deux modèles les plus performants avec le GBM légèrement meilleur que le GLM.

Dans la figure 37, notez qu'effectivement, avec les 30 variables explicatives sélectionnées en 4.3, les modèles GBM et GLM sont les deux modèles les plus performants avec le GBM légèrement meilleur que le GLM.

Ainsi, compte tenu de ces résultats, le modèle GLM est choisi comme modèle d'occurrence. En effet, même si le GBM semble être meilleur que le GLM, la différence reste très légère. De plus, le GBM est beaucoup moins interprétable que le GLM et dans la perspective de construire un produit d'assurance-réassurance, le GBM est plus difficile à contractualiser par sa complexité. Pour ces raisons, le choix est porté sur le GLM, ou plus exactement sur le modèle GLM ridge binomial *logit*, dont les coefficients sont facilement interprétables. Ainsi, pour la variable réponse binaire y_{occ}, le modèle d'occurrence modélise :

$$\text{logit}(y_{occ}) = \frac{y_{occ}}{1 - y_{occ}} = \beta X \tag{68}$$

Avec X le vecteur des variables explicatives et β les coefficients de régression ridge. Le modèle d'occurrence n'est pas davantage détaillé car une définition plus complète du modèle GLM est donnée dans la section 3.3.2.1 et la différence entre régression classique et régression ridge est aussi donnée dans la section 4.4.2.

4.5 Modèle de fréquence et méthode de sélection des variables

Dans l'approche occurrence-fréquence, il faut se focaliser à présent sur le modèle de fréquence et la construction de ce dernier. Nous rappelons que nous disposons de plus de 800 variables (821 plus exactement) qui décrivent les conditions climatiques, géologiques et socio-économiques des communes françaises et nous souhaitons que le modèle de fréquence, pour chaque commune, modélise la proportion de contrats sinistrés. Le passage de la proportion de sinistres à la charge de sinistres sera géré ensuite par un coefficient historique.

En d'autres termes, pour une commune c et une année t, si $n_{risk}(c,t)$ est le nombre de contrats localisés et $n_s(c,t)$ le nombre de sinistres observés, alors la variable réponse $y_{freq}(c,t)$ du modèle de fréquence est définie par :

$$y_{freq}(c,t) = \frac{n_s(c,t)}{n_{risk}(c,t)} \tag{69}$$

Notez que $y_{freq}(c,t)$ est bien une quantité observée car les contrats sont enregistrés dès leurs établissements. Pour les sinistres, il est nécessaire d'attendre la décision de la commission interministérielle et la publication de cette dernière dans le Journal Officiel. Néanmoins, dans la mesure où le délai d'attente est de l'ordre d'un an et demi pour les sinistrés de sécheresse et que les années présentes dans le portefeuille sont 2016, 2017 et 2018, il est raisonnable de penser qu'à la date d'aujourd'hui, les sinistres sont aussi bien observés et ne risque pas d'évoluer à trop grande échelle.

Avant d'aborder la question du quel modèle choisir, semblable à la construction du modèle d'occurrence, pour les mêmes raisons, il faut essayer de réduire le nombre de variables explicatives. Pour cela, il faut procéder en plusieurs étapes, à savoir dans l'ordre des opérations :

- Étape 1 : Supprimer les variables dont la variance est quasi nulle.
- Étape 2 : Pour les variables restantes, en cas de deux variables colinéaire, négliger la variable la plus corrélée aux autres variables.
- Étape 3 : Utiliser l'algorithme boruta ([36]) sur les variables indépendantes et récupérer celles estimées pertinentes.
- Étape 4 : Appliquer l'algorithme stepwise ([37])pour obtenir un modèle de régression réduit.
- Étape 5 : (Facultatif) Calculer l'importance des variables (4.3.1.3) et ne garder que les 30 variables les plus importantes.

Par la suite, dans l'ordre, chacune des étapes sera détaillée et illustrée avec des exemples si besoin. Le modèle de fréquence sera présenté après avoir décrit ces opérations de sélection des variables.

4.5.1 Variance quasi nulle

Soit X une variable quelconque. Alors X a une variance quasi nulle si les réalisations de X vérifient :

- C_1 : Le ratio r_1 du nombre d'occurrences de la valeur la plus fréquente sur le nombre d'occurrences de la seconde valeur la plus fréquente est plus grand qu'un seuil f_{cut}.
- C_2 : Le ratio r_2 du nombre de valeurs distinctes sur le nombre d'observations est plus petit qu'un seuil p_{cut}

Pour illustrer cela ces deux conditions, soient les 4 variables de 8 observations chacune suivantes :

Obs	X_1	X_2	X_3	X_4
$i=1$	1	1	1	1
$i=2$	1	3	1	2
$i=3$	1	3	2	8
$i=4$	1	2	1	4
$i=5$	1	3	2	5
$i=6$	1	3	1	3
$i=7$	1	2	2	6
$i=8$	1	3	2	7

Pour $f_{cut}=2$ et $p_{cut}=30\%$, la variable X_1 est à négliger car ne vérifie ni C_1, ni C_2. La variable X_2 et X_3 sont aussi à négliger car le premier vérifie C_1 mais pas C_2 et le second vérifie C_2 mais pas C_1. La dernière variable X_3 est à garder car elle vérifie les deux conditions.

Bien entendu, il faut préciser que cette notion de variance quasi nulle prend du sens surtout lorsque la variable est discrète, ce qui est le cas pour certaines des 821 variables de la base de données de

catastrophe naturelle. En effet, pour les variables continues, tous les valeurs sont en théorie uniques et par conséquent, les conditions C_1 et C_2 sont toujours vérifiées car $r_1 = r_2 = 1$.

Notez que cette opération est réalisée à l'aide de la fonction $nearZerVar()$ de la librairie *caret* de **R**. Par la suite f_{cut} est fixé à 19 et p_{cut} à 10%.

4.5.2 Variables colinéaires

Dans cette étape, soient $(X_1, X_2, ..., X_n)$ les variables explicatives et $\rho_{i,j}$ les coefficients de corrélation deux à deux. Alors les $\rho_{i,j}$ sont définis par :

$$\rho_{i,j} = \frac{\mathbf{E}((X_i - \mathbf{E}X_i)(X_j - \mathbf{E}X_j))}{\sigma_{X_i}\sigma_{X_j}} \tag{70}$$

Avec $\mathbf{E}$ la fonction espérance et les σ les écart-types. Bien entendu, d'un point de vue calculatoire, ces coefficients de corrélation sont calculés à l'aide des approximations classiques.

Par ailleurs, pour chaque variable X_i, soit ρ_i sa corrélations moyenne avec les autres variables. Alors ρ_i est défini par :

$$\rho_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} |\rho_{i,j}| \tag{71}$$

Ainsi, l'opération de l'étape 2 consiste à considérer successivement chaque $\rho_{i,j}$ tel qu'à chaque fois que $\rho_{i,j}$ est plus grand qu'un seuil s fixé au départ, l'une des deux variables X_i ou X_j sera considérer comme redondante et donc à négliger par la suite. Le choix de la variable redondante est basé sur la corrélation moyenne ρ_i, i.e la variable la plus corrélée aux autres sera négligée. En d'autres termes, si $\rho_{i,j} > s$ et si $\rho_i > \rho_j$, alors la variable X_i sera négligée.

Notez que puisque $\rho_{i,j} = \rho_{j,i}$, il suffit d'examiner $n(n-1)/2$ termes avec n le nombre de variables. Notez aussi que l'opération est réalisée par la fonction $findCorrelation()$ de la librairie *caret* de **R**. Par la suite, le seuil s est fixé à 0.90.

4.5.3 Algorithme Boruta

L'algorithme boruta est un algorithme de sélection des variables basé sur des forêts aléatoires. Avant de présenter les détails de cet algorithme, nous rappelons qu'une forêt aléatoire est un ensemble d'arbres de décision chacun construit à partir d'un sous-ensemble d'observations et d'un sous-ensembles de variables tous sélectionnés aléatoirement. Le fait de prendre uniquement des sous-ensembles permet de définir pour chaque arbre la notion de données OOB qui sont les données non utilisées dans la construction de l'arbre en question.

Notez aussi qu'il est possible de définir des mesures d'importance pour les variables à l'aide d'une forêt aléatoire. L'une des mesures d'importance est la décroissance moyenne de l'indice de Gini induite par la variable, définie précédemment lors de la présentation de l'algorithme de sélection ensembliste. Cette dernière mesure à quel point une variable contribue dans la décroissance de l'indice d'impureté. Une autre mesure d'importance possible est la décroissance moyenne du taux de vraie positive (dans le cas d'une classification) ou la décroissance moyenne de l'erreur de prédiction (dans le cas d'une régression).

En effet, pour chaque arbre de décision, l'erreur de prédiction est calculée à l'aide de ses données OOB associées, puis en moyennant sur tous les arbres, une erreur moyenne est obtenue. Ensuite, ces opérations sont reconduites avec cette fois les réalisations de la variable, dont nous souhaitons avoir l'importance, permutées. Cette permutation revient à supprimer les pouvoirs prédictifs de la variable tout en gardant les lois marginales intactes. L'importance de la variable est alors la différence de l'erreur moyenne entre le cas permuté et le cas non permuté. Cette mesure d'importance est communément notée MDA (Mean Decrease Accuracy). Nous ne rentrerons pas davantage dans les détails de la définition du MDA et de ses déclinaisons en fonction du problème considéré. Nous retenons simplement pour le besoin que, le MDA est à valeur réelle positive et plus le MDA est grand, plus la variable est

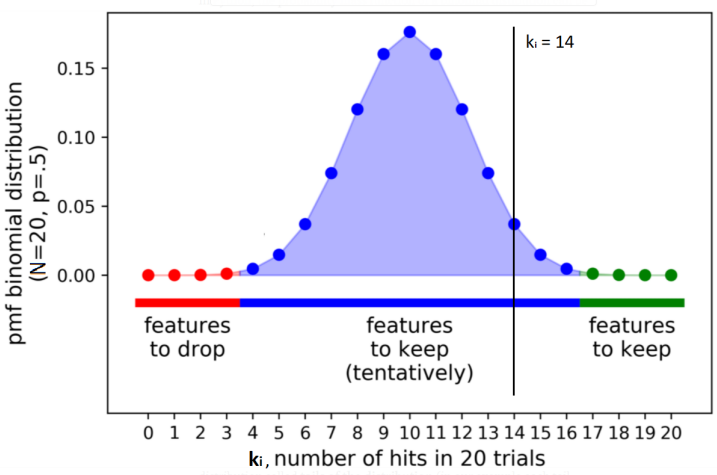

Figure 38 – La fonction de masse d'une loi binomiale $\mathcal{B}(N = 20, p = 0.5)$. Nous classons les k_i à l'aide de la fonction de masse en les mettant en abscisse. Les zones verte et rouge représentent chacune une masse de 0.5%. Les variables dont les k_i sont classés en rouge sont à supprimer et celles en vert sont à garder. Notez que dans cette étude, les variables dite tentatives seront aussi gardées.

importante.

À présent, soient $(X_1, X_2, ..., X_n)$ les variables explicatives, alors l'algorithme boruta consiste à :

- Étape 1 : Considérer un vecteur de taille n, $(k_1, k_2, ..., k_n)$ avec tous les $k_i = 0$ initialement.
- Étape 2 : Construire $(X_1^s, X_2^s, ..., X_n^s)$ avec X_i^s la variable X_i dont les réalisations ont été permutées aléatoirement.
- Étape 3 : Construire une forêt aléatoire à partir des données $(X_1, X_2, ..., X_n, X_1^s, X_2^s, ..., X_n^s)$ et calculer le MDA des $2n$ variables.
- Étape 4 : Calculer $m = \max_{i=1,2,...,n} \text{MDA}(X_i^s)$.
- Étape 5 : Pour chaque i allant de 1 à n, si $\text{MDA}(X_i) > m$ alors incrémenter k_i de 1.
- Étape 6 : Réitérer N fois l'étape 3 et 4.
- Étape 7 : Classer les variables en fonction de leurs k_i et en utilisant la fonction de masse d'une loi binomiale $\mathcal{B}(N, p = 0.5)$ (figure 38).

Durant l'étape 7 de l'algorithme, les variables explicatives X_i associées aux k_i sont classées à l'aide d'une loi binomiale $\mathcal{B}(N, p = 0.5)$ avec N le nombre d'itérations à l'étape 6 comme l'illustre la figure 38. À la sortie de l'algorithme boruta, les variables sont classées en trois catégories, à savoir : tentatives, à supprimer ou à garder. Il est possible de considérer un N plus grand ou des zones rouge et verte plus grande (figure 38) pour ne plus avoir de variables indécises. Néanmoins, augmenter N demande plus de temps de calcul et agrandir les zones verte et rouge, i.e augmenter la p-value, nuit au degré de confiance du classement. Dans le cas présent, il est préférable de garder les variables avec un statut indécis s'il y en a. Par la suite, par prudence donc, les variables tentatives sont traitées de la même manière que les variables à garder.

Il est intéressant de noter que la puissance de cet algorithme repose sur la construction aléatoire des forêts aléatoires ainsi que sur une décision non paramétrique mais plutôt statistique. Une variable est potentiellement à garder si elle a une importance plus grande que la meilleure des variables permutées.

Puis, avec une loi binomiale où $p = 0.5$, correspondant à un cas d'absence total d'information, les $k_i - 0.5N$, dans cette perspective, peuvent être vu comme des déviations par rapport à la situation d'absence d'information. Si $k_i - 0.5N \ll 0$, alors la variable X_i n'a pas de pouvoir explicatif puisque la variable X_i n'est guère mieux que la variable X_i permuter aléatoirement. Au contraire si $k_i - 0.5N \gg 0$, alors X_i a un réel pouvoir explicatif puisque son MDA est significativement plus grand que les MDA des variables permutées.

Par la suite, la fonction *Boruta*() de la librairie *Boruta* de **R** est utilisée pour cette opération. De plus, N est fixé à 1000 et les zones verte et rouge restent avec une masse de 0.5% chacune.

4.5.4 Sélection stepwise

Après l'algorithme boruta, en fonction des approches, il reste souvent encore une centaine de variables explicatives. Cela ne correspond qu'à environ 1/8 du nombre de variables explicatives initiales mais le modèle reste trop complexe. Il est nécessaire d'aller plus loin dans la sélection des variables.

Dans ce but, l'algorithme bien connu de régression stepwise ([37]) est utilisé. L'idée de la méthode est de réduire le nombre de variables explicatives tout en essayant de garder intact la performance du modèle. Plus exactement, un modèle initial à l'entrée de l'algorithme est spécifié. L'algorithme doit ensuite enlever des variables du modèle initial ou rajouter des variables au modèle initial tout en minimisant un critère. En pratique, le critère en question est souvent le critère d'information d'Akaike (AIC) défini, avec M un modèle quelconque, par :

$$\mathrm{AIC}(M) = 2k - 2\ln(L) \tag{72}$$

Avec k le nombre de paramètres dans le modèle M et L la valeur de la vraisemblance. Dans le cas présent où les la variable réponse correspond à des proportions, le modèle M initial est souvent un GLM binomial avec lien *logit* ou un GLM Poisson avec lien *log*. De plus, l'algorithme stepwise est utilisé en mode backward, i.e le modèle contient toutes les variables explicatives et l'algorithme devra minimiser l'AIC en enlevant au fur et à mesure les variables du modèle initial. En détails, si n est le nombre de variables et M le modèle initial (par exemple GLM poisson avec lien *log*) construit à partir des n variables explicatives, les étapes de l'algorithme de stepwise sont :

- Étape 1 : Initialiser deux listes de variables $l_{save} = (k_1, k_2, ..., k_n)$ et $l_{rem} = (\varnothing)$ avec les k_i le nom des variables et $\varnothing$ l'ensemble vide.

- Étape 2 : Construire le modèle de départ qui est le modèle initial (par exemple GLM poisson avec lien *log*) construit uniquement à partir des variables dans l_{save}.

- Étape 3 : Construire les n_{save} modèles avec n_{save} le nombre de variables dans l_{save}. Chaque modèle correspond au modèle de départ mais privé de la variable k_i, un élément de la liste l_{save}. Chaque modèle est donc construit avec les $n_{save} - 1$ variables restantes de l_{save}.

- Étape 4 : Construire les n_{rem} modèles avec n_{rem} le nombre de variables dans l_{rem}. Chaque modèle correspond au modèle de départ mais ajouté de la variable k_i présente dans l_{rem}. Chaque modèle est donc construit avec les variables déjà présentes dans le modèle de départ et une variable de l_{rem} en plus. Si $k_i = \varnothing$, alors le modèle de départ est retourné.

- Étape 5 : Parmi les $n_{save} + n_{rem}$ construits en 3 et 4, sélectionner le modèle dont l'AIC est le plus petit.

- Étape 6 : Si le modèle sélectionné est issu de l'étape 3, pour k_s la variable enlevée pour obtenir le modèle sélectionné, supprimer k_s de l_{save} et ajouter k_s à l_{rem}. Sinon, pour k_a la variable ajoutée pour obtenir le modèle sélectionné, supprimer k_a de l_{rem} et ajouter k_a à l_{save}.

- Étape 7 : Répéter les étapes 2 à 6 tant que la minimisation d'AIC est possible et que l_{save} n'est pas vide.

Notez que lors de la construction d'un modèle, par exemple GLM poisson avec lien *log*, il y a toujours un terme constant en plus des coefficients de régression. Ce terme n'est pas nécessairement nul dans cette étude même s'il est possible d'imposer cette contrainte. En pratique, la fonction $stepAIC()$ de la librairie $MASS$ de **R** est sollicité pour cette opération.

Il faut noter que la méthode de stepwise possède néanmoins des fragilités. Cette dernière utilise notamment une unique base de données tout au long de la procédure. En conséquence, le modèle final risque d'être trop simplifié, i.e problème de sur-apprentissage. De plus, lors de la procédure, uniquement un sous-ensemble de modèles possibles est examiné et non tout l'ensemble des modèles possibles (2^n modèles possible en tout). Il est tout à fait possible que le modèle final retenu ne soit pas le modèle optimal en termes de minimisation globale du critère d'AIC.

4.5.5 Définition du modèle et métrique de régression

À la sortie de l'algorithme de stepwise, il reste souvent une cinquantaine de variables. Notez qu'il est possible de réduire davantage, par exemple en utilisant la métrique de régression (section 4.3.1.3) définie précédemment au moment de la présentation de la méthode ensembliste de sélection des variables. Pour ce faire, il faut régresser la variable réponse $y_{freq}(c,t)$ sur les variables retenues par l'algorithme stepwise mais normalisées (centrées en 0 et réduites de telle sorte que la variance vaut 1). Il suffit ensuite, par exemple, de retenir les variables associées aux 30 coefficients de régression (sans compter le terme constant) les plus grandes pour avoir uniquement les 30 variables les plus importantes.

Notez que cette dernière étape n'est pas toujours nécessaire. Typiquement, une cinquantaine de variables semble raisonnable et par conséquent cette dernière étape peut être négligée. Cette dernière est tout de même mentionnée car il se trouve que parfois lors des tests, des cas où à la sortie de l'algorithme stepwise, il reste encore une centaine de variables. Ces cas se présentent notamment lorsque le seuil s utilisé au cours de l'étape 2 est trop grand.

La figure 39 représente les 20 variables explicatives les plus importantes au sens de la métrique de régression parmi les cinquante variables sélectionnées. Notez que les variables climatiques, géologiques et socio-économiques sont toutes présentes même s'il y a beaucoup plus de variables climatiques. Notez aussi que, de nouveau, les variables explicatives climatiques avec une fenêtre de calcul de 3 mois, 6 mois ou 12 mois semblent être majoritaires.

À partir de ces variables explicatives et en utilisant les données des communes reconnues uniquement, le modèle de fréquence construit est un modèle GLM Poisson avec lien *log* et un terme d'exposition $\log(n_{risk})$ où n_{risk} est le nombre de contrats. En d'autres termes, la variable réponse y_{freq} est donnée par :

$$\log(y_{freq}) = \log\left(\frac{n_s}{n_{risk}}\right) = \beta X \Longleftrightarrow \log(n_s) = \beta X + \log(n_{risk}) \tag{73}$$

Avec β les coefficients de régression classique, X le vecteur des variables explicatives, n_s le nombre de sinistres et n_{risk} le nombre de contrats. Il n'est pas nécessaire de détailler davantage car une définition plus complète du GLM est donnée dans la section 3.3.2.1.

4.6 Modèle d'occurrence-fréquence, analyse et interprétation des résultats

Dans cette section, nous allons d'abord définir le modèle d'occurrence-fréquence à partir du modèle d'occurrence (4.3 et 4.4) et du modèle de fréquence (4.5). Une première analyse des résultats sera réalisée et montrera qu'il est nécessaire d'ajouter une correction à la sortie du modèle d'occurrence. Pour cela, une fonction déterministe dite d'influence sera proposée dans un second temps. In fine, les résultats corrigés de l'approche occurrence-fréquence sont très encourageants mais malheureusement, en anticipation de la dernière partie du rapport, aussi incomplets en raison des nouvelles données publiées par Météo-France.

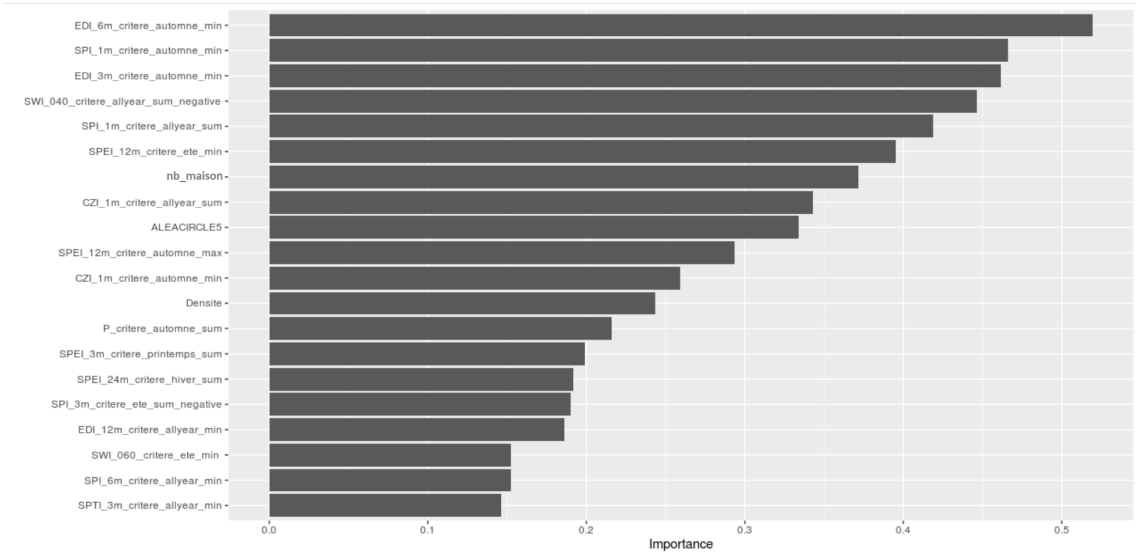

Figure 39 – Plot représentant les 20 variables explicatives les plus importantes au sens de la métrique de régression (section 4.3.1.3) parmi les cinquante variables explicatives sélectionnées. Notez de nouveau que les variables explicatives avec une fenêtre de calcul de 3 mois, 6 mois ou 12 mois semblent être majoritaires.

4.6.1 Principe du modèle

Nous rappelons que nous souhaitons modéliser la charge totale des sinistres pour chaque année. Pour cela, une approche en deux étapes dite occurrence-fréquence est proposée avec la modélisation d'abord de la probabilité y_{occ} qu'une commune, après formulation de la demande, soit reconnue en situation de catastrophe nature par la commission interministérielle, puis ensuite de la proportion y_{freq} de contrats sinistrés. La base de données utilisée est la base de données de catastrophe naturelle construite précédemment et comporte plus de 800 variables explicatives et environ 3×35000 observations.

Le modèle d'occurrence utilise 30 variables explicatives construites à la section 4.2 et sélectionnées à la section 4.3. La construction exacte du modèle d'occurrence, quant à elle, est présentée dans la section 4.4. Pour le modèle de fréquence, ce dernier comporte une cinquantaine de variables explicatives construites à la section 4.2 et sélectionnées via les opérations de la section 4.5. La forme exact du modèle de fréquence est aussi donnée à la fin de la section 4.5.

À présent, soit $A(c,t)$ le montant des sinistres observé pour la commune c durant l'année t, alors le modèle occurrence-fréquence s'écrit :

$$A(c,t) = y_{occ}(c,t) \times y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \tag{74}$$

Avec $n_{risk}(c,t)$ le nombre de contrats signés pour la commune c et l'année t et C_{year} le montant moyen d'un sinistre défini par :

$$C_{year} = \frac{1}{10} \sum_{t=2009}^{2018} \frac{\sum\limits_{c} A(c,t)}{\sum\limits_{c} n_{risk}(c,t)} \tag{75}$$

Notez que le choix de la période 2009-2018 est déterminé par les contraintes sur les données, notamment la disponibilité de ces dernières. Notez aussi que les données de sinistres sont à l'échelle communale pour la période 2009-2018 mais les données de contrats sont à l'échelle communale uniquement pour les années 2016, 2017 et 2018. Ainsi, dans le modèle occurrence-fréquence, y_{occ} et y_{freq} sont des valeurs modélisées, n_{risk} est une variable observée et C_{year} est constant dans le temps.

4.6.2 Analyse des résultats

A présent que le modèle complet est introduit, il est possible d'analyser les résultats de ce dernier. Une première analyse sur un échantillon équilibré est réalisée. L'idée étant que comme le modèle d'occurrence correspond à un problème de classification, il faut par conséquent gérer le problème de déséquilibre entre les différentes classes. En effet, la surreprésentation ou la sous-représentation d'une classe particulière peut nuire à la calibration du modèle. L'échantillon d'entraînement est donc extrait de la base de données de catastrophe naturelle de telle sorte qu'il y a autant de communes déclarées en état de catastrophe naturelle que de communes non déclarés sur chacun des blocs suivants :

- Bloc 1 : les données des années 2009 et 2011.
- Bloc 2 : les données des années 2010, 2013, 2014, 2015 et 2016.
- Bloc 3 : les données des années 2012 et 2017.
- Bloc 4 : les données des années 2018.

Notez que les années sont regroupées de telle sorte que chaque bloc ait approximativement le même nombre de sinistres. L'idée dernière est donc de sous-échantillonner des communes non déclarées afin d'avoir une base équilibrée, i.e autant de communes déclarées en état de catastrophe naturelle que de communes non déclarées.

La figure 40 représente le montant total des charges (IBNR inclus), le nombre de communes déclarées en état de catastrophe naturelle, le nombre de sinistres sous le modèle d'occurrence et le nombre de sinistres sous hypothèse que le modèle d'occurrence est parfait (prédit exactement 1 pour les communes ayant eu l'arrêté interministérielle et 0 sinon). Il est possible réaliser ces graphes sous différentes hypothèses de modèle d'occurrence car y_{occ}, en plus d'être estimé par le modèle, est aussi observé à posteriori. En effet, le délai d'attente pour les demandes d'arrêté est de l'ordre d'un an et demi et donc à ce jour, les arrêtés des années 2016, 2017 et 2018 sont quasiment tous tombés. Notez que la charge totale en 2018 est très sous-estimée du fait que le nombre de communes ayant obtenu l'arrêté de catastrophe naturelle est sous-estimé en 2018. En effet, le nombre de sinistres sous le modèle d'occurrence est sous-estimé pour l'année 2018 mais cette sous-estimation est directement héritée du modèle d'occurrence. En supposant un modèle d'occurrence parfait, la sous-estimation dans le modèle de fréquence pour l'année 2018 est beaucoup moins importante.

Le modèle occurrence-fréquence estime mal la charge totale pour l'année 2018, ce qui est critique puisqu'il s'agit d'une année très sinistrée. En comparant les estimations pour l'année 2018 en fonction du modèle d'occurrence utilisé, le problème semble venir notamment du modèle d'occurrence. Il faut donc essayer d'améliorer ce dernier en premier.

Pour avoir une idée de comment réaliser cette amélioration, nous représentons dans la figure 41, les cartes de France avec pour chacune des 35 000 communes, i.e l'ensemble de la base de données de catastrophe naturelle et non plus le sous-ensemble équilibré, la probabilité observée et prédite pour une commune d'être décrétée en état de catastrophe naturelle en échelle de couleur pour les années 2017 et 2018. Il faut noter que les zones sinistrées prédites sont très similaires aux zones réellement sinistrées. La différence majeure entre les cartes prédites et les cartes réelles est que les zones sinistrées sont plus ou moins étalées. Ainsi, par la suite, nous allons chercher une fonction dite d'influence qui a pour but de rétrécir ou d'élargir les zones sinistrées en fonction des années.

4.6.3 Fonction d'influence

L'idée derrière cette fonction d'influence est que les facteurs sociaux et politiques ne sont pas suffisamment pris en compte. En effet, parmi les variables explicatives sélectionnées lors de la construction du modèle d'occurrence, il n'y a pas de variable socio-économique. Pour remédier à cela, il est possible bien entendu de revenir sur la sélection des variables mais cela impliquerait qu'il faut probablement

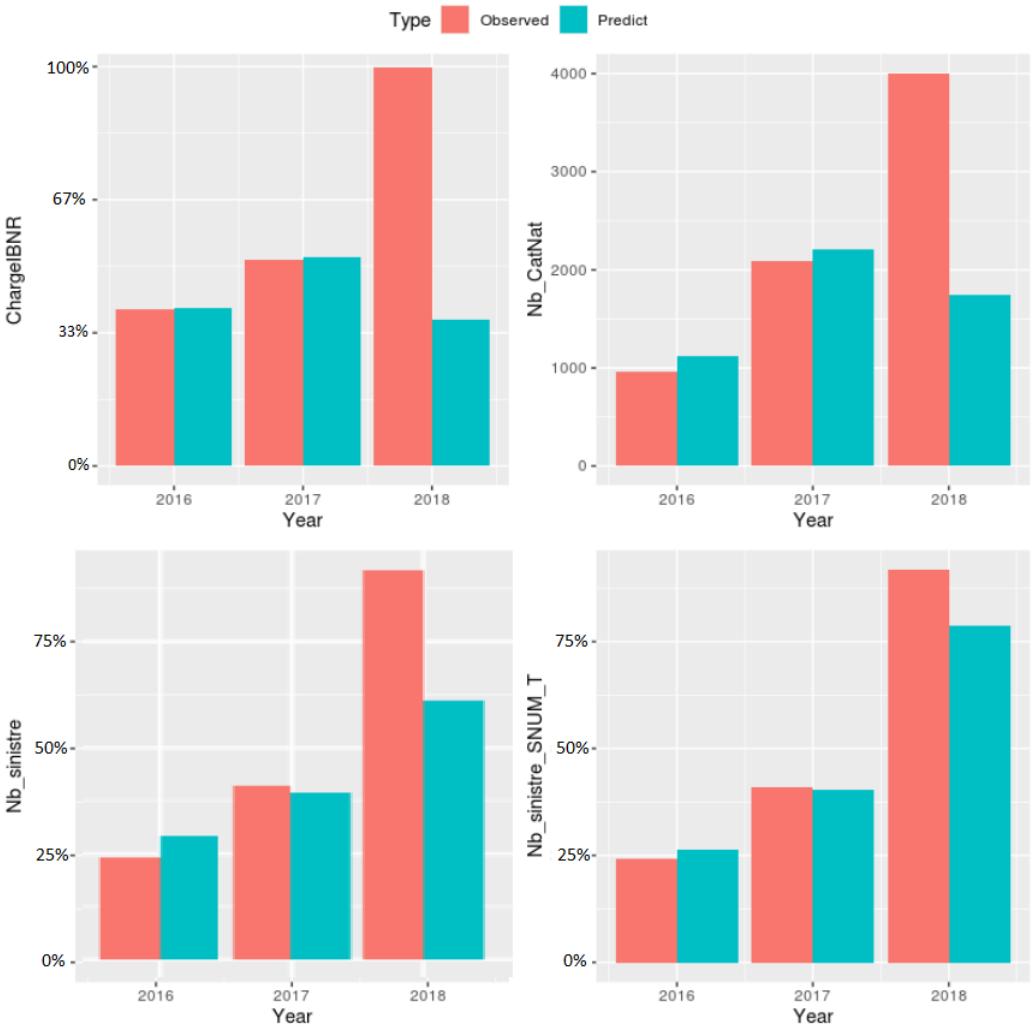

Figure 40 – De gauche à droite et de haut en bas, sur le sous-ensemble des données équilibrées, nous avons le montant total des charges (IBNR inclus) observé et prédit par année, le nombre de communes déclarées en état de catastrophe naturelle par année, le nombre de sinistres sous le modèle d'occurrence par année et le nombre de sinistres sous hypothèse que le modèle d'occurrence est parfait (prédit exactement 1 pour les communes ayant eu l'arrêté interministérielle et 0 sinon). Il faut noter que la charge en 2018 est très sous-estimée du fait que le nombre de communes ayant obtenu l'arrêté de catastrophe naturelle est sous-estimé en 2018. En effet, le nombre de sinistres sous le modèle d'occurrence est sous-estimé pour l'année 2018 mais cette sous-estimation est directement héritée du modèle d'occurrence. En supposant un modèle d'occurrence parfait, la sous-estimation dans le modèle de fréquence pour l'année 2018 est beaucoup moins importante.

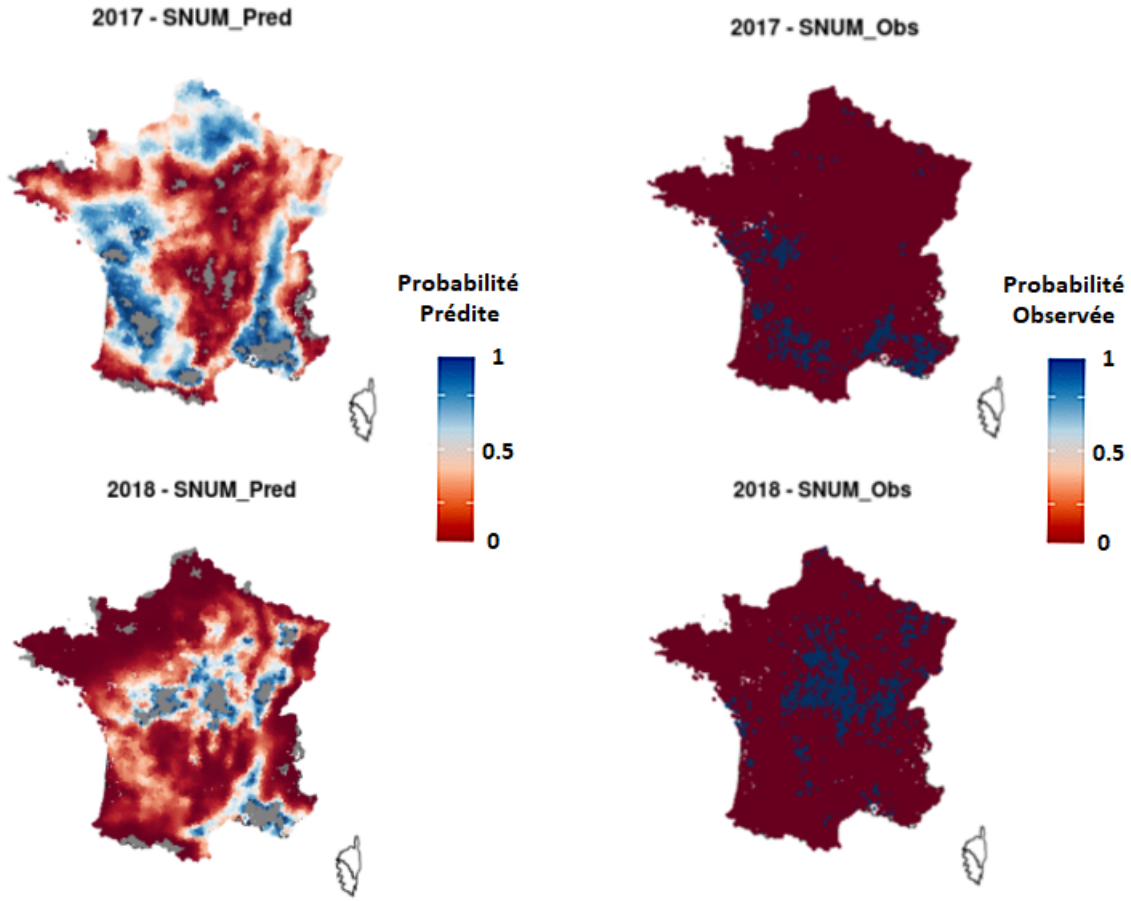

Figure 41 – Cartes de France avec pour chacune des 35 000 communes, i.e l'ensemble de la base de données de catastrophe naturelle et non plus le sous-ensemble équilibré, la probabilité observée (droite) et prédite (gauche) pour une commune d'être décrétée en état de catastrophe naturelle en échelle de couleur pour les années 2017 (haut) et 2018 (bas). Il faut noter que les zones sinistrées prédites (les deux cartes de gauche) sont très similaires aux zones réellement sinistrées (les deux cartes de droite). La différence majeure entre les cartes prédites et les cartes réelles est que les zones sinistrées sont plus ou moins étalées. Ainsi, nous allons chercher une fonction dite d'influence qui a pour but de rétrécir ou d'élargir les zones sinistrées en fonction des années.

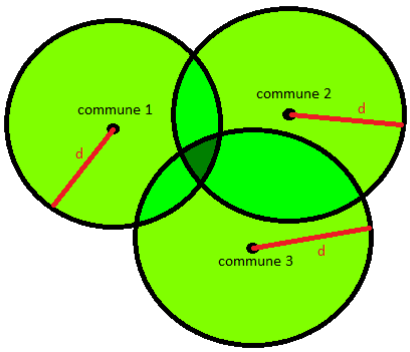

Figure 42 – Illustration des cercles d'influence pour trois communes ainsi que la définition de la distance typique d'influence d, identique pour toutes les communes, qui ne dépend que de l'année considérée.

plus de variables explicatives pour décrire le problème, chose qui est à éviter car cela complexifie le modèle. Ainsi, l'idée est plutôt de proposer une fonction d'influence qui va être appliquée aux sorties du modèle d'occurrence.

Intuitivement, chaque année a sa particularité et le modèle étant calibré sur toutes les années, correspond en fait à une situation moyenne. Il est donc nécessaire d'adapter la prédiction en fonction de l'année à l'aide d'une variable caractéristique de chaque année. Dans un premier temps, nous allons prendre le nombre de communes demanderesses comme variable caractéristique de chaque année. En effet, naïvement nous estimons que s'il y a plus de demandes, en raison par exemple des élections, il y aura plus de pression sur les maires et par effet domino sur les différents intermédiaires jusqu'à la commission interministérielle. Ainsi, cette variable de nombre de communes demanderesses est potentiellement un bon reflet de la tension sociale propre à chaque année.

Sur la forme exacte de cette fonction d'influence, notez d'abord que les variables des modèles d'occurrence et de fréquence sont surtout climatiques et donc sont continues dans l'espace. Par conséquent, si une commune initiale a de grande chance d'être déclarée en état de catastrophe naturelle, i.e le modèle d'occurrence prédit une valeur proche de 1, alors ses voisins doivent aussi avoir une grande chance d'être déclarés en état de catastrophe naturelle, puis les voisins de ses voisins aussi, et ainsi de suite avec la probabilité qui décroît lorsque nous nous éloignons de la commune initiale. Dans la figure 42, cette idée est illustrée en représentant les cercles d'influence de plusieurs communes ainsi que la distance typique, identique pour toutes les communes, qui ne dépend que de l'année considérée.

Il faut préciser que dans cette perspective, plusieurs cercles d'influence peuvent se croiser et par conséquent, la distance typique ne peut s'observer qu'au niveau de la frontière entre les communes décrétées en état de catastrophe naturelle et les communes non décrétées, comme l'illustre la figure 43. Ainsi, cette dernière s'apparente à l'épaisseur d'une frontière séparant les communes décrétées et non décrétées. Cette distance typique peut schématiquement aussi être vue comme l'épaisseur des contours blancs dans les cartes de la figure 41 bien que les bornes exactes des probabilités des communes situées au niveau des contours sont à précisées.

À présent, avec les idées suggérées précédemment, il est possible de proposer une fonction d'influence. Bien entendu, ce dernier est arbitraire et de fondement théorique fragile si ce n'est que nous faisons une analogie entre la sécheresse et le séisme. Nous allons supposer que la décroissance de la probabilité d'être décrétée en état de catastrophe naturelle est exponentielle dans l'expression de la fonction d'influence et que la pression sociale et politique s'exprimera à travers la distance typique de décroissance illustrée dans les schémas précédents. Ainsi, en reprenant les notations du 4.6.1, le modèle occurrence-fréquence corrigé par la fonction d'influence est défini par :

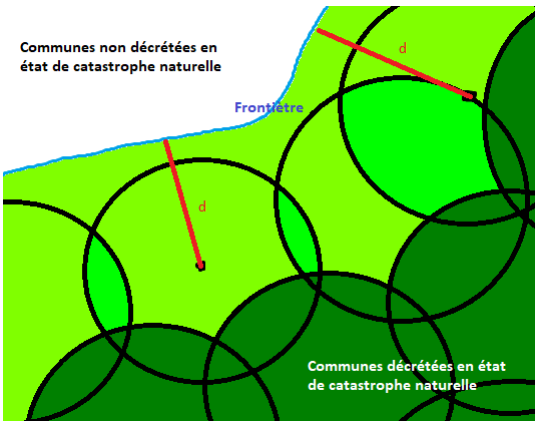

Figure 43 – Illustration de l'interaction des cercles d'influence des communes et de l'observation de la distance typique d'influence d au niveau de la frontière entre les communes décrétées en état de catastrophe naturelle et les communes non décrétées.

$$A(c,t) = f_{infl}(y_{occ}(c,t)) \times y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \tag{76}$$

Avec :

$$f_{infl}(y_{occ}(c,t)) = \left(\frac{y_{occ}(c,t) - m_t}{M_t - m_t}\right)\left(\frac{d_{ref}}{d_t}\right)^{\alpha} \tag{77}$$

Où :

$$\begin{cases} \alpha = 0.5 \\ M_t = \max\limits_{c} y_{occ}(c,t) \\ m_t = \min\limits_{c} y_{occ}(c,t) \end{cases} \tag{78}$$

Les m_t et M_t servent à normaliser les probabilités prédites par le modèle d'occurrence, i.e les projeter dans l'intervalle $[0,1]$. Une fois projetées, la probabilité de chaque commune d'être déclarée en état de catastrophe naturelle est calibrée à l'aide de d_t, i.e le nombre de communes demanderesses, qui reflète la pression sociale et politique de l'année t et s'interprète comme une distance typique de décroissance de l'influence des communes sur leurs voisins. Le d_{ref} correspond ainsi à la situation moyenne sur laquelle est calibrée le modèle sans fonction d'influence. Bien que la distance d_{ref} peut potentiellement être mesurée (figure 43), cette dernière sera considérée comme un paramètre de la fonction d'influence au même titre que α qui est estimé à 0.5 et qui décrit la puissance de décroissance de l'influence d'une commune sur ses voisins. L'estimation de d_{ref} est réalisée à partir de la figure 44 représentant le nombre de communes demanderesses par année. En regardant quelles sont les années devant être corrigées vers le haut ou vers le bas et quelles sont les années qui ne nécessitent pas de correction, la distance d_{ref} est fixée approximativement à 2800.

4.6.4 Résultats corrigés

La figure 45 représente le nombre de communes décrétées en état de catastrophe naturelle respectivement observé et prédit sans et avec fonction d'influence. Il faut remarquer que lorsque la fonction d'influence est appliquée, les prédictions semblent se dégrader légèrement pour les années 2009 et 2017 mais s'améliorent nettement pour les autres années, et notamment l'année 2018. L'approche avec la fonction d'influence semble plutôt encourageante dans le cadre du sous-ensemble équilibré.

Nous allons appliquer à présent le modèle d'occurrence-fréquence avec la fonction d'influence sur l'ensemble des données et non plus sur le sous-ensemble des données équilibrées. Dans la figure 46,

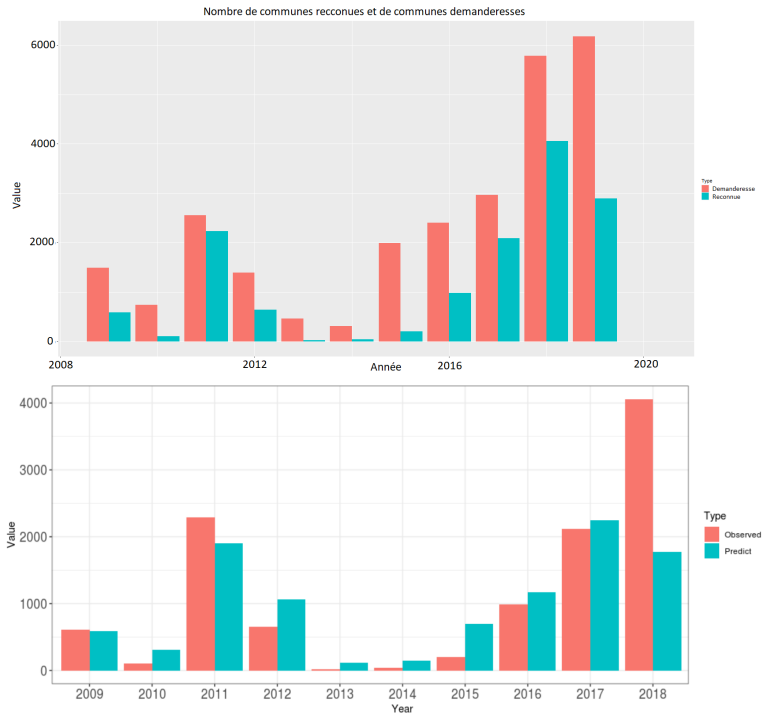

Figure 44 – En haut, nous avons l'évolution du nombre de communes ayant formulé une demande de reconnaissance de l'état de catastrophe naturelle (rouge) et parmi ces demandes, le nombre de communes effectivement reconnues en état de catastrophe naturelle (bleu) par la commission interministérielle. En bas, nous avons le nombre de communes décrétées en état de catastrophe naturelle observé (rouge) et prédit (bleu) par le modèle occurrence-fréquence sans fonction d'influence. En regardant quelles sont les années devant être corrigées vers le haut ou vers le bas et quelles sont les années qui ne nécessitent pas de correction, la distance d_{ref} est fixée approximativement à 2800.

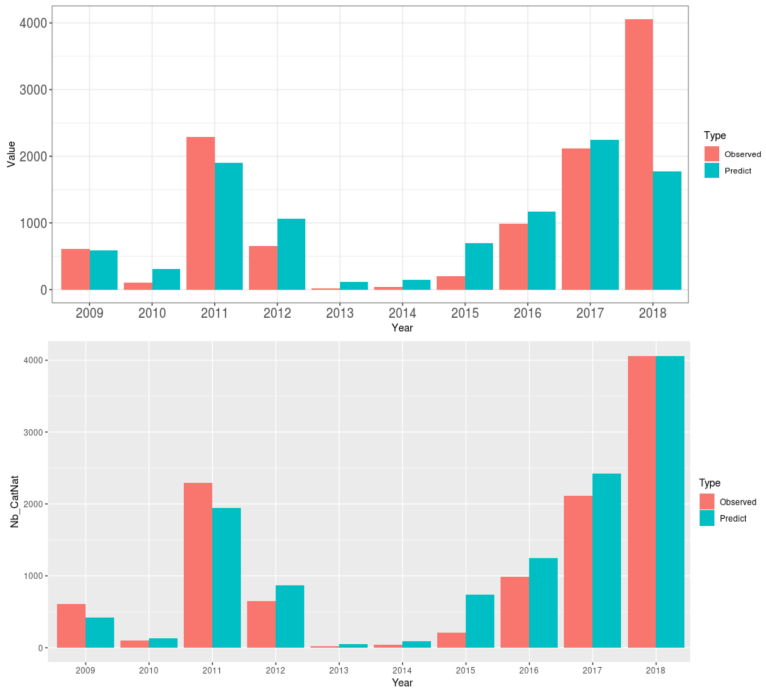

Figure 45 – De haut en bas, le nombre de communes décrétées en état de catastrophe naturelle observé (rouge) et prédit (bleu) respectivement sans et avec fonction d'influence sur le sous-ensemble des données équilibrées. Il faut noter que lorsque la fonction d'influence est appliquée, les prédictions semblent se dégrader légèrement pour les années 2009 et 2017 mais s'améliorent nettement pour les autres années, et notamment l'année 2018. L'approche avec la fonction d'influence semble plutôt encourageante dans le cadre du sous-ensemble équilibré.

Figure 46 – De haut en bas, le nombre de communes décrétées en état de catastrophe naturelle observé (rouge) et prédit (bleu) respectivement sans et avec fonction d'influence sur l'ensemble des données. L'application de la fonction d'influence apporte clairement une amélioration mais ne corrige pas pour autant l'ensemble des surestimations. Les années 2018 et 2019 semblent être bien corrigées mais pas les années antérieures à 2018.

nous représentons le nombre de communes décrétées en état de catastrophe naturelle observé et prédit respectivement sans et avec fonction d'influence sur l'ensemble des données. L'application de la fonction d'influence apporte clairement une amélioration mais ne corrige pas pour autant l'ensemble des surestimations. Les années 2018 et 2019 semblent être bien corrigées mais pas les années antérieures à 2018.

Pour expliquer cela, en anticipation de la suite, nous savons qu'en 2018, il y a eu une grande réforme des règles de décret d'arrêté de catastrophe naturelle avec l'utilisation de l'indice SWI (Soil Wetness Index) de Météo France. Cette réforme peut expliquer cette différence de la qualité de correction avant et après 2018 apportée par la fonction d'influence. De plus, comme le montre la figure 41, même sans la fonction d'influence, le modèle semble bien situer les zones sinistrées. Ainsi, de manière générale, cette approche est plutôt prometteuse, notamment après l'application des nouveaux critères de reconnaissance.

Néanmoins, même si en théorie le nombre de communes demanderesses sont disponibles dès lors que les communes font la demande de décret, en pratique, ces données sont publiées en même temps que la décision de la commission interministérielle et donc indisponible au moment de l'application du modèle, ce qui nous incite à chercher d'autres variables pour caractériser cette notion de pression sociale et politique. D'autre part, au milieu de l'année 2021, Météo France a publié ses propres données SWI Uniforme qui servent de base dans les décisions de la commission interministérielle ([38]). Ainsi, à partir de 2018 et selon la nouvelle procédure de reconnaissance de l'état de catastrophe naturelle ([39]), l'éligibilité des communes au décret est calculée de manière déterministe à l'aide du SWI Uniforme, i.e dès lors qu'une commune éligible fait une demande d'arrêté, le décret est à priori accordée automatiquement. Inversement, pour les communes non éligibles, le refus sera aussi automatique s'il y a demande d'arrêté.

La nouvelle procédure rend le modèle d'occurrence peu utile puisque l'inconnue porte désormais non plus sur la décision de la commission interministérielle mais sur les communes et si ces dernières vont faire la demande d'arrêté ou pas. L'approche du modèle d'occurrence avec la fonction d'influence, bien que prometteuse, n'est malheureusement plus parfaitement à jour. Il est donc nécessaire de le mettre à jour voire même proposer une modélisation alternative.

4.7 Conclusion

Dans cette approche annuelle à résolution communale, nous avons d'abord construit une base communale appelée base de catastrophe naturelle. Cette base contient des variables climatiques, des variables géologiques, et des variables socio-économiques. Notez que par rapport à la base départementale, les variables climatiques sont construites à l'aide des données de température et de précipitation de bonne qualité, i.e mensuelle, et les variables socio-économiques sont nouvelles. Nous avons construit ensuite un modèle d'occurrence et un modèle de fréquence en définissant au préalable des méthodes de sélection propre à chaque modèle basées sur des principes ensemblistes, des distances statistiques, ou des algorithmes qui ont déjà fait leurs preuves dans la littérature. Notez que les méthodes de sélection établies sont applicables de manière générale dans d'autres problèmes de modélisation. Le modèle complet occurrence-fréquence obtenu, moyennant la définition d'une fonction déterministe appelée fonction d'influence, produit des résultats cohérents et encourageants. Les mauvaises estimations pour les années antérieures à 2018 s'expliquent par l'application des nouveaux critères de reconnaissance en 2018 et témoignent de la nécessité de prendre en compte l'évolution de la réglementation dans la modélisation du risque de subsidence au sein du régime CatNat de la CCR.

Néanmoins, vers avril 2021, en soutien aux évolutions règlementaires, Météo-France rend publiques les données de SWI Uniforme. Ces données permettent de calculer exactement les critères de reconnaissance actuels et compte tenu du circulaire associé à la réforme réglementaire ([39]) qui invite la commission interministérielle à une automatisation des reconnaissances de catastrophe naturelle sur la base de ces nouveaux critères, le modèle d'occurrence devient peu intéressant puisque l'inconnue porte désormais non plus sur la décision de la commission interministérielle mais sur les communes et si ces dernières vont faire la demande d'arrêté ou pas. Ainsi, le modèle occurrence-fréquence n'est plus parfaitement à jour et il est nécessaire de le mettre à jour ou de proposer une modélisation alternative en intégrant les nouvelles données de SWI Uniforme.

5 Apport des données SWI, approche via les critères d'éligibilité

5.1 Introduction

Dans la continuité de la démarche parcimonieuse de l'étude, nous allons apporter une réponse au fait que le modèle occurrence-fréquence n'est plus à jour et qu'il est faut intégrer les nouvelles données de SWI Uniforme dans la modélisation. L'idée générale de cette partie est que nous allons tenter différentes modélisations et comparer les résultats aux estimations de charge IBNR annuelle totale de la CCR ([43]). Le modèle final retenu sera le modèle le plus cohérent dans l'ensemble et qui répond à l'objectif actuariel de l'étude.

Ainsi, dans cette partie, nous allons dans un premier temps intégrer les nouvelles données publiées par Météo France ([38]) dans la base de données de catastrophe naturelle construite précédemment. Notez que ces données portent sur l'indice de SWI Uniforme et permet de calculer exactement les critères de reconnaissance actuels. Dans un second temps, nous proposerons plusieurs modèles afin de mettre en évidence l'intérêt et la pertinence de la théorie des modèles inflatés pour cette étude. Puis dans un troisième temps, nous utiliserons cette théorie pour construire un modèle saisonnier composé de quatre modèles sous-jacents et qui est plus adapté aux nouveaux critères de reconnaissance appliqués depuis 2018. Il faut noter que cette modélisation saisonnière nécessitera notamment de diviser l'ensemble des données en quatre sous-ensembles. Enfin dans un dernier temps, nous résumerons le rôle de l'actuaire et les points importants de son travail dans la perspective d'un produit sur le risque de subsidence.

5.2 Intégration des nouvelles données

Afin d'intégrer les nouvelles données de Météo-France, nous allons d'abord rappeler l'évolution historique des critères de reconnaissance dans le régime de catastrophe naturelle français afin de mettre en évidence le contexte dans lequel les nouveaux critères de reconnaissance ont été introduits. Nous détaillerons ensuite le calcul exact des critères de reconnaissance utilisés par la commission interministérielle depuis 2018. Les variables explicatives qui en découlent seront bien entendu intégrées dans les modélisations des sections suivantes.

5.2.1 Évolution des critères de reconnaissance et SWI Uniforme

Nous allons dans cette section décrire brièvement l'évolution des critères de reconnaissance d'état de catastrophe naturelle ([1],[39]). Il faut d'abord rappeler qu'il n'y a pas encore de consensus scientifique sur la caractérisation du phénomène de sécheresse. La plupart des pays utilisent jusqu'à l'heure actuelle un ensemble d'indices construit sur mesure en fonction des données disponibles et adapté à leurs situations météorologiques et géologiques. Il s'agit donc des critères empiriques sans véritable fondement scientifique incontestable et qui évoluent fréquemment en fonction des besoins et de la politique.

En France jusqu'en 1999, la sécheresse était appréhendée uniquement selon une approche strictement météorologique, i.e sur la base des déficits pluviométriques observés. À partir de 2000, l'autorité a introduit un critère reposant sur le calcul du bilan hydrique des sols superficiels avec l'aide de 200 stations météorologiques et de leurs 20 ans de données d'observations. Chaque commune étant rattachée à une station de référence, la France est découpée en zone de pluviométrie homogène appelée zonage "Aurore". Les critères de reconnaissance, appelé "critères 2000", reposent sur trois conditions, à savoir :

- La présence d'argile sur le territoire de la commune.
- Une période de quatre trimestres consécutifs durant laquelle la réserve en eau des sols est inférieure à la normale.

- Une décade de la période de recharge des nappes, i.e de janvier à mars, où la réserve hydrique du sol est inférieure à 50% de la réserve normale.

En raison de l'épisode de sécheresse de 2003 marqué par une sécheresse en été très différente des années précédentes et qui, en utilisant les critères en vigueur, aurait conduit la commission interministérielle à refuser la reconnaissance à presque la totalité des communes demanderesses alors même que d'importants dégâts étaient observés sur une grande partie du territoire métropolitain. Ainsi, Météo-France, sous la demande de l'autorité, élabore des nouveaux critères afin de prendre en compte la sécheresse estivale. En 2004, les critères sont devenus :

- La présence d'argile sur le territoire de la commune.
- Le rapport de la moyenne de la réserve hydrique du 3è trimestre sur la moyenne hydrique normale inférieur à 20% .
- Le nombre de décades pendant lesquelles le réservoir hydrique était égal à zéro se situe au 1er ou 2è rang par rapport à la période 1989-2003.

Un premier assouplissement en janvier 2005 augmente le 20% à 21% et étend le critère de classement jusqu'au 3è rang. Un nouvel assouplissement s'opère cinq mois plus tard avec le critère de 21% remplacé par une durée de retour de la moyenne des réserves en eau du sol du troisième trimestre supérieure à 25 ans.

Au fur et à mesure des années, l'ajout et la modification sans cesse des critères rendent ces derniers complexes et difficilement compréhensibles pour le public. Ainsi, en 2009, grâce aux innovations techniques, et notamment en matière de récolte et simulation des données, sous l'impulsion du gouvernement, Météo-France introduit un indice d'humidité du sol appelé SWI (Soil Wetness Index) pour remplacer tous les précédents critères. Les critères basés sur le SWI a notamment pour objectif d'être plus simple et plus systématique dans leurs applications. Néanmoins, avant de préciser ces derniers, nous allons définir brièvement l'indice SWI.

L'indice SWI est issu d'un modèle hydrométéorologique de Météo-France appelé SIM et qui est composé de trois modules, à savoir SAFRAN, ISBA et MODCOU. Le module SAFRAN ou Système d'Analyse Fournissant des Renseignements Adaptés à la Nivologie est un module qui permet de reconstruire des profils verticaux de l'atmosphère (température, vent, ...) sur des zones climatiques homogènes avec un pas de temps horaire. Le module ISBA ou Interaction Sol-Biosphère-Atmosphère est un module pour décrire l'état du sol et les échanges sol-plante-atmosphère. Et le module MODCOU est un modèle hydrogéologique qui permet de simuler les débits des rivières et les niveaux piézométriques des aquifères. L'indice SWI ([40], [20]) est produit à l'aide des données du module ISBA à l'échelle du maillage SAFRAN (8 981 mailles de 8 km de côté couvrant la France entière) au pas de temps horaire et est défini, pour une maille m, par :

$$SWI(m) = \frac{w - w_{wilt}}{w_{fc} - w_{wilt}} \tag{79}$$

Avec w le contenu en eau du sol (sur environ 10 m de profondeur comportant la couche de surface, la couche racinaire et la couche profonde), w_{fc} le contenu en eau au point de flétrissement et w_{wilt} le contenu en eau à la capacité au champ. Il faut préciser que w_{fc} correspond à une situation limite où les plantes ne peuvent plus prélever l'eau du sol et w_{wilt} à la situation opposée où le sol est saturé en eau. Ainsi, il faut noter que l'indice SWI est positif et plus le SWI est petit, plus le sol est sec.

Les critères de reconnaissance définis à l'aide de SWI en 2009 est au nombre de deux, à savoir un critère hivernal pour capturer les sécheresses longues comme en 1989-1990 et un critère estival pour capturer les sécheresses estivales comme en 2003. Il suffit ainsi de vérifier l'un des deux critères climatiques et le critère géotechnique pour que la commune soit reconnue en situation de catastrophe naturelle. Plus précisément, ces critères sont :

- Critère géotechnique : présence d'argile sur la commune et au moins 10% de la surface communale est impactée.

- Critère hivernal : un indice SWI, calculé sur une période de 4 trimestres consécutifs, inférieur à la normale (période 1971-2000), et comportant une décade du trimestre de fin de recharge (janvier, février et mars) inférieur à 80% de l'indice normal (choc hivernal).

- Critère estival : le rapport de la moyenne de SWI du 3è trimestre à la moyenne de SWI normal inférieur à 70%, le nombre de décades pendant lesquelles l'indice est inférieur à 0,27 se situe au 1er, 2è ou 3è rang comparé à la période 1989-2011 et la durée de retour de la moyenne des SWI des 9 décades de juillet à septembre supérieure à 25 ans

Néanmoins, la sécheresse de 2011 contrairement aux précédents épisodes de sécheresse est marquée par un caractère printanier. Par conséquent, la commission interministérielle a retenu un nouveau critère climatique de reconnaissance qui s'ajoute aux précédents, à savoir :

- Critère printanier : la durée de retour de la moyenne des SWI des 9 décades d'avril à juin supérieure à 25 ans.

Avec ces critères, un assouplissement intervient en 2016 sur le critère géologique avec la suppression de la contrainte sur la superficie impactée.

En 2019 intervient une nouvelle tentative de simplification des critères avec l'utilisation du SWI Uniforme ([39]) calculé sur une fenêtre glissante de taille 3 mois sur les SWI journalières avec un pas de temps mensuel. Une année est ainsi décrite à l'aide de 12 indices SWI Uniforme et celui du mois i comporte les informations du mois i mais aussi les informations des deux mois précédents. Par ailleurs, il faut noter que les indices sont définis pour chaque maille et une commune est attachée à une ou plusieurs mailles. Les critères de reconnaissance actuels, appliqués pour les demandes d'arrêté concernant les années 2018 et suivantes, sont donc :

- Critère géotechnique : surface argileuse avérée de la commune supérieure ou égale à 3%.

- Critère hivernal : l'un des SWI Uniforme de janvier, février ou mars d'une des mailles attachées à la commune demanderesse a une durée de retour de 25 ans.

- Critère printanier : l'un des SWI Uniforme de avril, mai ou juin d'une des mailles attachées à la commune demanderesse a une durée de retour de 25 ans.

- Critère estival : l'un des SWI Uniforme de juillet, août ou septembre d'une des mailles attachées à la commune demanderesse a une durée de retour de 25 ans.

- Critère automnal : l'un des SWI Uniforme de octobre, novembre ou décembre d'une des mailles attachées à la commune demanderesse a une durée de retour de 25 ans.

Nous précisons qu'une durée de retour est calculée en comparant le SWI Uniforme établi pour un mois donné avec les indicateurs établis pour ce même mois au cours des cinquante dernières années. Une durée de retour de 25 ans correspond ainsi à une position de l'indice au 1er ou au 2è rang lorsque nous ordonnons dans l'ordre décroissant les 50 valeurs de SWI.

Notez que les quatre critères climatiques sont favorables aux communes puisqu'il suffit que l'un des SWI Uniforme vérifie le critère de durée de retour de 25 ans sur les $12n$ SWI Uniforme de la commune avec n le nombre de mailles attachées à la commune. Cette définition explique aussi pourquoi dans certain cas une commune est reconnue en état de catastrophe naturelle alors qu'une commune limitrophe

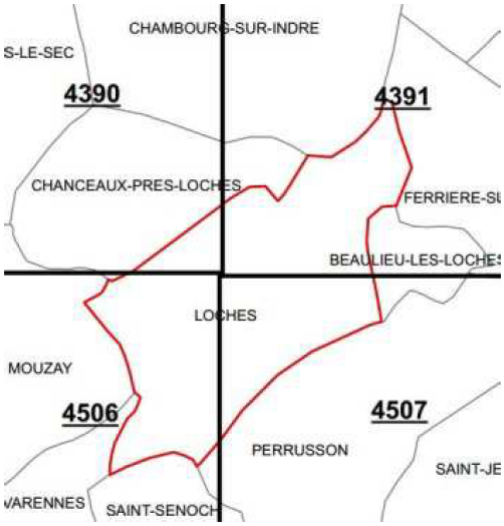

Figure 47 – Illustration extraite du circulaire ([39]). Dans cette configuration, si seule la maille 4390 remplit le critère météorologique, alors l'ensemble de la commune de Loche est reconnue en état de catastrophe naturelle. Ce n'est pas le cas de la commune Beaulieu-les-Loches car elle est rattachée aux mailles 4391 et 4507, qui elles ne remplissent pas les critères météorologiques.

ne l'est pas, car cette dernière est associée à des mailles géographiques différentes ne remplissant pas le critère météorologique. Cette remarque est illustrée dans la figure 47 extraite du circulaire (([39])).

Par ailleurs, notez que le critère géotechnique portant sur la composition argileuse du sol est évalué à partir des données du Bureau de Recherche Géologique et Minière (BRGM) et complété par des études du sol si la surface exposée est inférieure à 3%. Dans les figures 48 et 49, nous représentons respectivement la carte avant 2019 (extrait de [1]) et après 2019 (extrait de [41]) de l'exposition du territoire français au phénomène de retrait-gonflement des argiles. Notez qu'il y a clairement une différence en terme de territoire exposé avec notamment une réévaluation des zones de fortes expositions, qui tend à favoriser les communes. Cette mise à jour s'explique par une avancée technique dans les méthodes de mesure et de traitement des données mais aussi par un besoin politique suite à loi ELAN qui introduit dans l'article 68 des dispositions réglementaires nécessitant une identification des zones exposées au phénomène.

In fine, en parcourant l'histoire de l'évolution des critères de reconnaissance de situation de catastrophe naturelle, nous constatons que c'est loin d'une science exacte. Les critères évoluent en fonction des besoins et de la politique et à chaque nouveau cas de sécheresse non observé historiquement, les critères sont amenés à être modifiés ou complétés. À cette constante évolution des critères de reconnaissance s'ajoute probablement par dessus un arbitrage dans les décisions puisque les données pour vérifier les critères des années antérieures à 2018 ne sont pas disponibles. Ceci implique notamment une incompréhension des décisions d'arrêté pour les maires et les habitants des communes, d'où les simplifications des critères de reconnaissance en 2019.

D'autre part, cette évolution constante rend aussi les études historiques des reconnaissances des sécheresses difficiles car la corrélation entre les reconnaissances des sécheresses d'une année à l'autre est grandement affaiblie. En d'autres termes, une sécheresse d'hiver de l'année passée peut ne pas être une sécheresse d'hiver si nous utilisons les critères de cette année. Ainsi, afin d'avoir des données homogènes, l'un des soucis de l'étude concerne notamment l'actualisation des reconnaissances passées dans les termes d'aujourd'hui. Notez qu'il est nécessaire de faire cette mise à jour car sinon, deux tiers des données disponibles dans le portefeuille seront inexploitables. Néanmoins, cette actualisation a ses limites car plus il faut remonter dans le temps, plus les critères de reconnaissance diffèrent des critères

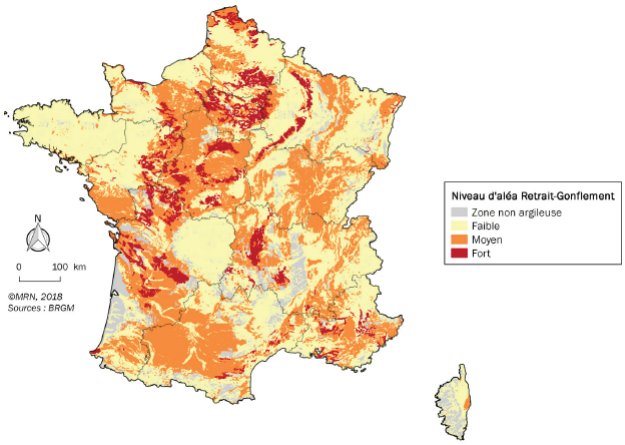

Figure 48 – Ancienne carte avant 2019 de l'exposition du territoire français au phénomène de retrait-gonflement des argiles (extrait de [1]).

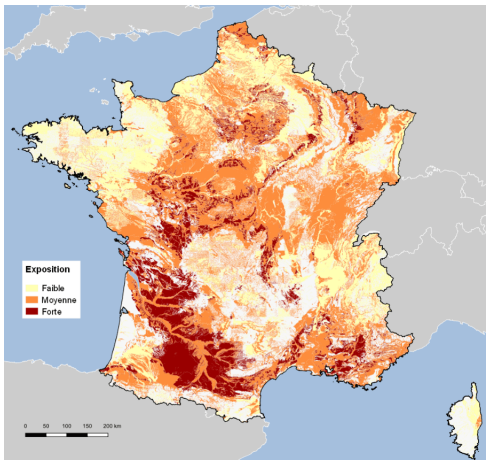

Figure 49 – Nouvelle carte après 2019 de l'exposition du territoire français au phénomène de retrait-gonflement des argiles (extrait de [41]).

actuels.

5.2.2 Récapitulatif de la base complète

La mise à disposition des données de SWI Uniforme par Météo-France([38]) nous invite à reconsidérer la base de données de catastrophe naturelle. Typiquement, les données de SWI modélisées via des instruments de satellite ne sont plus utiles à priori puisque nous pouvons désormais remplacer ces données modélisées par des vraies données produites par Météo-France et utilisées par la commission interministérielle.

Les données de Météo-France comportent bien entendu les 50 ans de données nécessaire pour calculer les critères actuels de reconnaissance de catastrophe naturelle (section 5.2.1). À l'aide de ces données, pour un mois m d'une année t, nous disposons des valeurs de SWI Uniforme du mois m, que nous rappelons est un indice mensuel, pour les 50 années avant t. Par exemple, pour le mois de janvier de l'année 2018, nous disposons aussi des SWI Uniforme du mois de janvier des années 1969, 1970, ..., 2018. Et ceci pour chaque commune c de la France métropolitaine. Ainsi, pour chaque mois m et chaque couple (c,t), en notant $\mathcal{P}$ l'ensemble des 50 valeurs de SWI Uniforme associées, nous définissons les indices suivants :

$$\begin{cases} MIN.SWI.m(c,t) = SWI^U(m,c,t) \\ MIN.RANK.m(c,t) = \dfrac{rg(SWI^U,\mathcal{P})}{50} \\ MIN.DELTA.MEAN.m(c,t) = \overline{SWI^U}^{\mathcal{P}} \\ MIN.DELTA.MED.m(c,t) = med(SWI^U,\mathcal{P}) \\ MIN.ZSCORE.m(c,t) = \dfrac{SWI^U(m,c,t) - MIN.DELTA.MEAN.m(c,t)}{\sigma_{\mathcal{P}}} \end{cases} \tag{80}$$

Avec $SWI^U(m,c,t)$ la valeur du SWI Uniforme du mois m pour la commune c et l'année t, $rg(SWI^U,\mathcal{P})$ le rang de $SWI^U(m,c,t)$ parmi les 50 valeurs dans $\mathcal{P}$ (que nous rappelons dépend de m, c et t) avec 1 pour la plus petite valeur et 50 pour la plus grande, $\overline{SWI^U}^{\mathcal{P}}$ la moyenne des 50 valeurs de SWI Uniforme dans $\mathcal{P}$, $med(SWI^U,\mathcal{P})$ la médiane des 50 valeurs de SWI Uniforme dans $\mathcal{P}$, et $\sigma_{\mathcal{P}}$ l'écart-type des 50 valeurs de SWI Uniforme dans $\mathcal{P}$. À partir de ces quantités, nous définissons aussi les minimums de ces quantités sur l'année t, i.e :

$$\begin{cases} MIN.SWI = \min\limits_{m=1,2,...,12} MIN.SWI.m(c,t) \\ MIN.RANK = \min\limits_{m=1,2,...,12} MIN.RANK.m(c,t) \\ MIN.DELTA = \min\limits_{m=1,2,...,12} MIN.DELTA.MEAN.m(c,t) \\ MIN.DELTA = \min\limits_{m=1,2,...,12} MIN.DELTA.MED.m(c,t) \\ MIN.ZSCORE = \min\limits_{m=1,2,...,12} MIN.ZSCORE.m(c,t) \end{cases} \tag{81}$$

Parallèlement, nous définissons quatre variables catégorielles $0/1$ qui correspondent à la satisfaction ou non des quatre critères météorologiques de reconnaissance de catastrophe naturelle présentés à la section 5.2.1. Ces variables catégorielles sont :

$$\begin{cases} HIVER(c,t) = I\left(\left[\sum\limits_{m=1}^{3} I(MIN.RANK.m(c,t) \geq 0.04)\right] \geq 1\right) \\ PRINTEMPS(c,t) = I\left(\left[\sum\limits_{m=4}^{6} I(MIN.RANK.m(c,t) \geq 0.04)\right] \geq 1\right) \\ ETE(c,t) = I\left(\left[\sum\limits_{m=7}^{9} I(MIN.RANK.m(c,t) \geq 0.04)\right] \geq 1\right) \\ AUTOMNE(c,t) = I\left(\left[\sum\limits_{m=10}^{1} 2I(MIN.RANK.m(c,t) \geq 0.04)\right] \geq 1\right) \end{cases} \tag{82}$$

Avec I la fonction indice qui vaut 1 si la condition est satisfaite et 0 sinon. Ainsi, en combinant ces quatre variables catégorielles, il est possible de définir une variable catégorielle 0/1 qui indique si pour une année t donnée, la commune c est éligible ou pas au décret de catastrophe naturelle. Cette variable est :

$$Eligible(c,t) = I(HIVER(c,t) + PRINTEMPS(c,t) + ETE(c,t) + AUTOMNE(c,t) \geq 1) \quad (83)$$

Ainsi, toutes ces nouvelles variables explicatives climatiques sont ajoutées à la base de données de catastrophe naturelle construite dans la partie précédente.

Notez que dans la figure 29, la variable qui a obtenu la meilleure statistique DTS, qui est la plus robuste des statistiques de la section 4.3.2, est une variable SWI modélisé de quantile 5%. Cette variable correspond plus exactement au nombre de points de la période automnale qui sont inférieurs au quantile 5% calculé sur les points de l'année entière. En comparant cela à la variable catégorielle $AUTOMNE$, il faut remarquer que les deux définitions sont très proches. Par conséquent, il semble que cette variable est en fait une variable qui approche le critère automnal actuel, ce qui témoigne en faveur de l'intérêt de la statistique DTS qui a sélectionné cette dernière.

Dans la figure 50, le nombre de communes éligibles au sens des critères de reconnaissance actuels et le nombre de communes reconnues par la commission interministérielle mais non éligibles au sens des critères actuels sont représentés. Constatez que pour les années 2011 et 2018, il y a un grand nombre de communes éligibles. Les critères actuels mettent donc en évidence les sécheresses des années 2011 et 2018, ce qui est à priori l'effet recherché. Par ailleurs, le nombre de communes reconnues mais non éligibles au sens des critères de reconnaissance actuels est très faible pour l'année 2018 et diminue en 2019 mais est élevé pour certaines années antérieures à 2018. Ce qui est logique puisque les nouveaux critères ne sont entrées en vigueur qu'à partir de 2018. Néanmoins, cela rejoint la remarque de la section 5.2.1, à savoir qu'une sécheresse d'hiver d'une année passée peut ne pas être une sécheresse d'hiver si nous utilisons les critères actuels. Il faut aussi noter que pour 2011, malgré l'utilisation des anciens critères, il y a peu de communes reconnues mais non éligibles comparé au nombre de communes éligibles. En effet, il s'agit d'une année où il y a eu de nombreux recours et la commission a dû revoir les critères compte tenu de la situation particulière. Cela ne s'est pas réopéré de la même manière en 2016 et 2017 car d'une part les sécheresses semblent être moins importantes comparées à 2011 et d'autre part, il y a les nouveaux critères de reconnaissance qui sont déjà en discussion.

Pour corriger le problème d'incohérence entre les différents critères, lorsqu'il faudra calibrer les modèles par la suite, les communes reconnues mais non éligibles pour les années 2016 et 2017 sont considérées comme des communes éligibles. En effet, nous estimons pouvoir faire cela et que cela suffit à rendre les données homogènes car si en regardant l'évolution des critères de reconnaissance, les critères de reconnaissance avant et après 2018 sont en partie similaires. Autrement dit, pour des petites fenêtres temporelles, ce qui est le cas ici, cette actualisation peut convenir et la base de données résultante n'est à priori pas totalement absurde.

5.3 Approche naïve et approche annuelle

Dans cette section, nous allons mettre à profit les nouvelles données de SWI Uniforme. Nous allons construire d'abord un modèle de référence le plus simple possible qui servira de point de comparaison en plus des estimations de la CCR ([43]). Puis nous mettrons à jour le modèle d'occurrence de l'approche occurrence-fréquence de la partie précédente. Enfin, nous présenterons la théorie des modèles inflatés et l'amélioration qu'elle peut apporter comparée au modèle d'occurrence-fréquence mis à jour.

5.3.1 Modèle de référence

En comparant le nombre de communes éligibles dans la figure 50 et la charge totale observée dans la figure 40, nous constatons que pour les trois années, à savoir 2016, 2017 et 2018, il semble y avoir une forte corrélation. Par conséquent, le modèle de référence proposé consiste à expliquer la charge totale par le nombre de communes éligibles. En d'autres termes, la charge totale est modélisée par :

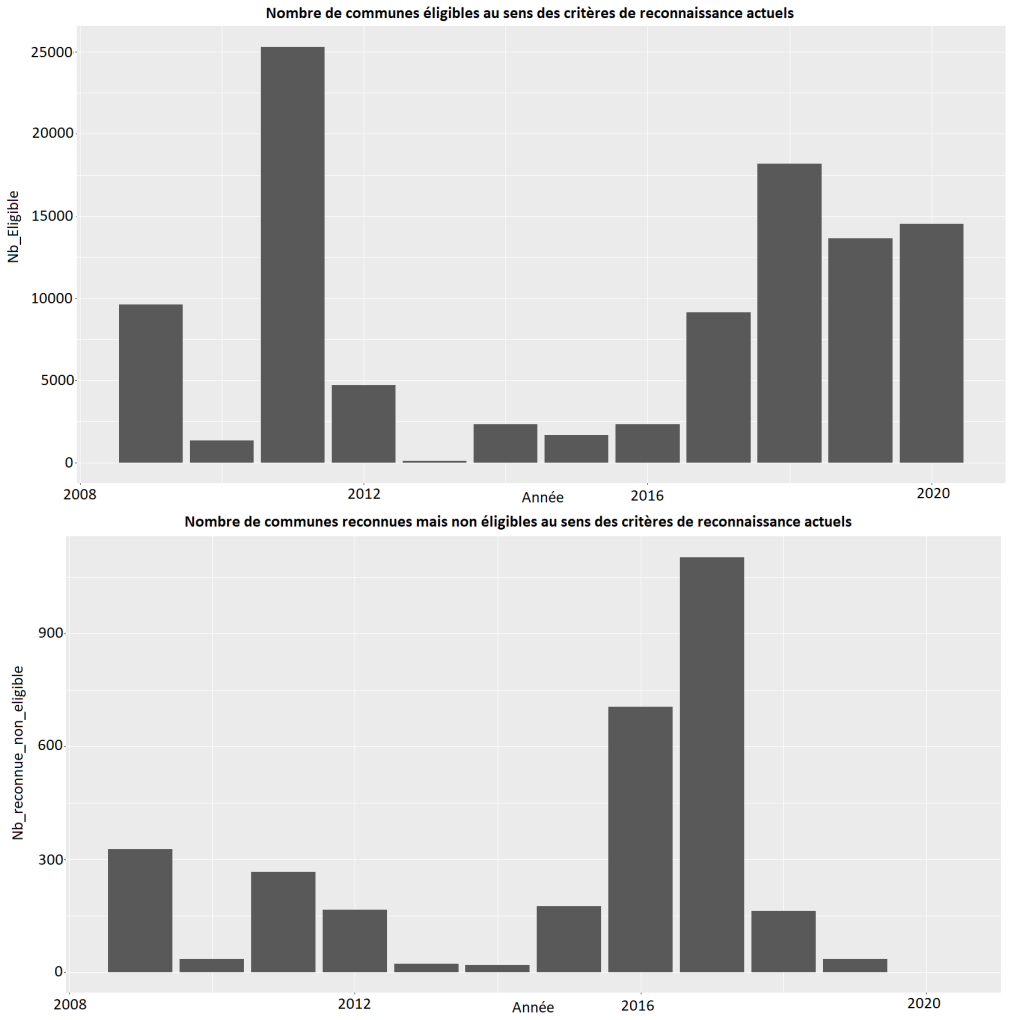

Figure 50 – De haut en bas, nous représentons le nombre de communes éligibles au sens des critères de reconnaissance actuels et le nombre de communes reconnues par la commission interministérielle mais non éligible au sens des critères actuels. Notez que pour les années 2011 et 2018, il y a un grand nombre de communes éligibles donc les critères actuels mettent en évidence les sécheresses des années 2011 et 2018, ce qui est à priori l'effet recherché. Le nombre de communes reconnues mais non éligibles au sens des critères de reconnaissance actuels est très faible pour l'année 2018 et diminue en 2019 mais est élevé pour certaines années antérieures à 2018. Ce qui est logique puisque les nouveaux critères ne sont entrés en vigueur qu'à partir de 2018. Néanmoins, cela rejoint la remarque de la section 5.2.1, à savoir qu'une sécheresse d'hiver d'une année passée peut ne pas être une sécheresse d'hiver si nous utilisons les critères actuels. Il faut aussi noter que pour 2011, malgré l'utilisation des anciens critères, il y a peu de communes reconnues mais non éligibles comparé au nombre de communes éligibles. En effet, il s'agit d'une année où il y a eu de nombreux recours et la commission a dû revoir les critères compte tenu de la situation particulière. Cela ne s'est pas opéré de la même manière en 2016 et 2017 car les sécheresses semblent être moins importantes comparées à 2011 et il y a les nouveaux critères de reconnaissance qui sont déjà en discussion.

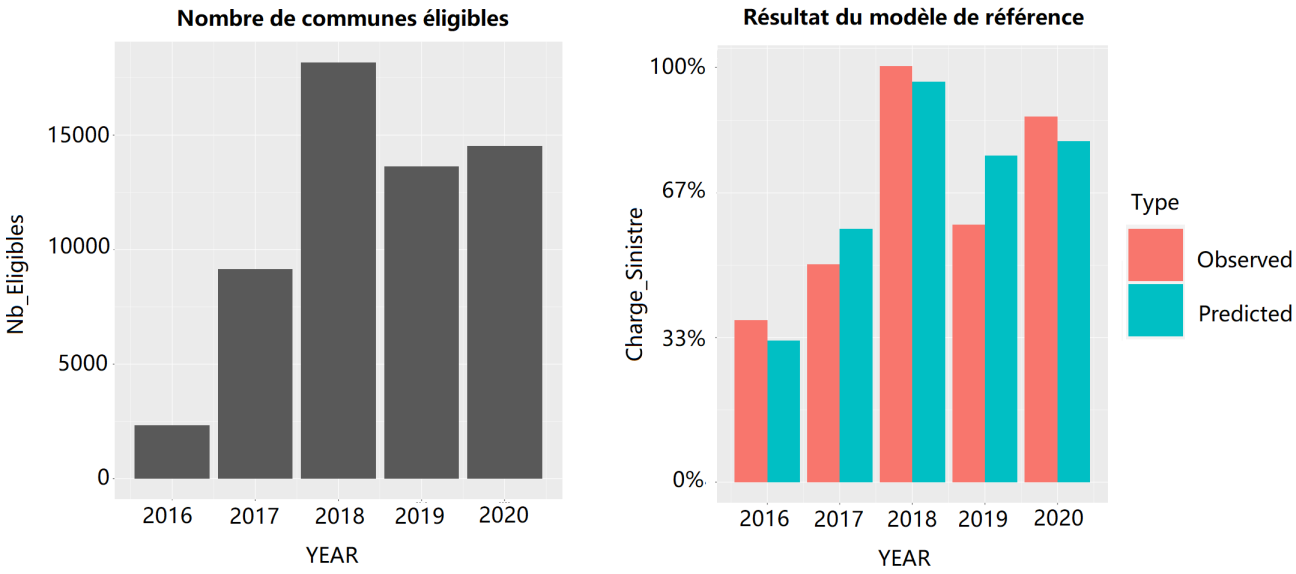

Figure 51 – Nous représentons le nombre de communes éligibles (gauche), la quantité de charge observée (droite, rouge) estimée par la CCR ([43]) et la quantité de charge prédite (droite, bleu) par le modèle de référence. Pour les années 2016, 2017, 2018 et 2020, il y a une bonne correspondance entre charge prédite et charge observée. Pour 2019, la charge prédite est surestimée. De manière générale, il faut noter que les prédictions sont plutôt correctes.

$$Charge\ totale\ =\ a \times nombre\ de\ communes\ eligible\ +\ b \tag{84}$$

Avec a le coefficient de régression et b la constante de régression. La figure 51 représente le nombre de communes éligibles, la quantité de charge observée et la quantité de charge prédite par le modèle de référence. Pour les années 2016, 2017, 2018 et 2020, il y a une bonne correspondance entre charge prédite et charge observée. Pour 2019, la charge prédite est surestimée. De manière générale, il faut noter que les prédictions sont plutôt correctes.

5.3.2 Modèle de commune demanderesse

Puisque nous avons vu que, en raison des évolutions réglementaires, il n'est plus nécessaire de construire un modèle d'occurrence pour prédire la probabilité qu'une commune soit décrétée en situation de catastrophe naturelle, nous allons par la suite tenter plusieurs approches différentes.

La première s'inscrit dans la même lignée que l'approche avec le modèle d'occurrence, i.e nous allons construire un modèle pour prédire la probabilité qu'une commune fasse la demande d'arrêté. En effet, si la commune est éligible, le décret sera automatiquement attribué, et si le décret n'est pas attribué, cela signifie que la commune ne répond pas aux critères d'attribution définis par le circulaire ([39]). Ainsi, l'idée derrière est que le choix de faire la demande d'arrêté ou pas est guidé par les conditions climatiques subites et donc possiblement expliqué par les indices de sécheresse. Ce modèle appelé modèle de probabilité demanderesse, va se substituer au modèle d'occurrence. Le modèle de fréquence, quant à lui, n'est pas modifié.

Le modèle de probabilité demanderesse est une forêt aléatoire et les variables utilisées sont les mêmes que celles utilisées dans le modèle de fréquence auxquelles nous ajoutons les variables de SWI Uniformes nouvellement construites. Le choix du forêt aléatoire est notamment motivé par l'analogie entre ce modèle et ce qui peut se passer lors de la formulation ou non d'une demande d'arrêté. En effet, parmi les quelque 35 000 communes françaises, la majorité sont des petites communes. Lors d'un événement de sécheresse, le maire constate souvent les dégâts directement auprès des habitants qui

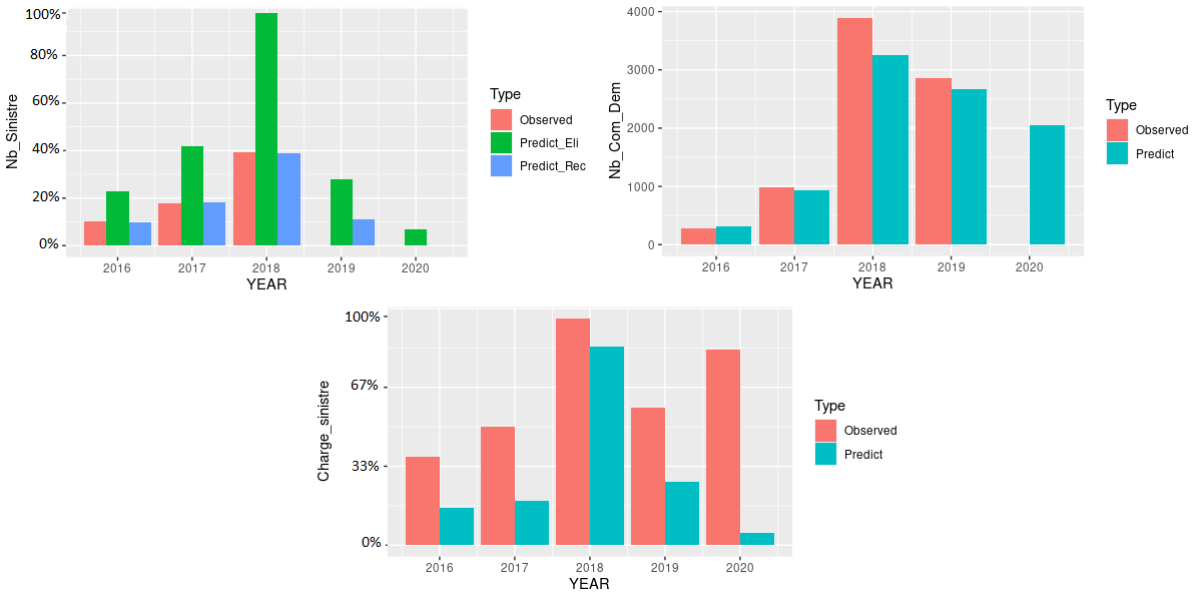

Figure 52 – De gauche à droite et de haut en bas, nous représentons le nombre total de sinistres (observé, prédit en sommant sur les communes éligibles et prédit en sommant sur les communes reconnues), le nombre de communes demanderesses (observé et prédit en sommant sur les communes éligibles) et la charge totale (observée et prédite en sommant sur les communes éligibles). Notez que les charges observées sont des estimations de la CCR ([43]) et que pour 2019 et 2020, nous ne disposons pas d'estimation du nombre de sinistres.

décident eux s'il faut formuler une demande d'arrêté. Cette consultation directe auprès des concernés pour formuler une décision finale est analogue au fonctionnement d'un modèle de forêt aléatoire qui agrège les résultats de chaque arbre de décision qui le constitue.

Pour calibrer ce modèle de probabilité demanderesse, les données des années 2009 à 2017 sont utilisées. Les données des années 2018 à 2020 constituent, quant à elles, l'ensemble test. Notez que les données d'apprentissage sont équilibrées, i.e les communes non demanderesses sont sous-échantillonnées de telle sorte qu'il y a autant de communes demanderesses que de communes non demanderesses. Ainsi, en notant y_{dem} la sortie du modèle de probabilité demanderesse, le modèle demanderesse-fréquence est :

$$A_{tot}(t) = \sum_{c \in \mathcal{E}} y_{dem}(c,t) \times y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \tag{85}$$

Avec $A_{tot}(t)$ la charge totale prédite de l'année t et $\mathcal{E} = \mathcal{E}_r \cup \mathcal{E}_{nr}$ l'ensemble des communes éligibles qui peut se décomposer en $\mathcal{E}_r$, communes éligibles et reconnues (car demande d'arrêté formulée), et $\mathcal{E}_{nr}$, communes éligibles mais non reconnues (car demande d'arrêté non formulée).

La figure 52 représente la nombre total de sinistres (observé, prédit en sommant sur les communes éligibles et prédit en sommant sur les communes reconnues), le nombre de communes demanderesses (observé et prédit en sommant sur les communes éligibles) et la charge totale (observée et prédite en sommant sur les communes éligibles). Notez que nous nous intéressons aux résultats sur les communes éligibles et non sur les communes reconnues puisque la reconnaissance n'est pas observable en temps réel. Nous mentionnons les prédictions sur les communes éligibles pour le modèle de fréquence uniquement dans un but d'analyse du modèle. Ainsi, il faut noter que le nombre de sinistres prédit, en sommant sur les communes éligibles, est largement surestimé. Ceci est logique car lors de la calibration du modèle, uniquement les données des communes reconnues sont utilisées en raison de l'approche en deux étapes, d'où une bonne prédiction sur les communes reconnues. Pour le modèle de probabilité demanderesse, nous avons plutôt une bonne prédiction. Une question légitime est alors pourquoi la surestimation du nombre de sinistres sur les communes éligibles ne se retrouve pas dans la charge totale

prédite. Pour comprendre cela, il faut reprendre l'équation 85 et supposer que le modèle de probabilité demanderesse est parfait, i.e y_{dem} prédit exactement 1 pour les communes éligibles demanderesses et donc reconnues et exactement 0 pour les communes éligibles mais non demanderesses. Dans ce cas, la charge totale prédite est vaut :

$$\begin{aligned} A_{tot}(t) &= \sum_{c\in\mathcal{E}} y_{dem}(c,t) \times y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \\ &= \sum_{c\in\mathcal{E}_r} y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \end{aligned} \tag{86}$$

Ce qui correspond effectivement à ce que le modèle complet souhaite avoir. Cependant si y_{dem} n'est pas parfait mais correspond plutôt à une distribution très centrée autour d'une valeur x tout en gardant une très bonne prédiction lorsque nous agrégeons au niveau annuel, ce qui est possible, alors en approximant y_{dem} par x, nous avons :

$$A_{tot}(t) \approx x \times \left(\sum_{c\in\mathcal{E}} y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year} \right) \tag{87}$$

Il suffirait alors d'avoir x proche de $x^\star$ pour que malgré la surestimation du nombre de sinistres, il est tout de même possible d'avoir une charge totale prédite correcte. Le $x^\star$ étant défini par :

$$x^\star \approx \frac{\sum\limits_{c\in\mathcal{E}_r} y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year}}{\sum\limits_{c\in\mathcal{E}} y_{freq}(c,t) \times n_{risk}(c,t) \times C_{year}} \tag{88}$$

Ainsi, le cas de la figure 52 avec une estimation 2018 plutôt correcte est tout à fait possible. Cependant, le modèle sous-estime pour les autres années, ce qui n'est pas désiré.

Ainsi, il semble que l'approche par le modèle demanderesse-fréquence est prometteuse mais difficile à mettre en place. Par ailleurs, il semble aussi que l'idée de modéliser la décision qu'un maire fasse une demande d'arrêté est difficilement acceptable par les cédantes d'un point de vue commercial. Pour répondre à ces problèmes soulevés, au lieu de continuer l'approche en deux parties, nous décidons de réaliser une modélisation en un seul tenant, i.e modéliser directement le nombre de sinistres en utilisant les données des communes éligibles. Cela fait perdre au modèle une part d'interprétabilité car tout est caché dans un unique modèle mais l'avantage est que le modèle devient ainsi plus simple et à priori plus acceptable car répond en partie aux critiques potentielles des cédantes sur le fait qu'il est difficile de modéliser une décision humaine. Notez que cette simplicité est aussi un atout dans le cas d'un éventuel produit d'assurrance-réassurance.

5.3.3 Modèle de zero inflaté annuel

Dans cette section, nous allons considérer non pas deux mais un seul et unique modèle qui prédit directement le nombre de sinistres pour l'ensemble des communes éligibles et pas seulement pour les communes reconnues. Notez que cette approche, en plus d'être moins complexe, i.e un modèle au lieu de deux, a l'avantage de correspondre davantage à la vision commune, à savoir que le choix de faire une demande d'arrêté de catastrophe naturelle est un choix politique et social et ne peut pas être modéliser aisément.

5.3.3.1 Tentative et limite du GLM poisson

Pour la modélisation, nous commençons d'abord par employer un modèle GLM avec une distribution de poisson, un lien *log* et un terme d'exposition qui est le $\log(n_{risk})$. En effet, les données d'apprentissage sont les données des communes éligibles et pour ces communes, les valeurs non nulles sont plutôt rares. Ainsi, une loi d'événement rare comme la loi de poisson est à priori plutôt adaptée.

Ensuite les méthodes de sélection des variables présentées et utilisées lors de la construction du modèle de fréquence dans la section 4.5 sont appliquées pour construire le modèle. Néanmoins, deux étapes de sélection au préalable supplémentaire sont ajoutées. En effet, dans les résultats précédents, de manière générale, il y a une surreprésentation des variables explicatives calculées sur une fenêtre glissante de 3 mois, 6 mois ou 12 mois parmi les variables sélectionnées. Les variables explicatives calculées sur une fenêtre de 1 mois et 24 mois ne sont donc pas pertinentes à priori. Ce qui semble logique puisque les premières sont probablement très bruitées en raison de la petite fenêtre temporelle et les secondes sont calculées sur une fenêtre trop longue et capture davantage la sécheresse hydrologique plutôt que la sécheresse météorologique ou agricole. Ainsi les variables climatiques avec une fenêtre de calcul de 1 mois et 24 mois sont négligées dès le début. Par ailleurs, pour avoir un modèle qui reste cohérent et facilement interprétable, dans un premier temps, les variables climatiques saisonnières sont négligées. En termes de variables climatiques, seules les variables annuelles sont gardées. En effet, cela limite l'effet de compensation qui peut y avoir entre les saisons. La compensation entre nombre de sinistres total positif pour une saison et négatif pour une autre fait à priori fonctionner le modèle mais n'a pas réellement de sens physique.

Ainsi, en sélectionnant au préalable les variables selon les deux précédents critères, en appliquant les mêmes méthodes de sélection des variables explicatives que pour le modèle de fréquence, à savoir dans l'ordre, supprimer les variables dont la variance est quasi nulle, supprimer les variables colinéaires, appliquer l'algorithme boruta, et appliquer l'algorithme de stepwise, nous obtenons un modèle GLM poisson avec 116 variables, un lien *log* et $\log(n_{risk})$ comme terme d'exposition. À titre indicatif, le calcul du modèle est réalisé à l'aide de la fonction $glm(.)$ de **R**.

Avant d'analyser les résultats de ce modèle, notez que parmi les données, les communes ayant un nombre de sinistres non nul, i.e $n_s(c,t) \neq 0$, sont plutôt rares. En d'autres termes, la question qui se pose est est-ce que le modèle GLM a bien pris en compte la sur-représentation des zéros dans les observations. Pour répondre à cela,il faut rappeler que pour une loi de poisson de paramètre λ, la moyenne est aussi λ. Autrement dit, une estimation du nombre de zéros prédit par le modèle, $\hat{n}_{zero}$, est donnée par :

$$\hat{n}_{zero} = \sum_{t=2016}^{2018} \sum_{c \in \mathcal{E}} n_{risk} \times \exp(\mu(c,t)) \tag{89}$$

Où $\mu(c,t)$ est la prédiction du nombre de sinistres par le modèle GLM poisson pour la commune c et l'année t. Notez que $\exp(-\mu(c,t))$ est la probabilité de tirer zéro pour une variable aléatoire suivant une loi de poisson de paramètre $\lambda = \mu(c,t)$. Ainsi, pour n_{zero} le nombre de zéros réellement observé, nous constatons que :

$$\frac{n_{zero} - \hat{n}_{zero}}{n_{zero}} \approx 11\% \tag{90}$$

Autrement dit, le nombre de zéros est sous-estimé à hauteur de 11% comparé au nombre de zéros réellement présent dans la base de données, ce qui est grand par rapport au seuil de quelques pourcents communément utilisé dans la littérature. D'un point de vue théorique, cette sous-estimation semble logique car les zéros sont à priori générés par deux phénomènes différents à savoir des communes ayant obtenue l'arrêté mais sans sinistralité réelle et des communes n'ayant tout simplement pas fait de demande d'arrêté. Ainsi, pour tenir compte de cela, il est plus adéquat d'utiliser un modèle de poisson inflaté en zéro plutôt qu'un modèle de poisson simple.

5.3.3.2 Définition du modèle zero inflaté

Soit Y la variable réponse, alors cette dernière suit une loi de poisson zéro inflaté si :

$$\begin{cases} \mathbb{P}(Y=0) = \pi + (1-\pi)\exp(-\lambda) \\ \mathbb{P}(Y=k) = (1-\pi)\dfrac{\lambda^k}{k!}\exp(-\lambda), \quad k = 1,2,3,... \end{cases} \tag{91}$$

Avec :

$$\begin{cases} \mathbb{E}(Y) = (1-\pi)\lambda \\ \text{Var}(Y) = (1-\pi)(\lambda + \pi\lambda^2) \end{cases} \tag{92}$$

Où λ et π sont des paramètres à estimer. Notez que ce modèle est différent du modèle zéro modifié (ou modèle Hurdle) car dans ce dernier, les zéros sont générés seulement par une masse en zéro alors que dans le modèle zéro inflaté, les zéros sont générés par la masse en zéro π et la distribution de poisson en zéro $\exp(-\lambda)$.

En pratique, la masse en zéro, i.e π, est modélisée le plus souvent par un GLM binomial avec lien *logit* et les coefficients de régression sont obtenus par maximum de vraisemblance. En notant θ_{zero} les coefficients de la partie binomiale et X les variables explicatives associées, θ_{count} les coefficients de la partie poisson et Z les variables explicatives associées, alors pour $Y \sim \mathcal{ZI}(N, \pi, \lambda)$ (loi zéro modifié de paramètre N, π et λ) avec lien *logit* pour la partie binomiale, lien *log* pour la partie poisson et n_{risk} comme exposition, la prédiction μ s'écrit :

$$\begin{cases} \pi = \dfrac{\exp(\theta_{zero}X)}{1+\exp(\theta_{zero}X)} \\ \mu = \mathbb{E}(Y) = (1-\pi) \times n_{risk} \times \exp(\theta_{count}Z) \end{cases} \tag{93}$$

Il faut noter que les variables explicatives utilisées dans la partie binomiale du modèle peuvent parfaitement être différentes des variables explicatives utilisées dans la partie poisson.

Ainsi, le modèle de poisson zéro inflaté est privilégié au lieu du modèle de poisson simple. La construction du modèle est réalisée à l'aide de la fonction $zeroinfl()$ de la librairie *pscl* de **R**. Pour cette construction, en termes de données, les données des communes éligibles des années 2016, 2017 et 2018 sont utilisées comme pour le modèle poisson simple. Pour les variables explicatives de la partie poisson du modèle zéro inflaté, nous utilisons les variables annuelles avec une fenêtre temporelle de 3 mois, 6 mois ou 12 mois, les variables géologiques, et les variables socio-économiques. Nous appliquons ensuite sur ces dernières les méthodes de sélection des étapes 1, 2 et 3 de la section 4.5. Pour les variables explicatives de la partie binomiale du modèle zéro inflaté, nous utilisons les mêmes variables mais sur lesquelles nous appliquons les méthodes de sélection des étapes 1, 2, 3 et 4 de la section 4.5. Notez que l'étape 4 s'applique en spécifiant un modèle initial qui est ici un GLM binomial lien *logit* avec comme réponse $I(n_s > 0)$ où I est la fonction indicatrice. Il n'est pas possible d'appliquer l'étape 4 pour les variables de la partie poisson du modèle zéro inflaté car l'algorithme stepwise n'est pas défini.

En faisant cette construction, le modèle zero inflaté annuel obtenu donne pour chaque commune c et chaque année t, la proportion de contrats sinistrés y_{an} définie par :

$$y_{an}(c,t) = (1-\pi(c,t)) \times n_{risk}(c,t) \times \exp(\theta_{count}Z(c,t)) \tag{94}$$

Avec les différentes quantités définies par l'équation 93 et les variables correctement évaluées en l'observation (c,t). Ce modèle donne notamment :

$$\frac{n_{zero} - \hat{n}_{zero}}{n_{zero}} \approx 2\% \tag{95}$$

Ainsi, le nombre de zéros est correctement estimé puisque l'écart n'est plus que de l'ordre de 2% comparé au nombre de zéros observé, ce qui est parfaitement acceptable.

5.3.3.3 Analyse des résultats

Avant de montrer les résultats, il faut définir un coefficient de charge moyenne par département, notée C_{dep}. Ce dernier, pour chaque département d, est défini par :

$$C_{dep}(d) = \frac{1}{10}\sum_{t=2009}^{2018} \frac{\sum\limits_{c\in d} A(c,t)}{\sum\limits_{c\in d} n_{risk}(c,t)} \tag{96}$$

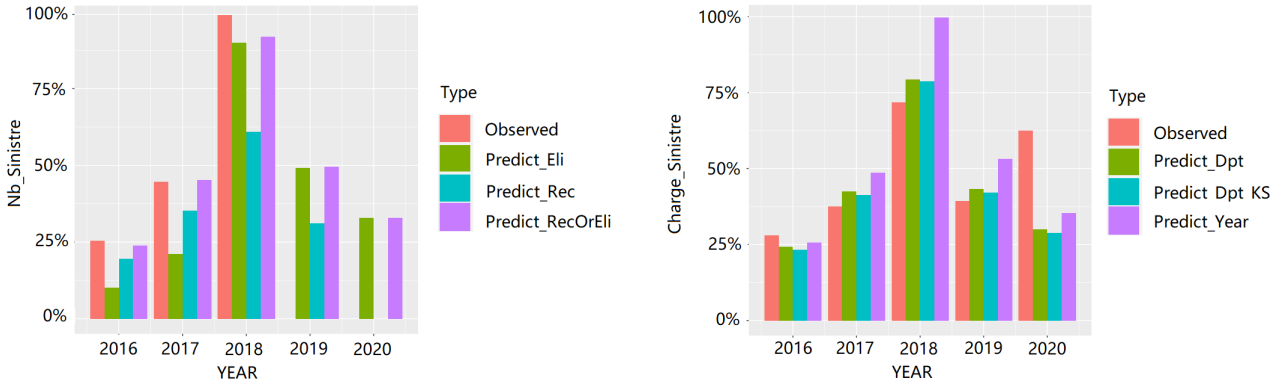

Figure 53 – Nous représentons le nombre de sinistres (observé en rouge, prédit en sommant sur les communes éligibles en vert, prédit en sommant sur les communes reconnues en bleu et prédit en sommant sur les communes éligibles ou reconnues en violet) à gauche et la charge totale (observée en rouge, prédite en sommant sur les communes éligibles et en appliquant soit C_{year} en violet, soit C_{dep} en vert ou soit C^L_{dep} en bleu) à droite pour le modèle zero inflaté annuel. Notez que pour le nombre de sinistres, nous nous intéressons à la prédiction sommée sur les communes éligibles ou reconnues. Les deux autres prédictions montrent la pertinence de l'approximation évoquée à la fin de la section 5.2.2, i.e pour 2016 et 2017, il est important de considérer les communes reconnues mais non éligibles comme des communes éligibles. Notez que de manière générale, le modèle de poisson inflaté en zéro avec un terme d'exposition $\log(n_{risk})$ donne plutôt une bonne prédiction des charges totales pour les années 2016, 2017, 2018 et 2019. L'utilisation du coefficient C^L_{dep} est probablement à privilégier comparée à C_{year} ou C_{dep}. Néanmoins, la charge 2020 prédite est surestimée, ce qui n'est pas désiré. Notez que les charges observées sont des estimations de la CCR ([43]) et que pour 2019 et 2020, nous ne disposons pas d'estimation du nombre de sinistres.

Dans la mesure où il existe certains départements qui ont très peu de sinistres voire pas de sinistre, il faut lisser géographiquement C_{dep} par exemple à l'aide d'un noyau gaussien dans le cas présent. Soit C^L_{dep} la version lissée de C_{dep}, alors pour le département d, ce dernier est défini par :

$$C^L_{dep}(d) = \frac{\sum\limits_{k \neq d} C_{dep}(d) \times \exp\left(-\dfrac{h^2(k,d)}{2b^2}\right)}{\sum\limits_{k \neq d} \exp\left(-\dfrac{h^2(k,d)}{2b^2}\right)} \tag{97}$$

Avec $h(k,d)$ la distance géographique entre le département k et le département d et b la distance moyenne entre deux départements voisins. Ce lissage permet notamment d'avoir une approximation de la charge moyenne pour les départements jamais sinistrés dans le portefeuille.

Ainsi, dans la figure 53, nous représentons le nombre de sinistres et la charge totale pour par le modèle zéro inflaté annuel. Notez que pour le nombre de sinistres, nous nous intéressons à la prédiction sommée sur les communes éligibles ou reconnues. Les deux autres prédictions ne sont là que pour confirmer la pertinence de l'approximation évoquée à la fin de la partie 5.2.2. En effet, pour 2016 et 2017, il est important de considérer les communes reconnues mais non éligibles comme des communes éligibles, sans quoi les prédictions risquent d'être grandement biaisées pour ces deux années.

Notez que de manière générale, le modèle de poisson inflaté en zéro avec un terme d'exposition $\log(n_{risk})$ donne plutôt une bonne prédiction des charges totales pour les années 2016, 2017, 2018 et 2019. L'utilisation du coefficient C^L_{dep} est probablement à privilégier comparée à C_{year} ou C_{dep}.Néanmoins, la charge 2020 prédite est surestimée, ce qui n'est pas désiré.

Par ailleurs, pour avoir une idée de la pertinence des prédictions 2019 et 2020, nous faisons une ACP et nous allons analyser la position des données des années 2019 et 2020 par rapport aux données des

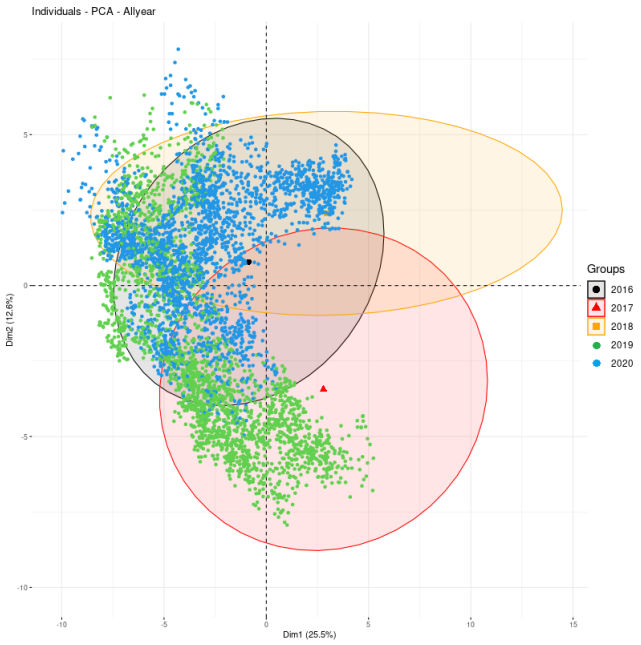

Figure 54 – Nous représentons l'emplacement des données des années 2016 (noir), 2017 (rouge) et 2018 (orange) par des ellipses de niveau 95% (i.e l'ellipse contient 95% des données) et les données des années 2019 (vert) et 2020 (bleu) par des points dans la plan formé par les deux composantes principales les plus importantes. Bien entendu, pour raison de visibilité, seule une partie des données sont représentées. De manière générale, il faut noter que les points sont majoritairement situés dans les ellipses, ce qui signifie que les données des années 2019 et 2020 sont probablement situées parmi les données des années précédentes. Autrement dit, ces nouvelles données sont plus ou moins des données déjà vues par le modèle et par conséquent, il est possible d'espérer que les prédictions du modèle ne soient pas totalement aberrantes.

années 2016, 2017 et 2018 dans les deux composantes principales les plus importantes. Une définition détaillée de la méthode d'ACP se trouve dans la partie 3.4.1.

Dans la figure 54, nous représentons l'emplacement des données des années 2016, 2017 et 2018 par des ellipses de niveau 95% (i.e l'ellipse contient 95% des données) et les données des années 2019 et 2020 par des points dans la plan formé par les deux composantes principales les plus importantes. Bien entendu, pour raison de visibilité, seule une partie des données sont représentées.

De manière générale, il faut noter que les points sont majoritairement situés dans les ellipses, ce qui signifie que les données des années 2019 et 2020 sont probablement situées parmi les données des années précédentes. Autrement dit, ces nouvelles données sont plus ou moins des données déjà vues par le modèle et par conséquent, il est possible d'espérer que les prédictions du modèle ne soient pas totalement aberrantes.

Ainsi, la conclusion de ce modèle zéro inflaté annuel est que la prédiction 2019 est probablement acceptable mais la prédiction 2020 est potentiellement trop basse. Cette sous estimation s'explique en partie par l'éloignement temporel mais probablement aussi par le fait qu'il n'y a pas suffisamment d'années de données. Néanmoins, il est tout de même possible d'envisager une amélioration de ce modèle. L'idée est qu'il faut revenir sur les critères de reconnaissance actuels (5.2.1) et essayer de modéliser chacun des critères séparément. En effet, il faut noter qu'il y a une certaine indépendance dans les critères climatiques, par exemple le critère estival et hivernal sont à priori indépendants les données SWI Uniforme utilisées ne sont pas les mêmes, mais que cette indépendance n'est pas suffisamment prise en compte dans le modèle annuel.

5.4 Segmentation des données et approche saisonnière

La dernière approche considérée est une approche saisonnière et à échelle communale. Il faut noter d'abord que l'ensemble des modèles construits jusqu'alors sont issus d'une approche annuelle et non saisonnière. Par ailleurs, dans le papier de Charpentier et al. ([42]), les auteurs évoquent déjà un éventuel intérêt de cette approche dans leur conclusion mais faute de données, ils se sont arrêtés au modèle zero inflaté annuel.

L'idée de la démarche est de construire quatre modèles, i.e un par saison, et d'agréger ensuite les résultats. Chaque modèle tente donc de prédire le nombre de sinistres survenus lors de la saison associée. Cela a notamment l'avantage d'éviter l'effet de compensation non désiré remarqué précédemment dans la construction du modèle zero inflaté annuel. Plus exactement, cette modélisation saisonnière prend en compte explicitement le caractère indépendant des critères climatiques. Ainsi, dans cette approche, il est possible de dire par exemple quelle est la saison la plus sinistrée pour chaque année.

Pour ce faire, nous présenterons d'abord la division de la base complète construite dans la partie précédente en quatre sous-ensembles. Nous replacerons ensuite le modèle de référence de la section précédente dans cette approche saisonnière afin de garder un point de comparaison. Puis nous définirons le modèle saisonnier composé de quatre modèles sous-jacents qui sont tous des modèles inflatés en zéro. Et in fine, en comparant les résultats du modèle saisonnier au modèle de référence et au modèle zero inflaté annuel, nous concluons que le modèle saisonnier est probablement le plus prometteur parmi tous les modèles que nous avons construit jusqu'à présent, malgré l'incertitude de ses prédictions pour 2020 en raison d'un manque de diversité dans les années et un manque de données pour certaines saisons.

5.4.1 Division de la base de données

Pour réaliser cette approche, il faut diviser la base de données de catastrophe naturelle mise à jour avec les données de SWI Uniforme en quatre bases saisonnières. Pour cela, nous considérons les mois de janvier, février et mars comme saison d'hiver, les mois de avril, mai et juin comme saison de printemps, les mois de juillet, août et septembre comme saison d'été et les mois de octobre, novembre et décembre comme saison d'automne.

Pour les données d'arrêté, nous disposons pour chaque arrêté la période de validité de ce dernier. Cette information est utilisée pour savoir si une commune donnée a obtenu ou non la reconnaissance d'état de catastrophe naturelle pour une saison donnée d'une année donnée, ce qui ramène les données d'arrêté à l'échelle saisonnière. Dans la figure 55, le nombre d'arrêtés par années et par saison est représenté. Notez qu'avant les années 2000, les arrêtés sont donnés pour des périodes très longues, d'où une espèce de continuité dans les barres sur la partie gauche du plot. En 2003, il y a eu une sécheresse estivale importante, ce qui explique l'explosion du nombre d'arrêtés en été 2003. De même, la figure reflète bien la sécheresse printanière en 2011 qui a impliqué l'introduction d'un critère printanier par les autorités. De manière générale, le passage de l'échelle annuelle à l'échelle saisonnière est donc bien effectué pour les données d'arrêtés.

Pour les données de sinistres, nous disposons des dates approximatives des sinistres. Cette information est donc utilisée pour attacher chaque événement de sinistre et chaque charge d'événement à l'arrêté qui lui est le plus proche (au sens temporel). Dans la figure 56, le nombre de sinistres et de la charge totale par année et par saison est représenté. Notez que les deux plots de la figure sont très semblables, ce qui est à priori logique, mais surtout il faut remarquer que l'été 2003 et le printemps 2011 ressort bien. Ceci montre que le passage de l'échelle annuelle à l'échelle saisonnière est plutôt bien effectué aussi.

Et enfin pour les variables explicatives, par définitions, les variables climatiques sont soit des variables annuelles, soit des variables saisonnières. Il suffit donc d'enlever les variables annuelles et faire correspondre les variables saisonnières à leurs saisons respectives. Les variables géologiques et socio-économiques étant indépendantes des saisons, nous les gardons dans chacune des bases saisonnières, de même que pour le nombre de contrats n_{risk}.

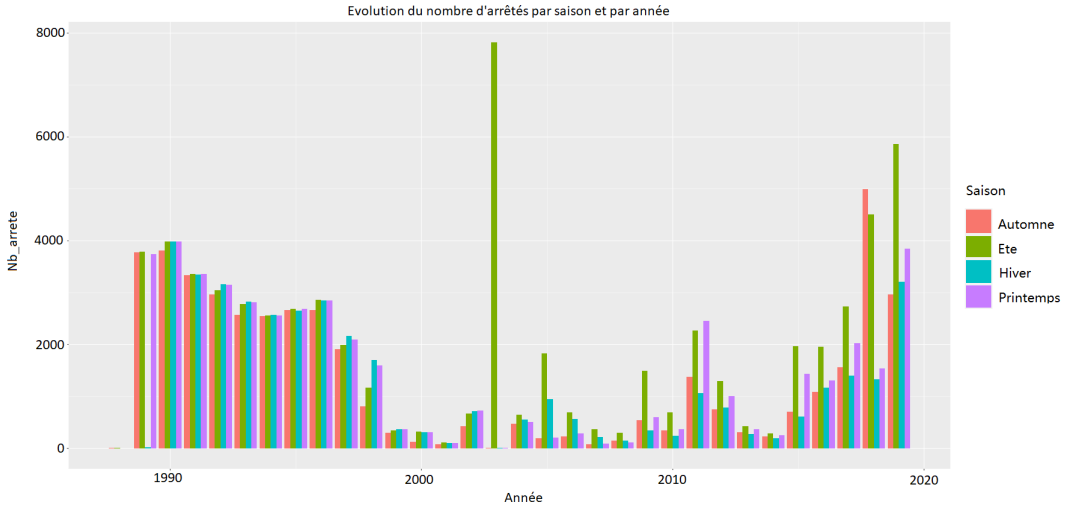

Figure 55 – Nous représentons le nombre d'arrêtés par années (abscisse) et par saison (rouge pour l'automne, vert pour l'été, bleu pour l'hiver et violet pour le printemps). Certains chiffres sont en pourcentage du maximum observé pour raison de confidentialité. Noter qu'avant les années 2000, les arrêtés sont donnés pour des périodes très longues, d'où une espèce de continuité dans les barres sur la partie gauche du plot. En 2003, il y a eu une sécheresse estivale importante, d'où le nombre d'arrêtés en été 2003 qui explose. De même, la figure reflète bien la sécheresse printanière en 2011 qui a impliqué l'introduction d'un critère printanier par les autorités. De manière générale, le passage de l'échelle annuelle à l'échelle saisonnière est donc bien effectué pour les données d'arrêtés.

5.4.2 Revisite du modèle de référence

Avant de présenter le modèle saisonnier, il faut faire un rapide retour sur le modèle de référence. En effet, à l'aide des bases saisonnières, il est possible d'avoir le nombre de communes éligibles par saison. Dans la figure 57, ce nombre est représenté ainsi que le nombre de communes éligibles par saison en fonction de la charge totale par saison pour les années 2016, 2017 et 2018.

Dans cette figure, notez que dans le plot d'en bas, malgré le peu de points disponibles, il semble avoir une relation linéaire entre le nombre de communes éligibles et le nombre de sinistres, individuellement pour chaque saison (sauf hiver) mais aussi pour les quatre saisons confondues. Ceci montre que le modèle de référence, malgré sa simplicité, semble avoir tout de même une certaine pertinence.

5.4.3 Définition des modèles saisonniers et analyse des résultats

Pour les quatre modèles sous-jacents du modèle saisonnier, le modèle de poisson inflaté en zéro est à nouveau sollicité. Notez que chaque modèle sous-jacent donne la proportion de contrats sinistrés durant la saison correspondante. En multipliant la sortie par le nombre de contrats, chaque modèle prédit en fait le nombre de sinistres par commune. Ainsi, les proportions de contrats sinistrés pour les quatre modèles sous-jacents, notées y_{hiv}, y_{pri}, y_{ete} et y_{aut}, sont définies comme dans l'équation 94.

Ainsi, pour chaque saison, la même construction que pour le modèle zero inflaté annuel est faite, i.e nous utilisons les données des communes éligibles, au sens des variables $HIVER$, $PRINTEMPS$, ETE ou $AUTOMNE$ (définies dans la section 5.2.2), des années 2016, 2017 et 2018. Pour chaque modèle zéro inflaté sous-jacent, pour les variables explicatives de la partie poisson du modèle, nous utilisons les variables saisonnières de la saison correspondante avec une fenêtre temporelle de 3 mois, 6 mois ou 12 mois, les variables géologiques et les variables socio-économiques. Nous appliquons ensuite sur ces dernières les méthodes de sélection des étapes 1, 2 et 3 de la section 4.5. Pour les variables explicatives de la partie binomiale du modèle zéro inflaté, nous utilisons les mêmes variables mais sur lesquelles nous appliquons les méthodes de sélection des étapes 1, 2, 3 et 4 de la section 4.5. Notez que l'étape 4 s'applique en spécifiant un modèle initial qui est un GLM binomial lien *logit* avec comme

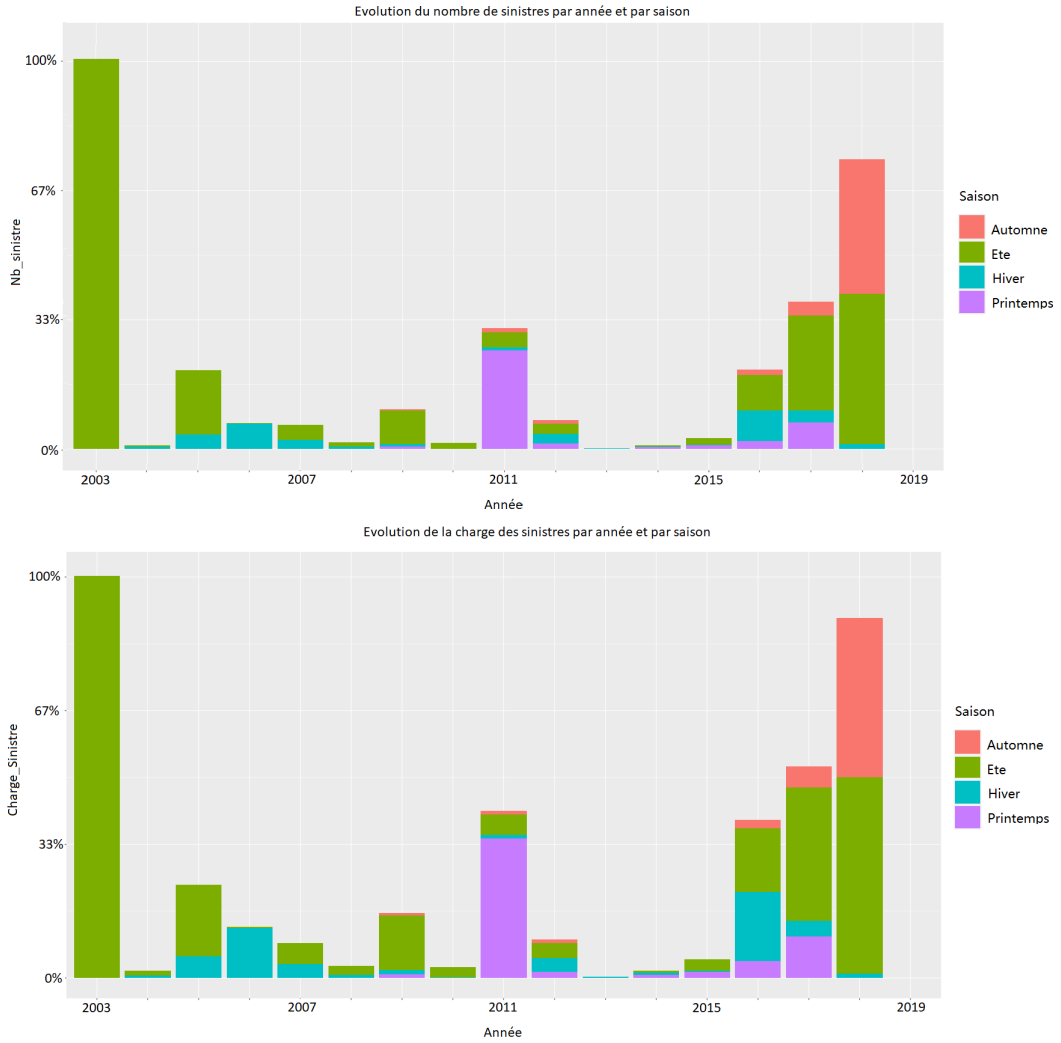

Figure 56 – Nous représentons l'évolution du nombre de sinistres (en haut) et de la charge totale (en bas) par année et par saison (rouge pour l'automne, vert pour l'été, bleu pour l'hiver et violet pour le printemps). Nous constatons d'abord que les deux plots sont très semblables, ce qui est logique, mais surtout nous notons que l'été 2003 et le printemps 2011 ressort bien. Ceci montre que le passage de l'échelle annuelle à l'échelle saisonnière est plutôt bien effectué.

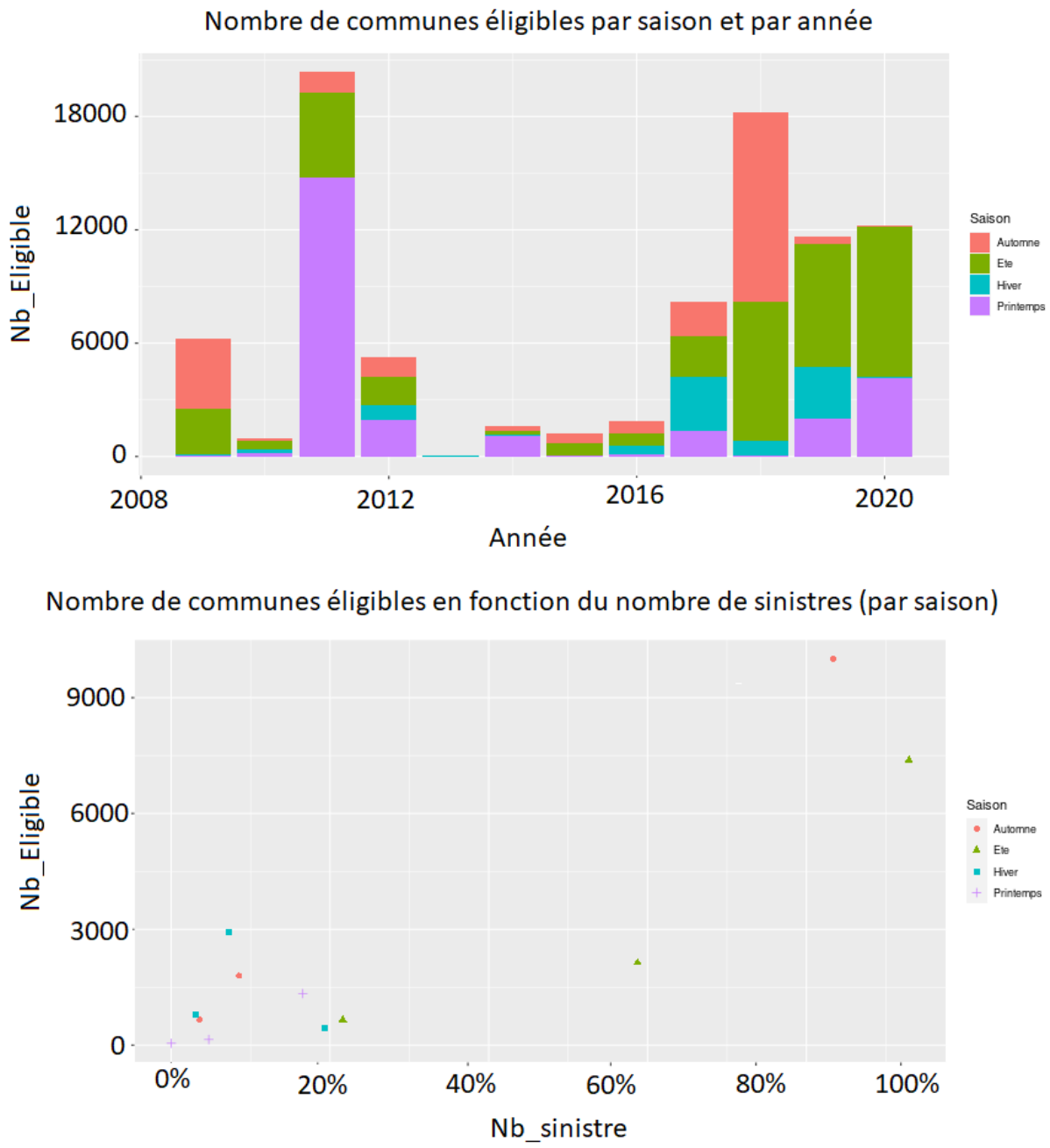

Figure 57 – Nous représentons le nombre de communes éligibles par saison (en haut) ainsi que le nombre de communes éligibles par saison (rouge pour l'automne, vert pour l'été, bleu pour l'hiver et violet pour le printemps) en fonction de la charge totale par saison pour les années 2016, 2017 et 2018 (en bas). Notez que dans le plot d'en bas, malgré le peu de points disponibles, il semble avoir une relation linéaire entre le nombre de communes éligibles et le nombre de sinistres, individuellement pour chaque saison (sauf hiver) mais aussi pour les quatre saisons confondues. Ceci montre que le modèle de référence, malgré sa simplicité, semble avoir tout de même une certaine pertinence.

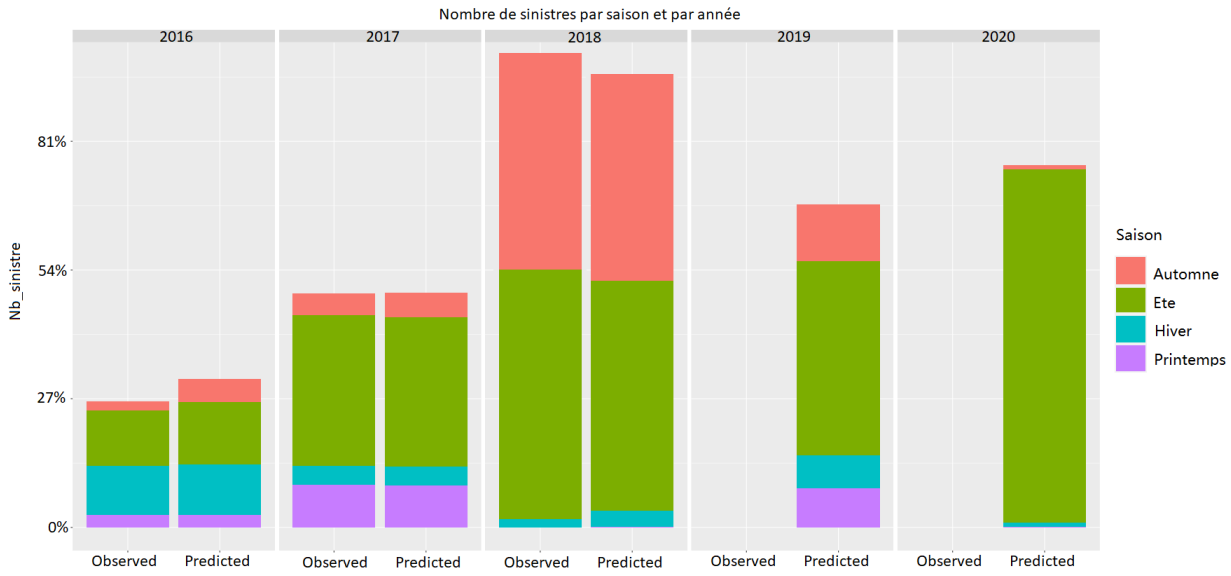

Figure 58 – Plot représentant le nombre de sinistres observé et prédit pour 2016-2018 et le nombre de sinistres prédit pour 2019-2020 par le modèle saisonnier combinant les quatre modèles sous-jacents (rouge pour l'automne, vert pour l'été, bleu pour l'hiver et violet pour le printemps). Notez que les prédictions sont globalement correctes pour les années 2016, 2017 et 2018 qu'importe la saison (avec éventuellement l'été 2018 qui est un peu sous-estimé). Notez aussi, et surtout, que le modèle prédit pour les années 2019 et 2020 une domination des sinistres en été. Il n'y a pas de valeur observée en 2019 et 2020 car nous avons uniquement des estimations en euros à l'échelle annuelle pour ces deux années.

réponse $I(n_s > 0)$ où I est la fonction indicatrice.

In fine, en combinant la sortie des quatre modèles, la proportion de contrats sinistrés totale, notée y_{saison}, est définie comme la somme des proportions de contrats sinistrés de chaque saison. Ainsi :

$$y_{saison} = y_{hiv} + y_{pri} + y_{ete} + y_{aut} \tag{98}$$

Notez que pour passer de la proportion au nombre de contrats sinistrés, il suffit de multiplier par n_{risk} qui est annuel et ne dépend pas des saisons (i.e même n_{risk} dans les quatre bases saisonnières).

La figure 58 représente le nombre de sinistres prédit par le modèle saisonnier combinant les quatre modèles sous-jacents. Notez que les prédictions sont globalement plutôt correctes pour les années 2016, 2017 et 2018 qu'importe la saison (avec éventuellement l'été 2018 qui est un peu sous-estimé). Notez aussi, et surtout, que le modèle prédit pour les années 2019 et 2020 une domination des sinistres en été. Il n'y a pas de valeur observée en 2019 et 2020 car nous avons uniquement des estimations en euros à l'échelle annuelle pour ces deux années (CCR, [43]).

En regardant à présent les résultats du modèle saisonnier dans la figure 59, qui représente le nombre de sinistres pour différents conditionnements et la charge totale pour différents coefficients de charge moyenne, notez que de manière générale, les prédictions des charges totales sont plutôt correctes avec l'utilisation du coefficient C^L_{dep} qui à privilégier comparée à C_{year} ou C_{dep}.

En comparant ces résultats au modèle de zero inflaté annuel de la figure 53, notez qu'avec le modèle saisonnier, la prédiction de charge totale 2019 est un peu surestimée mais la prédiction 2020 est tout à fait correcte. En analysant à présent, dans la figure 60, les positions des points comme dans la figure 54, notez que de manière générale, les données des années 2019 et 2020 font plutôt partie des données déjà vu par le modèle et par conséquent, il est possible d'espérer que les prédictions du modèle saisonnier ne sont pas totalement aberrantes.

Ainsi, l'ensemble de ces observations nous invite à penser que le modèle saisonnier, basé sur quatre modèles sous-jacents, est à retenir. En effet, les prédictions du modèle saisonnier sont cohérentes avec

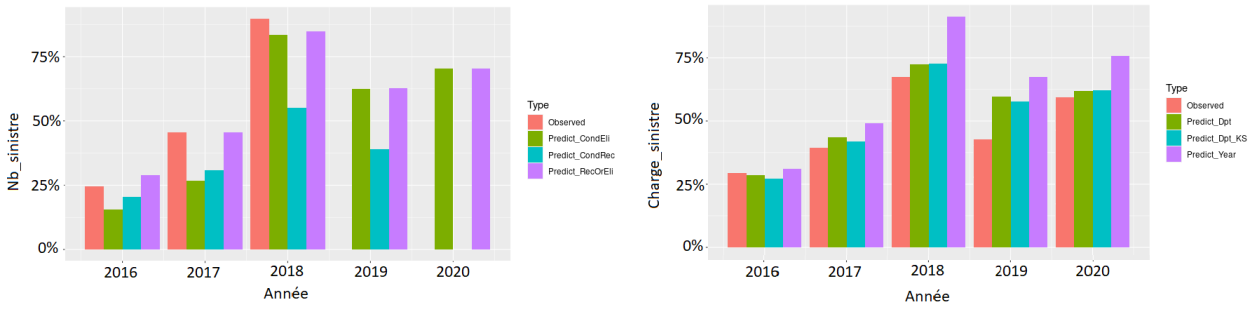

Figure 59 – Nous représentons le nombre de sinistres (observé en rouge, prédit en sommant sur les communes éligibles en vert, prédit en sommant sur les communes reconnues en bleu et prédit en sommant sur les communes éligibles ou reconnues en violet) à gauche et la charge totale (observée en rouge, prédite en sommant sur les communes éligibles et en appliquant soit C_{year} en violet, soit C_{dep} en vert ou soit C^L_{dep} en bleu) à droite pour par le modèle saisonnier. Notez sans surprise que de manière générale, les prédictions des charges totales sont plutôt correctes sauf pour 2019 où il y a un peu de surestimation. Par ailleurs, l'utilisation du coefficient C^L_{dep} est à privilégier comme remarqué précédemment. Puis, comparé au modèle de zero inflaté annuel de la figure 53, notez que la prédiction 2020 est correcte avec le modèle saisonnier.

les estimations de la CCR ([43]), i.e à l'échelle nationale, 2020 est une année plus sinistrée comparée à 2019 en termes de sinistralité subsidence. De plus, le modèle saisonnier montre que la charge des sinistres en 2020 s'explique surtout par une sinistralité estivale plus grave.

Néanmoins, il est nécessaire de nuancer la lecture des résultats. En effet, dans la figure 57, en 2020, le nombre de communes éligibles pour la période de printemps est important mais ceci n'est pas reflété dans la figure 58. Le modèle saisonnier prédit quasiment aucun sinistre pour le printemps 2020 alors qu'un tiers des communes éligibles en 2020 sont des communes éligibles pour la saison de printemps. De plus, en regardant de plus près les données utilisées par le modèle de printemps dans la figure 58, notez qu'il s'agit surtout des données qui proviennent de l'année 2017. Or si nous regardons les points de l'année 2019 et 2020 par rapport à l'ellipse de l'année 2017 pour la période de printemps dans la figure 60, nous constatons qu'une bonne partie des points 2020 sont en dehors de l'ellipse. Cela montre donc qu'il faut considérer les prédictions du modèle de printemps pour l'année 2020 avec précaution.

Par ailleurs, nous pouvons faire ces mêmes constats avec les données d'automne 2019 et l'ellipse de l'année 2018, ce qui justifie en partie la surestimation du nombre de sinistres 2019 et donc de la charge 2019 observée précédemment. Pour le modèle d'hiver, il faut noter que les données 2019-2020 sont surtout situées dans l'ellipse 2017, qui est l'année où il y a le plus de commune éligible en hiver (figure 58), et par conséquent ses prédictions 2019-2020 posent moins de problèmes aussi. Pour le modèle d'été, le problème ne se pose pas à priori car les données sont plutôt conséquentes qu'importe les années. Ainsi, tout ceci montre qu'il y a probablement une incertitude dans les prédictions liée au manque de donnée.

In fine, en considérant l'ensemble des remarques et le nombre de communes éligibles par saison de la figure 57, il semblerait assez normal qu'il y a une meilleure prédiction pour l'année 2019 que pour l'année 2020 car en 2019, c'est surtout en période d'été et d'hiver qu'il y a des communes éligibles (donc à priori sinistrées) et nous avons vu que les modèles de ces deux saisons sont à priori moins problématiques. En 2020, c'est surtout en période d'été et de printemps qu'il y a des communes éligibles, d'où une plus grande incertitude car pour le modèle de printemps, il n'y a à priori pas assez de données pour bien calibrer.

5.4.4 Comparaison des modèles

Dans le tableau de la figure 61, nous récapitulons les prédictions des charges totales pour les années 2019 et 2020 des trois modèles, à savoir modèle de référence, modèle zero inflaté annuel, et modèle

Figure 60 – De gauche à droite et de haut en bas, nous avons respectivement la base de données d'hiver, de printemps, d'été et d'automne. Nous représentons l'emplacement des données des années 2016 (noir), 2017 (rouge) et 2018 (orange) par des ellipses de niveau 95% (i.e l'ellipse contient 95% des données) et les données des années 2019 (vert) et 2020 (bleu) par des points dans la plan formé par les deux composantes principales les plus importantes. Bien entendu, pour raison de visibilité, seule une partie des données sont représentées.

saisonnier exprimées en pourcentage de la charge totale des sinistres observée en 2018.

Charge totale	Early IBNR	Référence	$\mathcal{ZI}$ annuel	$\mathcal{ZI}$ saison
2019	57.7%	79.2%	61.8%	80.7%
2020	84.6%	82.9%	44.3%	87.2%

Figure 61 – Tableau récapitulatif des charges totales observées et des charges totales prédites pour les années 2019 et 2020 par les trois modèles, à savoir modèle de référence, modèle zero inflaté annuel, et modèle saisonnier exprimées en pourcentage de la charge totale des sinistres observée en 2018.

De manière générale, les prédictions du modèle saisonnier sont les plus cohérentes lorsque nous les comparons aux prédictions du modèle de référence et aux estimations de la CCR ([?]). En effet, la charge prédite 2019 un peu surestimée (figure 59) a pu être expliquée par l'ACP, i.e un modèle d'automne qui surestime le nombre de sinistres 2019. Ainsi, compte tenu de nos précédentes analyses et comparaisons, le modèle saisonnier est le modèle à retenir. Les charges totales 2019 et 2020 prédites sont certes plus incertaines mais l'ensemble est cohérent avec les informations fournies par les différentes sources. De plus, le modèle est très prometteur puisque nous rappelons qu'ici, nous avons seulement trois années de données exploitable à une maille relativement fine avec en plus les données des années 2016-2017 qui sont des données actualisées. Ainsi, avec davantage d'années où les nouveaux critères de reconnaissance sont appliqués, les prédictions du modèle saisonnier avec quatre modèles zero inflatés sous-jacents peuvent être grandement améliorées.

Notez aussi que le modèle saisonnier répond à l'objectif actuariel de l'étude. En effet, pour une année donnée, les variables explicatives sont toutes des variables qui peuvent être connues au plus tard un mois après que l'année soit écoulée. En d'autres termes, avec ce modèle, nous avons une estimation raisonnable de la charge IBNR totale annuelle au plus tard un mois après que l'année soit écoulée. Ce délai de 1 mois est à comparer avec les 6 mois de la CCR ou les 12 mois pour un assureur. Il s'agit donc d'un réel avantage puisque cela permet une gestion plus fine des provisions au niveau du bilan.

5.5 Le rôle d'un actuaire

Lors de cette étude, nous avons vu d'une part des problématiques liées à la disponibilité des données comme par exemple les données utilisables qui ne sont à une résolution optimale que pour 2016, 2017 et 2018, mais aussi des problématiques de modélisation comme l'excès de zéro. La démarche adoptée a toujours été d'y répondre clairement à l'aide d'une approche plus fine et d'une théorie plus adéquate tout en accordant une attention particulière à l'évolution de la réglementation. Au cours de cette démarche, nous avons vu de nombreux points sur lesquels l'actuaire est sollicité.

Le rôle de ce dernier consiste d'abord à apporter une attention toute particulière aux données en sa possession. Il doit être en mesure d'évaluer la qualité et la pertinence de ces dernières, ses limites et surtout ses particularités. En effet, par exemple dans le cas de cette étude, nous utilisons un portefeuille représentatif de la France, ce qui rend le travail plus facile puisque l'approche est globale. Dans le cas où le portefeuille serait régional, ce dernier doit être en mesure d'identifier les particularités liées à la région et notamment les biais que cela peut introduire. Sa démarche par la suite serait donc adaptée à ces particularités qu'il a identifié.

Un second point sur lequel l'actuaire est sollicité consiste dans la modélisation du risque du portefeuille. Plus précisément, ce dernier doit identifier quels sont les facteurs pertinents pour modéliser le risque. Dans le cas du risque de subsidence en particulier, comme il n'y a pas d'indice universelle pour la sécheresse, nous avons vu que le choix des indices est crucial. En effet, ces derniers sont souvent adaptés pour capturer un certain type de climat et il faut alors utiliser les indices de sécheresse adaptés à chaque zone géographique. De plus, il faut aussi prendre en considération la dimension temporelle du risque modélisé, puisque pour la subsidence par exemple, nous avons vu que les fenêtres temporelles à privilégier sont des fenêtres de 3 mois, 6 mois ou 12 mois car il s'agit d'un phénomène avec une

cinétique lente. Ce dernier est donc tenu de connaître le risque de son portefeuille, notamment les phénomènes sous-jacents qui génèrent ce risque, et élaborer une approche qui le sied.

Un autre point crucial nécessitant l'expertise d'un actuaire concerne la réglementation. En effet, pour la subsidence, nous avons vu que l'instabilité des critères de reconnaissance rend des approches historiques difficiles notamment lorsque nous essayons de construire un modèle prédictif. En d'autres termes, ce dernier doit rester dynamique en termes de suivi des réglementations. Il doit être en mesure de les comprendre, les interpréter voire les calculer quantitativement afin de s'assurer que les produits dont il a la charge restent valables et pertinents dans le nouveau contexte règlementaire.

En somme, l'actuaire est attendu de manière générale en matière d'expertise, par exemple en statistique, en mathématique ou en techniques de modélisation. Ce dernier n'est pas tenu de comprendre en détails la moindre des subtilités des avancées techniques récentes mais doit être en mesure de les appliquer si besoin et d'apporter un regard critique sur les résultats obtenus. Ainsi, il est invité à maintenir à jour son niveau de connaissances professionnelles, ce qui est notamment le cas dans le contexte actuel avec une évolution rapide des contraintes réglementaires.

5.6 Conclusion

Dans cette partie, nous avons dans un premier temps intégré les nouvelles données de SWI Uniforme de Météo-France dans la base de données de catastrophe naturelle en mettant en évidence l'importance de prendre en compte l'évolution de la réglementation dans les modélisations. En effet, cette dernière change et s'adapte régulièrement en fonction des besoins politiques et rend les études historiques de prédiction difficile. Dans un second temps, nous avons mis à jour le modèle d'occurrence-fréquence, en définissant en passant un modèle de référence, mais surtout, nous avons montré l'intérêt de la théorie des modèles inflatés pour cette étude. L'utilisation de cette théorie dans un troisième temps en divisant adéquatement la base de données de catastrophe naturelle en quatre au préalable a permis de construire un modèle saisonnier composé de quatre modèles zéro inflaté sous-jacents. Chaque modèle sous-jacent caractérise une saison particulière et le modèle saisonnier combine ces derniers à l'aide d'une somme.

La comparaison des prédictions des différents modèles, combinée avec la visualisation des données dans l'espace des composantes principales, montre que les prédictions du modèles saisonnier sont les plus cohérentes dans l'ensemble. Il s'agit donc du modèle à retenir sachant qu'il reste encore une grande marge d'amélioration à ce dernier puisque pour le moment, en raison de la disponibilité des données, il n'utilise que trois ans de données (2016-2018) avec en plus les données des années 2016-2017 qui sont des données actualisées. Notez aussi qu'elle répond à l'objectif actuariel de l'étude car pour une année donnée, les variables explicatives peuvent être connues au plus tard un mois après que l'année soit écoulée. En d'autres termes, avec ce modèle, nous avons une estimation raisonnable de la charge IBNR totale annuelle au plus tard un mois après que l'année soit écoulée. Ce délai de 1 mois est à comparer avec les 6 mois de la CCR ou les 12 mois pour un assureur. Il s'agit donc d'un réel avantage puisque cela permet une gestion plus fine des provisions au niveau du bilan.

In fine, dans un quatrième temps, nous rappelons le parcours parcimonieux de cette étude en mettant en avant les différents aspects sur lesquels un actuaire est sollicité, à savoir une connaissance profonde de la réglementation, du risque et des données mises à sa disposition ainsi qu'une expertise techniques avancée.

6 Discussion

Lors de la première approche départementale, nous avons pu définir différents indices de sécheresse qui nous semblent intéressants ainsi que l'intérêt du modèle linéaire généralisé et de l'analyse en composantes principales dans l'étude tout en mettant en évidence l'insuffisance d'une approche départementale, notamment en terme de quantité de données.

Pour répondre à ces problèmes soulevés, nous affinons l'approche en passant de la maille départementale à la maille communale. Cette approche plus fine sur le plan spatial s'est accompagnée d'une tentative de modélisation du phénomène de subsidence en deux parties, ce qui nous a invité à établir toute une méthodologie dans la sélection des variables explicatives pour chaque partie du modèle. Ces méthodologies, basées souvent sur des approches ensemblistes, sont applicables dans d'autres problématiques de modélisations similaires nécessitant une étape de reconnaissance du risque. Cependant, un contrôle préliminaire des variables sur lesquelles les méthodes seront appliquées est parfois nécessaire afin de garantir la cohérence entre les variables sélectionnées par les méthodologies et le phénomène physique modélisé. Les résultats finaux obtenus à l'aide de ces outils, bien que prometteurs avec l'introduction de la fonction d'influence, ne sont malheureusement plus parfaitement à jour en raison de la mise à disposition des données de SWI Uniforme par Météo-France.

L'intégration de ces nouvelles données dans le projet, grandement facilitée par les méthodes déjà établies, nous invite à considérer des approches plus fines sur le plan temporel. À l'aide de la théorie des modèles inflatés, nous avons revu l'approche communale pour, in fine, aboutir à une modélisation saisonnière de la sinistralité subsidence. Ce modèle saisonnier est composé de quatre modèles sous-jacents inflaté en zéro et découle des quatre critères climatiques actuels de reconnaissance. L'analyse des différents résultats montre que ce modèle produit des résultats cohérents dans l'ensemble et estime correctement les charges totales des sinistres de manière générale. Par ailleurs, l'analyse des résultats montre qu'il y a un réel manque de diversité dans les années et un manque de données pour certaines saisons. Néanmoins, ceci est inhérent au fait que les nouveaux critères de reconnaissance ont été mis en place que très récemment, c'est-à-dire en 2018. Il faut noter aussi que ce modèle saisonnier répond à l'objectif actuariel de l'étude puisque les variables explicatives peuvent être connues au plus tard un mois après que l'année soit écoulée. Autrement dit, nous avons une estimation raisonnable de la charge IBNR totale annuelle au plus tard un mois après que l'année soit écoulée. Ce délai de 1 mois est à comparer avec les 6 mois de la CCR ou les 12 mois pour un assureur. Il s'agit donc d'un réel avantage puisque cela permet une gestion plus fine des provisions au niveau du bilan.

In fine, à la question de la modélisation du risque de subsidence, nous proposons le modèle communal saisonnier comme réponse. Ce dernier, qui répond à l'objectif actuariel de l'étude, est en partie motivé par la qualité de ses prédictions actuelles mais surtout pour son potentiel futur. La construction de ce dernier nous a permis d'établir entre autres des méthodologies utilisables dans d'autres problématiques similaires nécessitant une étape de reconnaissance du risque. Ainsi, l'approche saisonnière est plutôt nouvelle dans son genre et le modèle saisonnier avec quatre modèles sous-jacents du même type est caractéristique du risque de subsidence au sein du régime CatNat français car il s'agit d'un héritage direct des critères de reconnaissance actuels. Les résultats obtenus sont prometteurs tout en révélant les différentes limites comme le problème de manque de données. Cependant, cette dernière limitation sera rapidement surmontée dans les années à venir et il sera clairement intéressant de revisiter l'approche saisonnière à ce moment-là et peut-être même envisager une approche mensuelle si le besoin se révèle et si les données le permettent.

7 Note de synthèse

Introduction

Ce mémoire d'actuariat est réalisé au sein de l'équipe Agriculture & Parametrics de Liberty Mutual Re. Le sujet du mémoire porte sur l'étude et la modélisation du phénomène de subsidence. Plus précisément, nous nous intéressons au phénomène de retrait-gonflement des argiles et aux conséquences de ce dernier sur les constructions dans le cadre du régime de catastrophe naturelle français et à travers un portefeuille représentatif de ce risque en France. L'objectif du projet est de modéliser la charge IBNR totale nationale des sinistres de sécheresse pour une année donnée dès lors que cette dernière se termine à l'aide des variables dérivant des caractéristiques des contrats, des données climatiques, des données géologiques et des données socio-économiques. L'intérêt actuariel de l'étude réside donc dans cette connaissance rapide de la charge IBNR car cela permet une gestion plus fine des provisions au niveau du bilan.

Pour y parvenir, le travail est organisé en différentes approches. Dans une première partie, nous faisons une approche annuelle et à l'échelle départementale tout en présentant le phénomène de subsidence, le contexte de l'étude, les données brutes à disposition et les indices de sécheresses sollicités. L'exploration des données à travers d'abord un modèle GLM de base, puis des modèles toujours de régression mais améliorés à partir de cette base GLM, montre que l'approche départementale est trop restrictive en raison de la résolution de la maille départementale. Pour répondre à cela, une approche toujours annuelle mais à l'échelle communale est construite dans une seconde partie en intégrant les données du Journal Officiel portant sur l'obtention ou non des décrets de catastrophe naturelle par les communes, et aboutit à un modèle dit d'occurrence-fréquence qui comprend deux modèles sous-jacents. Au cours de cela, toute une méthodologie autour de la sélection des variables et construction des modèles basée sur des méthodes de machine learning est développée. Les résultats de prédictions obtenus lors de cette approche sont très prometteurs mais se trouvent être incomplets en raison de la publication des nouvelles données par Météo-France portant sur le SWI Uniforme. Ainsi, dans la troisième partie, nous intégrons dans l'étude ces nouvelles données utilisées par la commission interministérielle et à partir desquelles les critères de reconnaissance sont calculés. À l'aide de cela, nous mettons à jour le modèle occurrence-fréquence mais surtout, nous construisons un dernier modèle saisonnier en utilisant la théorie des modèles inflatés. Ce dernier, composé de quatre modèles sous-jacents, est plus adapté aux nouveaux critères de reconnaissance appliqués depuis 2018. In fine, l'approche saisonnière à l'échelle communale, avec des variables explicatives qui sont connues au plus tard un mois après qu'une année soit écoulée, conduit à des résultats satisfaisants tout en mettant en évidence un manque de diversité dans les données annuelles et un manque de données pour certaines saisons.

Approche départementale

L'approche annuelle à l'échelle départementale passe d'abord par un traitement des données brutes issues des sources qui a pour objectif d'éliminer les incohérences au sein des données et d'estimer les données manquantes à l'aide des méthodes de lissage géographique. Ces traitements permettent d'obtenir une base homogène et complète regroupant les informations portant sur le climat et la géologie, et à partir de laquelle il est possible de calculer des indices de sécheresse. Les indices de sécheresse calculés sont :

- Standardized Precipitation Index ou SPI.
- Standardized Precipitation Evapotranspiration Index ou SPEI.
- Standardized Precipitation Temperature Index ou SPTI.
- China-Z Index ou CZI.

- Effective Drought Index ou EDI.

Nous utilisons ensuite ces indices calculés ainsi que la base de données agrégée à la maille départementale pour expliquer le taux de destruction par département, i.e la charge totale des sinistres par département divisée par la valeur totale des constructions assurées.

Pour expliquer la relation entre le taux de destruction par département et les variables géologiques et climatiques, nous avons testé plusieurs modèles, à savoir le GLM, le modèle composé et le modèle de régression sur composantes principales. Le modèle GLM avec distribution gamma atteint rapidement ses limites lorsque nous analysons ses résultats mais il fournit néanmoins une référence de base. Une amélioration du modèle linéaire consiste à appliquer en amont un modèle binomial pour tenir compte du nombre important de zéros, i.e de la rareté des catastrophes naturelles. Néanmoins, les améliorations des prédictions ne sont pas suffisantes.

Le modèle de régression sur composantes principales est ensuite envisagé en dernier lieu de cette approche départementale. Ce modèle plus difficile à interpréter optimise le nombre de variables à utiliser dans la régression tout en préservant un maximum d'information. L'amélioration des prédictions reste insuffisante mais le modèle nous suggère qu'il y a un réel problème de manque des données lorsque nous représentons les points de données dans la base formée par les deux composantes principales les plus importantes. En effet, dans certaines régions de l'espace, les points ne proviennent que d'une année particulière, ce qui est problématique car cela suggère qu'il n'y a pas assez de données pour ces zones.

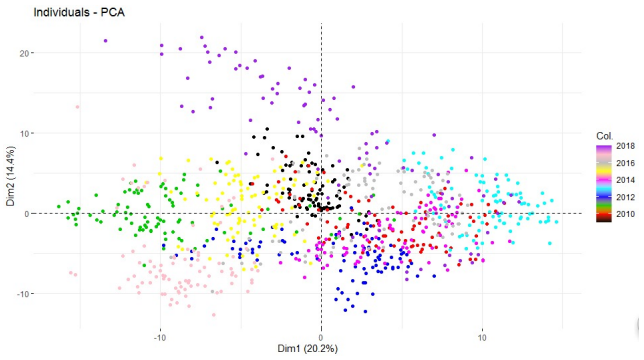

Figure : Nuage de points des données projetées dans les deux directions où les variances expliquées sont les plus grandes avec en échelle de couleurs les années. Les observations de l'année 2018 correspondent aux points violets et se distinguent des autres.

Approche communale

Dans cette seconde approche, moyennant quelques modifications dans la définition des variables climatiques et géologiques et l'ajout des variables socio-économiques, nous agrégeons la base de données à la maille communale. Cela nous permet d'avoir beaucoup plus de données d'observations puisque nous avons environ 35 000 points par années contre un peu moins de 100 points par année dans l'approche départementale. Néanmoins, en raison des limitations des données, la période d'étude n'est plus 2009-2018 mais 2016-2018.

Par ailleurs, nous modélisons non plus directement le taux de destruction mais faisons plutôt une approche en deux parties. En effet, afin de refléter au mieux le processus de décret, i.e reconnaissance de l'état de catastrophe naturelle par une commission interministérielle puis indemnisation, cette modélisation naturelle dite occurrence-fréquence modélise d'abord la probabilité qu'une commune soit décrétée en situation de catastrophe naturelle par la commission interministérielle puis ensuite la proportion de contrats sinistrés des communes. Pour obtenir la charge d'un sinistre, un coefficient de charge moyenne est estimé en amont. Au cours de la construction des deux modèles, nous définissons une méthodologie de sélection des variables applicables dans d'autres contextes similaires nécessitant une étape de reconnaissance du risque.

Pour le modèle d'occurrence, la sélection des variables consiste d'abord à faire une sélection à l'aide d'une méthode ensembliste, puis ensuite faire une sélection basée sur les distances entre distributions. Le modèle d'occurrence est ensuite construit en choisissant parmi les modèles de régression linéaire, les modèles de gradient boosting, les modèles de forêt aléatoire et les modèles de réseaux de neurones, celui qui minimise le critère d'AUC globale, moyennant une optimisation des hyperparamètres de ces modèles candidats au préalable. Le minimum d'AUC est atteint pour le modèle de gradient boosting mais la différence entre celui-ci et le modèle linéaire ridge étant minime, le choix est porté sur le second en raison de sa simplicité.

Pour construire le modèle de fréquence, la méthode de sélection des variables consiste à filtrer les variables de variance quasi nulle puis les variables colinéaires. L'algorithme boruta est ensuite appliqué aux variables restantes et une dernière sélection stepwise est appliquée à la sortie de cet algorithme. Le modèle de fréquence est ainsi un modèle GLM poisson avec un terme d'exposition.

Les résultats directs de la modélisation occurrence-fréquence ne sont pas totalement satisfaisants mais l'analyse des prédictions à l'aide des cartes de France montre que le modèle est convenable et nécessite finalement que peu d'ajustement. La correction est donc réalisée via une fonction déterministe dite d'influence visant à recalibrer la sortie du modèle d'occurrence. Les résultats de cette correction sont prometteurs puisque le nombre de communes sinistrées est correctement prédit pour 2018 et 2019. Les prédictions antérieures à 2018 sont incorrectes probablement en raison de la modification des critères de reconnaissance en 2018.

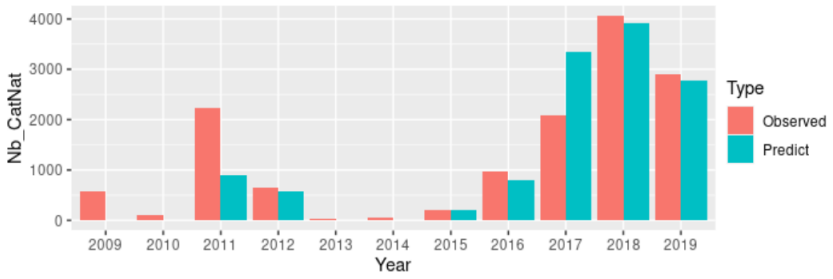

Figure : Le nombre de communes décrétées en état de catastrophe naturelle observé et prédit. Les prédictions sont corrigées par la fonction d'influence. Les années 2018 et 2019 sont bien corrigées mais pas les années antérieures à 2018 en raison de la modification des critères de reconnaissance en 2018.

Approche saisonnière

En raison de la publication des données de SWI Uniforme par Météo-France vers avril 2021, l'approche occurrence-fréquence précédente se trouve être incomplète. En effet, les critères de reconnaissance de l'état de catastrophe naturelle en vigueur depuis 2018 reposent sur l'utilisation des données de SWI Uniforme. Le modèle d'occurrence visant à approcher ces critères n'est donc plus parfaitement à jour puisque les critères peuvent désormais être calculés exactement.

Néanmoins, dans cette nouvelle procédure de reconnaissance, il est toujours nécessaire que la commune sinistrée fasse une demande d'arrêté, et c'est pourquoi nous remplaçons le modèle d'occurrence par un modèle de probabilité demanderesse, réalisé à l'aide d'une forêt aléatoire. Une analyse des résultats de cette mise à jour montre que le modèle biparti est loin d'être parfait notamment en raison de l'excès de zéros. Une amélioration envisagée est donc de regrouper les deux modèles en un à l'aide de la théorie des modèles inflatés.

À l'aide de cette théorie, un modèle zero inflaté dit annuel est calibré en supprimant pour les variables climatiques, les variables propres à une saison particulière. Ce modèle prédit le nombre de sinistres pour chaque commune et nous obtenons ainsi la charge totale en appliquant un coefficient de charge moyenne. Un second modèle consistant en une somme de quatre modèles zero inflatés est ensuite construit afin de tenir compte de manière individuelle les quatre critères de reconnaissance climatiques

utilisés par la commission. Chacun des quatre modèles caractérise donc une saison particulière et vise à prédire le nombre de sinistres observés lors de la saison associée. Pour cela, les variables climatiques, comme les communes éligibles, sont divisées en quatre catégories et pour chaque saison, nous utilisons les variables climatiques de la saison en plus des variables géologiques et socio-économiques qui sont indépendantes de la saison.

Au préalable de ces deux modèles zero inflatés annuel et saisonnier, un modèle de référence est construit et utilisé comme point de comparaison. Ce modèle simpliste consiste à régresser la charge totale de chaque année sur le nombre de communes éligibles de chaque année. Les charges prédites de ces trois modèles, en pourcentage de la charge totale des sinistres observée en 2018, sont fournies dans le tableau suivant. Ces chiffres sont à comparer aux estimations de la CCR dans la première colonne du tableau.

Charge totale	**Early IBNR**	**Référence**	$\mathcal{ZI}$ **annuel**	$\mathcal{ZI}$ **saison**
2019	57.7%	79.2%	61.8%	80.7%
2020	84.6%	82.9%	44.3%	87.2%

Figure : Tableau récapitulatif des charges totales observées et des charges totales prédites pour les années 2019 et 2020 par les trois modèles, à savoir modèle de référence, modèle zero inflaté annuel, et modèle de saison combiné. Les chiffres sont exprimés en pourcentage de la charge totale des sinistres observée en 2018.

Par ailleurs, comme dans l'approche départementale, en représentant les données dans l'espace des deux composantes principales les plus importantes, nous observons que le manque de diversité des données dans les années est partiellement corrigé notamment pour le modèle de la saison d'été. En comparant ces représentations spatiales entre eux et avec la connaissance du nombre de communes éligibles dans chaque saison, les prédictions du modèle saisonnier sont cohérentes dans l'ensemble et correctes de manière générale. De plus, il faut noter que nous disposons uniquement de trois années de données avec en plus les données des années 2016 et 2017 qui sont des données actualisées puisque les nouveaux critères de reconnaissance ne sont appliqués qu'à partir de 2018.

In fine, le modèle saisonnier est à retenir car d'une part, ce dernier a encore une grande marge d'amélioration possible, et d'autre part, ses résultats sont les plus cohérents avec les estimations de la CCR les plus récentes. Il répond aussi à l'objectif actuariel puisque les variables explicatives sont connues au plus tard un mois après qu'une année soit écoulée.

Conclusion

Nous avons réalisé une première approche annuelle à l'échelle départementale en construisant les indices de sécheresses qui semblent pertinents à l'étude. Cette première approche révèle rapidement ses limites et nous tentons d'y répondre dans une seconde approche annuelle mais à l'échelle communale tout en définissant des méthodologies de sélection des variables applicables dans d'autres contextes similaires nécessitant une étape de reconnaissance du risque. Les résultats de l'approche communale, bien que prometteurs avec l'introduction de la fonction d'influence, ne sont malheureusement plus parfaitement à jour avec la mise à disposition des données de SWI Uniforme par Météo-France. Par conséquent, une troisième approche sur la base des modèles zero inflatés est réalisée en intégrant ces nouvelles données. Le modèle saisonnier retenu est composé de quatre modèles zero inflatés sous-jacents dans le but refléter les quatre critères de reconnaissance climatiques utilisés par la commission interministérielle. L'analyse des résultats à l'aide de l'outil d'analyse en composantes principales montre que le modèle saisonnier est prometteur et possède une grande marge d'amélioration tout en répondant à l'objectif actuariel de l'étude.

8 Executive summary

Introduction

This report was written following our work within the Agriculture & Parametrics team of Liberty Mutual Re and deals with the modelling of risk of subsidence, a geological phenomenon. More precisely, we are interested in the shrinking and swelling of clays and its consequences over buildings within the French natural catastrophe regime (CatNat) provided by the Caisse Centrale de Réassurance (CCR) and with a portfolio representative of France. The aim of this project is to model the national IBNR amount of claims for a given year as soon as it ends with variables extracted from contracts, climate data, geological data and socioeconomic data. Thus, the actuarial interest of the study is the quick access to an estimate of the national IBNR amount of claims since it allows a better management of provisions for insurers.

For this purpose, the study is organized in several approaches. In the first part, we conduct an annual approach with regional spatial resolution by introducing the subsidence phenomenon, the context of this study, the availability of raw data and drought indices. An exploration of data through regression based models shows that the approach is too restrictive due to the regional resolution which limits the amount of data. To answer that, we propose a second approach, still annual, but with a local spatial resolution. That approach, by integrating CatNat data from the Journal Officiel, yields an occurrence-frequency two-step model. During the construction, a whole methodology about feature engineering and model development, based on machine learning techniques and which can be applied in other modelling problems with a recognition step, is proposed. Obtained results from the two-step model are promising but unfortunately incomplete with the publication of Uniform SWI data by Météo-France in April 2021. Therefore, in a third part, we integrate these new data used by the interministerial commission in the calculation of the actual recognition criteria. Thanks to that, using the inflated model framework, we proposed a seasonal model which mixed four zero inflated models, highly suitable for the current recognition criteria because each model deals with one seasonal recognition criterion specifically. An analysis through PCA shows that the lack of data isn't totally resolved but even in that case, the seasonal approach with local spatial resolution yields consistent results.

Regional approach

The annual approach with regional spatial resolution comes first with data processing. Indeed, we have to mix multiple data sets by removing some inconsistencies with geographical smoothing methods to finally obtain a usable data set. Through that, the complete data set contains climate and geological variables. The drought indices used during the construction are :

- Standardized Precipitation Index ou SPI.
- Standardized Precipitation Evapotranspiration Index ou SPEI.
- Standardized Precipitation Temperature Index ou SPTI.
- China-Z Index ou CZI.
- Effective Drought Index ou EDI.

Next to that, we calculate the destruction rate which is defined as, for each department, the total amount of claims divided by the total amount of insured constructions.

To explain the relationship between the destruction rate and the geological and climate variables, we tested multiple regression based models. The first tested model is a GLM model with gamma distribution but rapidly reaches its limits due to the excess of zeros. An improved version is to combine GLM gamma with another GLM binomial model to take into account the excess of zeros but the

improvement in terms of prediction isn't sufficient.

With the failure of our previous attempts, we wonder if we have correctly selected our explanatory variables. To verify that, we construct a GLM model using the principal components of the data set as explanatory variables. That model is hard to interpret but optimizes the number of variables without losing too much information. The results show that there is an improvement in terms of predictions but also, above all, a lack of data. Indeed, when we represent our data set on the base formed by the two main principal components, for some regions of the space, the data comes from just one particular year, which suggests a lack of data for this particular region.

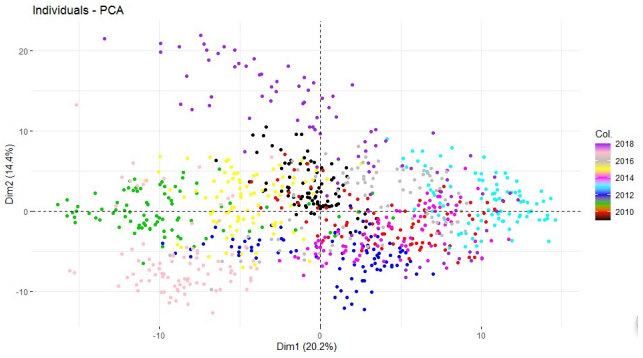

Figure : Cloud of dots representing the data set projected on the base formed by the two main principal components. The observations of the most recent year, namely 2018, are in purple and distingush from others.

Municipal approach

In this second approach, with some modifications of definitions of climate variables and geological variables and the addition of socioeconomic variables, we construct a data set with municipal spatial resolution. That means we have approximately 35 000 points instead of less than 100 points for each year. Of course, the period of the study changes from 2009-2018 to 2016-2018 due to the unavailability of portfolio data at municipal scale for other years.

Additionally, to model the recognition process of the French CatNat regime, we adopt a two-step modelling framework, namely occurrence-frequency model. The first part models the probability of a given town to be granted a natural disaster status from the interministerial commission and the second part models the number of contracts which are impacted by the disaster. To obtain the amount of claims, we use a coefficient of average claim amount calculated in advance and supposed to be constant over time. During the construction of the occurrence-frequency model, a whole methodology about feature engineering and model development, based on machine learning techniques and which can be applied in other modelling problems with a recognition step, is proposed.

More precisely, for the occurrence model, we select the explanatory variables first with an ensemble method and we filter the result through another selection method based on distances between distributions. The model is finally constructed by choosing between several candidates : elastic net GLM model, gradient boosting models, random forest models, and neural network models. The hyperparameters of candidate models are optimized and the selection criteria is the AUC. The minimum AUC is reached by the gradient boosting model but the difference between that and the GLM ridge model is negligible. Therefore, we choose the GLM ridge as the occurrence model because it is simpler, and thus easier to contractualize.

To construct the frequency model, we filter the explanatory variables through several operations. In order, we remove variables with a variance nearly zero, we remove variables which are collinear, we apply the boruta algorithm, and finally we apply a stepwise selection method. The obtained frequency model is a GLM model with poisson distribution and an exposition term constructed with the selected variables.

The direct results from the occurrence-frequency model are not totally satisfactory but a deeper analysis with French territory map reveals that we just need to bring some corrections to the occurrence model. Thus, the correction is done by applying a deterministic function to the outcome of the occurrence model and the results are promising. The total amounts of claims for 2018 and 2019 are correctly predicted. The predictions for years earlier than 2018 aren't correct but it is mainly due to the modification of recognition criteria by the government in 2019.

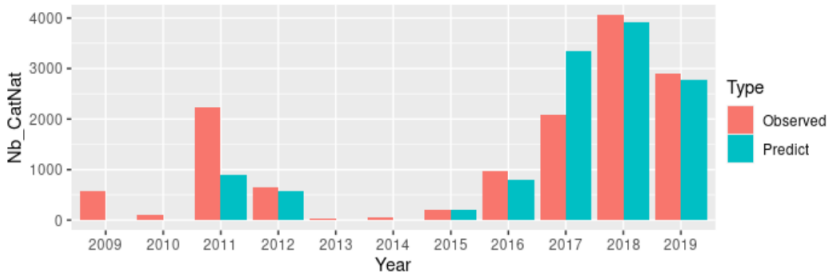

Figure : The predicted and observed number of towns which obtained the recognition. Predictions are corrected by the influence function and the total amounts of claims for 2018 and 2019 are correctly predicted.

Seasonal approach

In April 2021, Météo-France published the Uniform SWI data, which makes our previous approach incomplete. Indeed, the occurrence model is designed to approximate the recognition criteria but with the new data, these criteria can now be exactly calculated.

Nonetheless, the current recognition process still needs the mayor of the town to formulate a recognition request. Thus, the two-step approach is still relevant if we update the occurrence model. Indeed, the model has now to predict if the mayors will request or not a recognition of the state of natural disaster. The update is done by replacing the occurrence model by a random forest model. The results from the updated two-step approach is unfortunately not satisfactory due to the excess of zeros which is not explicitly modelled. Therefore, we proposed an improvement based on inflated model framework by bringing together the two-step in one.

Thanks to that framework, an annual zero inflated model is calibrated by removing climate variables which characterized one season in particular. After that, a second model which combines four zero inflated models is constructed. The model is highly suitable for the current recognition criteria because each model deals with one seasonal recognition criteria specifically. Of course, during the construction of each underlying model, in terms of climate variables, we use only variables which characterize the associated season. In terms of geological and socioeconomic variables, they remain the same because they are independent from the season. In other words, we spread our data set in four subsets.

Note that before these two models, we constructed a reference model which is simply the regression of total amount of claims by the total number of municipals which are eligible in terms of recognition criteria. Finally, the predicted amount of claims for 2019 and 2020 by the three models, in percentage of total amount of claims observed in 2018, is in the following table.

From the table, the seasonal model, more consistent with estimations of CCR, is the best. Indeed, like in our regional approach, by representing our data sets in the base formed by the two main principal components, we observe that the lack of data is partially corrected, especially for the summer data set. By comparing these spatial representations between them with knowledge of the number of eligible municipals by season, we conclude that the results from the seasonal model are consitent and its predictions are correct in general. Additionnally, the independant variables of the model can be known, and so the national IBNR amount of claims, at the latest one month after a year passed, which allow a better management of provisions for insurers. Note also that the model used only three years

Total claims	Early IBNR	Reference	$\mathcal{ZI}$ annual	$\mathcal{ZI}$ season
2019	57.7%	79.2%	61.8%	80.7%
2020	84.6%	82.9%	44.3%	87.2%

Figure : Table of total amounts of claims for 2019 and 2020 predicted by the reference model, by the annual zero inflated model and by the seasonal model. The amounts are expressed in percentage of total amount of claims observed in 2018.

of data. Among these data, data of 2016 and 2017 are updated data because the current recognition criteria is not applied before 2018. Therefore, this suggests that the seasonal model still has room for improvement.

Conclusion

We realized a first annual approach with regional spatial resolution by defining drought indices which are adapted to our portfolio. That first approach reached its limits rapidly and we override them by introducing a second approach, still annual, but with a municipal spatial resolution. During that, we construct a whole methodology about feature engineering and model development, based on machine learning techniques and which can be applied in other modelling problems with a recognition step. The results of the occurrence-frequency model are promising with the introduction of an influence function but are also incomplete due the Uniform SWI data published by Météo-France in april 2021. Therefore, by integrating these new data, we proposed a third approach based on inflated model framework. The seasonal model finally obtained consists of four underlying zero inflated models which aim at modelling the four seasonal recognition criteria used by the interministerial commission. The results of this model are satisfactory because it still has great room for improvement while providing tools for a better management of provisions for insurers.

9 Annexe

9.1 Indices complémentaires

9.1.1 Pinna Combinative Index

Il s'agit d'un indice développé vers 1992 et est défini par ([17]) :

$$I_P = \frac{1}{2}\left(\frac{P}{10+T} + \frac{12P_d}{10+T_d}\right) \tag{99}$$

Avec P et T la précipitation et la température moyenne annuelle et P_d et T_d la précipitation et la température moyenne du mois le plus sèche.

9.1.2 SMDI : Soil Moisture Deficit Index

Ce dernier utilise comme entrée la quantité d'eau stockée dans le sol notée SW, qui est une sortie du modèle SWAT et qui nécessite de nombreuses données complexes. L'indice SMDI est ensuite défini par ([21], [19]) :

$$SMDI_j = 0.5SMDI_{j-1} + \frac{SD_j}{50} \quad \text{avec} \quad SD_{ij} = \begin{cases} \dfrac{SW_{ij} - MSW_j}{\max(SW_j) - MSW_j} & \text{si} \quad SW_{ij} > MSW_j \\ \dfrac{SW_{ij} - MSW_j}{MSW_j - \min(SW_j)} & \text{si} \quad SW_{ij} = MSW_j \end{cases} \tag{100}$$

Avec i l'année, j la semaine, SW_{ij} la quantité d'eau moyenne hebdomadaire stockée dans le sol, MSW_j la médiane et $SMDI_0 = 0$.

Cet indice est développé par les auteurs notamment pour la gestion de sécheresse agricole et a notamment une meilleure résolution temporelle que l'indice SPI (données hebdomadaires au lieu de données mensuelles).

9.1.3 WASP : Weighted Anomaly of Standardized Precipitation

Cet indice utilise les données de précipitation mensuelle et tend à être calculé pour $n = 3, 6, 12$ mois. Il est notamment employé pour la gestion des sécheresses dans les régions tropicales humides et est défini par ([22], [19]) :

$$WASP_n = \frac{S_n}{\sigma_{S_n}} \quad \text{avec} \quad S_n = \sum_{i=1}^{n}\left(\frac{P_i - \overline{P_i}}{\sigma_i}\right)\frac{\overline{P_i}}{\overline{P_A}} \tag{101}$$

Où P_i est la précipitation du mois i, $\overline{P_i}$ et σ_i respectivement la moyenne et l'écart-type, et $\overline{P_A}$ la moyenne annuelle. Nous notons que le coefficient $\overline{P_i}/\overline{P_A}$ sert notamment à prévenir les anomalies lors des saisons sèches.

9.1.4 RDI : Reconnaissance Drought Index

Cet indice utilise des données de précipitation mensuelle et des données d'évapotranspiration qui peuvent être estimées comme dans le calcul de SPEI. Il est défini par ([23], [19]) :

$$RDI^{(i)} = \frac{P_i}{PET_i} \quad \text{ou} \quad RDI_n^{(i)} = \frac{RDI^{(i)}}{\overline{RDI}} - 1 \quad \text{ou} \quad RDI_{st}^{(i)} = \frac{\ln(RDI^{(i)}) - \overline{\ln(RDI^{(i)})}}{\sigma(\ln(RDI^{(i)}))} \tag{102}$$

Avec P_i la précipitation du mois i, PET_i le potentiel d'évapotranspiration du mois i, $\sigma(.)$ la fonction écart-type et $\overline{(.)}$ la fonction moyenne.

La seconde expression, appelée indice RDI normalisé, a un comportement similaire au SPI et la troisième expression suppose notamment que l'indice RDI initial suit une distribution lognormale.

9.1.5 CDI : Combined Drought Index

Il s'agit d'un indice qui prend en compte la précipitation et sa persistance, la température et sa persistance et l'humidité du sol et sa persistance ([24], [19]). Cet indice essaye donc de rendre compte des trois types de sécheresse (météorologique, agricole, hydrologique) de manière générale. Ainsi, les auteurs définissent un indice de précipitation, noté PDI, un indice de température, noté TDI, et un indice d'humidité, noté VDI, par :

$$PDI_{i,m} = \frac{\frac{1}{IP}\sum_{j=0}^{IP-1} P^*_{i,(m-j)}}{\frac{1}{n \times IP}\sum_{k=1}^{n}\sum_{j=0}^{IP-1} P^*_{k,(m-j)}} \sqrt{\frac{RL^*_{i,m}(P)}{\frac{1}{n}\sum_{k=1}^{n} RL^*_{k,m}(P)}} \tag{103}$$

$$TDI_{i,m} = \frac{\frac{1}{IP}\sum_{j=0}^{IP-1} T^*_{i,(m-j)}}{\frac{1}{n \times IP}\sum_{k=1}^{n}\sum_{j=0}^{IP-1} T^*_{k,(m-j)}} \sqrt{\frac{RL^*_{i,m}(T)}{\frac{1}{n}\sum_{k=1}^{n} RL^*_{k,m}(T)}} \tag{104}$$

$$VDI_{i,m} = \frac{\frac{1}{IP}\sum_{j=0}^{IP-1} NDVI^*_{i,(m-j)}}{\frac{1}{n \times IP}\sum_{k=1}^{n}\sum_{j=0}^{IP-1} NDVI^*_{k,(m-j)}} \sqrt{\frac{RL^*_{i,m}(NDVI)}{\frac{1}{n}\sum_{k=1}^{n} RL^*_{l,m}(NDVI)}} \tag{105}$$

Avec IP la fenêtre temporelle, n le nombre d'années de données disponibles, $P^*_{i,m} = P_{i,m} + 1$ où $P_{i,m}$ est la précipitation du mois m de l'année i, $T^*_{i,m} = T_{max} - T_{i,m} + 1$ où T_{max} est la température maximale dans toute la base de données, et $NDVI^*_{i,m} = NDVI_{i,m} - NDVI_{min} + 0.01$ où $NDVI$ est l'indice de différence de végétation normalisé. Pour prendre en compte la persistance, les auteurs définissent $RL^*_{i,m}(P) = RL_{max}(P) - RL_{i,m}(P) + 1$ (respectivement pour T et $NDVI$) avec $RL_{i,m}(P)$ (respectivement $RL_{i,m}(T)$ et $RL_{i,m}(NDVI)$) qui est le nombre de mois consécutif le plus grand où la précipitation (respectivement température et $NDVI$) est en-dessous (respectivement au-dessus et en-dessous) de la moyenne de la fenêtre.

Ainsi, pour des données mensuelles, l'indice CDI est défini par :

$$CDI_{i,m} = 0.5PDI_{i,m} + 0.25TDI_{i,m} + 0.25VDI_{i,m} \tag{106}$$

Nous précisons qu'il existe un délai de latence entre les différents phénomènes et pour des données plus fines comme dix jours, les auteurs estiment que les phénomènes liés à la température et aux précipitations sont en avance de deux décades, i.e m devient $m-2$ pour PDI et TDI dans les formules précédentes.

Il faut aussi noter que l'indice CDI mesure à quel point les conditions actuelles sont déviées des conditions de référence, définies par les tendances à long terme des séries temporelles, et ne mesure donc pas les phénomènes physiques en soi.

9.2 Sortie du modèle linéaire de l'approche départementale

Dans la figure 62, nous représentons le tableau, sous **R**, récapitulant les variables explicatives du modèle linéaire généralisé (partie 3.3.2.2). Le modèle est composé de 10 variables dont un terme d'interaction.

9.3 Sortie du modèle logit de l'approche départementale

Dans la figure 63, nous représentons le tableau, sous **R**, récapitulant les variables explicatives du modèle logistique (partie 3.3.3). Le modèle est composé de 28 variables explicatives dont 14 termes d'interaction.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.301e+01	2.531e+00	5.140	5.35e-07	***
taux_zonier	1.665e+01	4.441e+00	3.749	0.000218	***
SPTI_180j_critere_sum	1.370e-01	4.794e-02	2.858	0.004602	**
latitude_moyen	-4.274e-01	6.258e-02	-6.830	5.85e-11	***
SA_BATI_TOT	-4.010e-10	1.333e-10	-3.009	0.002878	**
SPTI_30j_critere_ete	1.903e-01	5.024e-02	3.789	0.000188	***
SPI_180j_hiver_max	-4.415e-01	1.589e-01	-2.779	0.005850	**
SPI_180j_ete_diff_min	-4.991e-01	1.595e-01	-3.129	0.001952	**
Prop_Zone_3	1.285e-02	5.848e-03	2.198	0.028804	*
SPTI_30j_critere_sum_moyenne	3.732e-01	1.643e-01	2.272	0.023905	*
`SA_BATI_TOT:SPTI_180j_critere_sum_moyenne`	1.413e-10	6.132e-11	2.304	0.021988	*

Figure 62 – Tableau, sous **R**, récapitulant les variables explicatives du modèle linéaire généralisé. Le modèle est composé de 10 variables explicatives dont un terme d'interaction.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.217e+01	3.670e+00	6.042	1.53e-09	***
taux_zonier	3.883e+01	7.996e+00	4.856	1.20e-06	***
SPTI_180j_critere_sum	5.468e-01	8.377e-02	6.527	6.72e-11	***
latitude_moyen	-5.872e-01	8.320e-02	-7.057	1.70e-12	***
RGA_0	6.221e-02	2.852e-02	2.181	0.029175	*
SA_BATI_TOT	8.842e-10	1.642e-10	5.384	7.29e-08	***
Prop_Zone_3	5.951e-02	2.243e-02	2.653	0.007978	**
SPEI_30j_printemps_moy	-7.340e+00	1.350e+00	-5.435	5.47e-08	***
SPTI_30j_printemps_diff_amplitude	4.791e-01	2.335e-01	2.052	0.040199	*
NB_RISQUES_TOT	-5.805e-04	1.717e-04	-3.382	0.000721	***
SPTI_90j_ete_max	4.600e+00	1.154e+00	3.987	6.68e-05	***
SPTI_30j_printemps_min	9.356e-01	3.759e-01	2.489	0.012809	*
RGA_1	-1.348e-02	6.330e-03	-2.129	0.033228	*
SPEI_90j_ete_diff_max	-4.198e+00	9.816e-01	-4.277	1.90e-05	***
taux_souscription_dpt	1.927e+01	6.472e+00	2.978	0.002903	**
`RGA_0:SPEI_90j_printemps_min`	7.500e-02	2.282e-02	3.286	0.001016	**
`SA_BATI_TOT:SPEI_30j_hiver_max`	-1.414e-10	7.186e-11	-1.968	0.049103	*
`SPTI_90j_ete_max:SPEI_90j_hiver_max`	-4.491e+00	1.023e+00	-4.389	1.14e-05	***
`SPTI_30j_printemps_diff_amplitude:SPI_180j_hiver_max`	-2.717e-01	1.068e-01	-2.543	0.010988	*
`SPEI_90j_ete_diff_max:taux_destruction`	3.640e+01	1.216e+01	2.993	0.002764	**
`taux_souscription_dpt:SPTI_90j_ete_diff_min`	-1.999e-01	9.761e-02	-2.048	0.040571	*
`SPTI_90j_ete_max:SPI_90j_hiver_amplitude`	1.021e+00	3.565e-01	2.863	0.004195	**
`Prop_Zone_3:taux_zonier`	-1.589e+00	4.475e-01	-3.551	0.000384	***
`SPEI_30j_printemps_moy:SPEI_180j_hiver_diff_amplitude`	5.700e+00	1.789e+00	3.186	0.001440	**
`taux_zonier:SPTI_30j_ete_diff_max`	-3.915e+01	9.324e+00	-4.199	2.69e-05	***
`Prop_Zone_3:SPTI_30j_hiver_diff_min`	-3.330e-02	1.371e-02	-2.430	0.015118	*
`latitude_moyen:SPI_90j_ete_diff_max`	2.153e-02	7.178e-03	3.000	0.002704	**
`taux_souscription_dpt:SPTI_90j_critere_sum_moyenne`	5.182e-01	2.123e-01	2.441	0.014643	*
`taux_zonier:SPTI_30j_printemps_moy`	-2.385e+01	1.112e+01	-2.145	0.031936	*

Figure 63 – Tableau, sous **R**, récapitulant les variables explicatives du modèle logit. Le modèle est composé de 28 variables explicatives dont 14 termes d'interaction.

9.4 Comparaison entre données climatiques modélisées et mesurées

Nous représentons dans la figure 64 (respectivement figure 65) les données de température maximale (respectivement minimale) mesurées en fonction des données de température maximale (respectivement minimale) modélisées pour les quatre stations météo qui sont les plus proches de leurs points de grille associés.

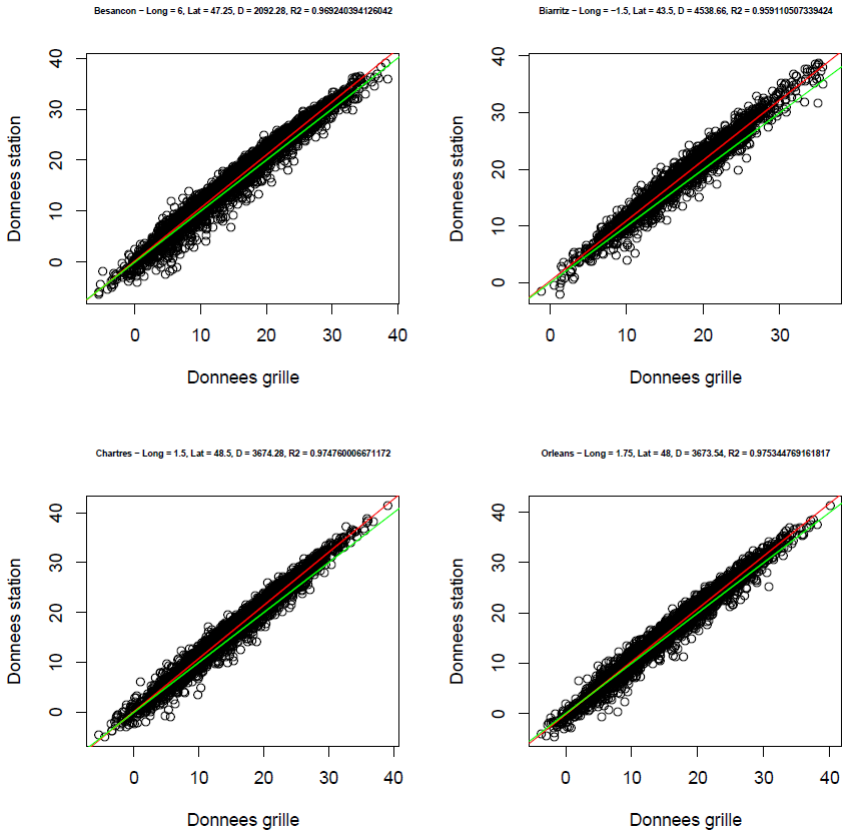

Figure 64 – La température maximale journalière de la station en fonction de la température maximale journalière modélisée. La droite verte correspond à la droite $y = x$ et la droite rouge correspond à la droite de régression. La distance entre la station et le point de la grille la plus proche est indiquée dans le titre en mètre accompagnée du nom de la station, de la position géographique de cette dernière et du R^2 de la régression linéaire. D'après ces graphes, les données de température maximale modélisées sont plutôt de bonne qualité

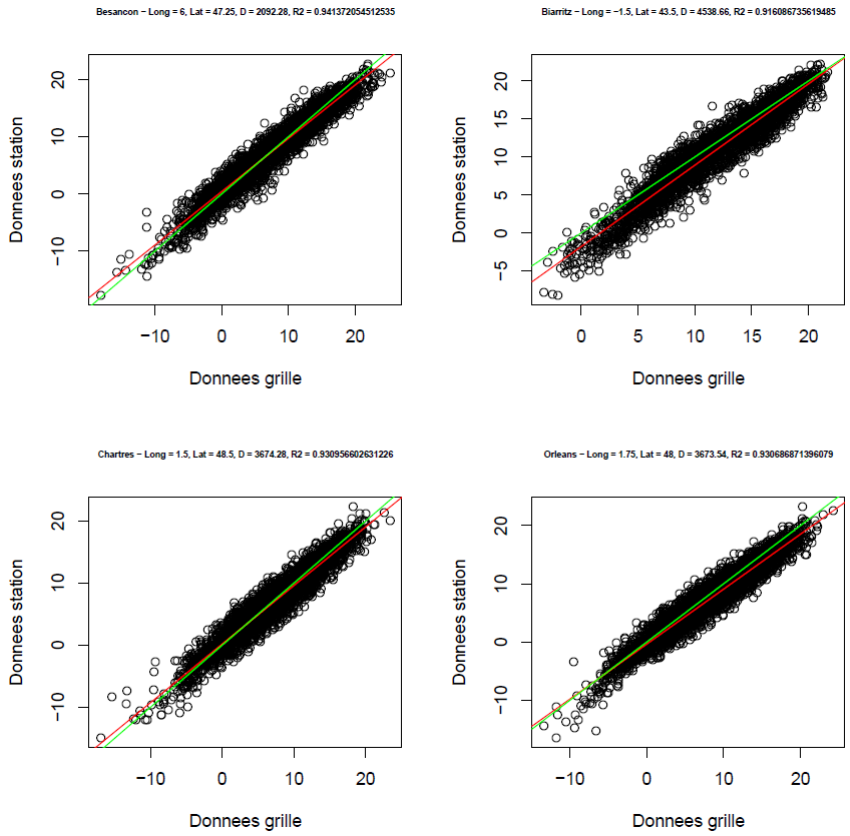

Figure 65 – La température minimale journalière de la station en fonction de la température minimale journalière modélisée. La droite verte correspond à la droite $y = x$ et la droite rouge correspond à la droite de régression. La distance entre la station et le point de la grille la plus proche est indiquée dans le titre en mètre accompagnée du nom de la station, de la position géographique de cette dernière et du R^2 de la régression linéaire. D'après ces graphes, les données de température minimale modélisées sont plutôt de bonne qualité.

Références

[1] Mission Risques Naturels, "Sécheresse Géotechnique - De la connaissance de l'aléa à l'analyse de l'endommagement du bâti", *rapport MRN* (2018).
URL : https://www.mrn.asso.fr/wp-content/uploads/2019/01/21-01-2018_rapport-mrn_secheresse-2018.pdf

[2] J. F. Schulte, "Modélisation du risque subsidence en France métropolitaine", *Mémoire, Institut des Actuaires* (2016).
URL : http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/

[3] Géorisque, "Base de données - Retrait / gonflement des argiles", *site visité le 16-09-2020.*
URL : https://www.georisques.gouv.fr/donnees/bases-de-donnees/

[4] E. Arnaud, "Modélisation du risque sécheresse en France", *Mémoire, Institut des Actuaires* (2016).
URL : http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/

[5] M. Hayes, M. Svoboda, N. Wall, and M. Widhalm, "The Lincoln Declaration on Drought Indices : Universal Meteorological Drought Index Recommended", *Bulletin of the American Meteorological Society 92(4), 485-488* (2016).
DOI : 10.1175/2010BAMS3103.1

[6] Z. Ali, I. Hussain, M. Faisal et al., "A Novel Multi-Scalar Drought Index for Monitoring Drought : the Standardized Precipitation Temperature Index", *Water Resour Manage 31, 4957–4969* (2017).
DOI : 10.1007/s11269-017-1788-1

[7] H. Wu, M. J. Hayes, A. Weiss and Q. Hu, "An evaluation of the Standardized Precipitation Index, the China-Z Index and the statistical Z-Score", *International Journal of Climatology 21, 745–758* (2001).
DOI : 10.1002/joc.658

[8] S. Morid, V. Smakhtin and M. Moghaddasi, "Comparison of seven meterological indices for drought monitoring in Iran", *International Journal of Climatology 26, 971–985* (2006).
DOI : 10.1002/joc.1264

[9] H.-R. Byun, D. A. Wilhite, "Objective Quantification of Drought Severity and Duration", *International Journal of Climatology 12, 2747-2756* (1999).
URL : https://doi.org/10.1175/1520-0442(1999)012<2747:OQODSA>2.0.CO;2

[10] W. C. Palmer, "Meteorological Drought", Research Paper No. 45, US Weather Bureau, Washington, DC (1965)

[11] T. B. McKee, N.J. Doesken and J. Kleist, "The Relationship of Drought Frequency and Duration to Time Scales", *Proceedings of the 8th Conference on Applied Climatology, 17-22 Jan. 1993, Anaheim, California.*
URL : https://www.droughtmanagement.info/literature/AMS_Relationship_Drought_Frequency_Duration_Time_Scales_1993.pdf

[12] D. C. Edwards, T. B. McKee, "Characteristics of 20th century drought in the United States at multiple time scales", *Colorado State University, Department of Atmospheric Science.* (1997)

[13] World Meteorological Organization, *Standardized Precipitation Index User Guide*, WMO-No. 1090, World Meteorological Organization, Geneva, Switzerland (2012).

[14] S. M. Vicente-Serrano, S. Begueria and J. I. Lopez-Moreno, "A multi-scalar drought index sensitive to global warming : the Standardized Precipitation Evapotranspiration Index", *International Journal of Climatology 23, 1696-1718.* (2010)
DOI : 10.1175/2009JCLI2909.1

[15] S. Begueria, S. M. Vicente-Serrano, F. Reig and B. Latorre, "Standardized precipitation evapotranspiration index (SPEI) revisited : parameter fitting, evapotranspiration models, tools, datasets and drought monitoring", *International Journal of Climatology 34, 3001–3023.* (2014)
DOI : 10.1002/joc.3887

[16] G. H. Hargreaves, "Defining and using reference evapotranspiration", *Journal of Irrigation and Drainage Engineering*, 120, 1132–1139 (1994)

[17] E. Balthas, *Spatial distribution of climatic indices in northern Greece* (2007).
DOI : 10.1002/MET.7

[18] E. De Martonne, *Traité de Géographie Physique*, Quatrième édition. A. Colin, Paris. (1925)

[19] WMO - GWP Integrated Drought management Programme, *Handbook on Drought Indicators and Indices*, WMO No. 1173. WMO, Geneva, Switzerland and GWP, Stockholm, Sweden (2016).
URL : https://www.droughtmanagement.info/indices/

[20] A. L. Barbu1 et al., *Assimilation of Soil Wetness Index and Leaf Area Index into the ISBA-A-gs land surface model : grassland case study* (2011).
DOI : 10.5194/bg-8-1971-2011.

[21] B. Narasimhan, R. Srinivasan, *Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring* (2005).
DOI : 10.1016/j.agrformet.2005.07.012.

[22] B. Lyon, A. G. Barnston, *ENSO and the Spatial Extent of Interannual Precipitation Extremes in Tropical Land Areas* (2005).
DOI : 10.1175/JCLI3598.1

[23] G. Tsakiris, D. Pangalou, H. Vangelis, *Regional Drought Assessment Based on the Reconnaissance Drought Index (RDI)* (2007).
DOI : 10.1007/s11269-006-9105-4

[24] Z. Balint, F. Mutua, P. Muchiri and C. T. Omuto, *Monitoring Drought with the Combined Drought Index in Kenya* (2013).
DOI : 10.1016/B978-0-444-59559-1.00023-2

[25] Ministère de l'économie, des finances et de la relance, "Demandes de valeurs foncières DVF" (2021).
URL : https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/

[26] INSEE, "Logements et résidences principales en 2017", *Recensement de la population* (2020).
URL : https://www.insee.fr/fr/statistiques/4515532?sommaire=4516107&q=recensement#dictionnaire/

[27] V. Vallès, "Le dynamisme démographique faiblit entre 2013 et 2018, avec la dégradation du solde naturel", *INSEE - Statistiques et études* (2020).
URL : https://www.insee.fr/fr/statistiques/4999744#consulter

[28] C. Arnold, "37 millions de logements en France au 1er janvier 2020", *INSEE - Statistiques et études* (2020).
URL : https://www.insee.fr/fr/statistiques/4985385

[29] U. Neumann, N. Genze et D. Heider, "EFS : an ensemble feature selection tool implemented as R-package and web-application", BioData Mining vol. 10, Article 21 (2017).

[30] L. Yu, H. Liu, "Efficient feature selection via analysis of relevance and redundancy", *Journal of Machine Learning Research*, 5, 1205–24 (2004)

[31] T. Hastie, T. Robert and J. H. Friedman, *The Elements of Statistical Learning*, Vol.1, page 339, Springer, New York (2001).

[32] C. Chen, C. Guestrin, "XGBoost : A Scalable Tree Boosting System", arXiv :1603.02754 (2016).
URL : https://arxiv.org/abs/1603.02754

[33] L. Mason, J. Baxter, P. Bartlett, M. Frean, "Boosting Algorithms as Gradient Descent", NIPS Conference, Denver, Colorado, USA, Nov. 29 - Dec. 4 (1999)
URL : https://www.researchgate.net/publication/221618845_Boosting_Algorithms_as_Gradient_Descent

[34] P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees", Machine Learning, 63(1), 3-42 (2006).

[35] C. Dowd, "A new ECDF Two-Sample Test Statistic", arXiv :2007.01360 (2020)

[36] B. K. Miron,R. R. Witold, "Feature Selection with the Boruta Package", Journal of Statistical Software, 36(11) (2010).
URL : http://www.jstatsoft.org/v36/i11/

[37] W. N. Venables and B. D. Ripley, "Modern Applied Statistics with S", Fourth edition, Springer. (2002).

[38] Météo France, Données mensuelles d'indice d'humidité des sols pour le dispositif CatNat (2021).
URL : https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=301&id_rubrique=40

[39] Ministère de l'intérieur, "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", *Circulaire du 10 mai 2019*, n° INTE1911312C. URL : http://circulaire.legifrance.gouv.fr/pdf/2019/05/cir_44648.pdf

[40] J.-M. Soubeyroux1, J.-P. Vidal , J. Najac, N. Kitova, M. Blanchard, P. Dandin, E. Martin, C. Pagé, F. Habets, "Impact du changement climatique en France sur la sécheresse et l'eau du sol", *Projet ClimSec - Rapport final du projet* (2011)

[41] Géorisque, "Exposition du territoire au phénomène", site visité le 5 juin 2021. URL : https://www.georisques.gouv.fr/articles-risques/exposition-du-territoire-au-phenomene

[42] A. Charpentier, M. James, H. Ali, "Predicting Drought and Subsidence Risks in France", arXiv :2107.07668 (2021). URL : https://arxiv.org/abs/2107.07668

[43] Caisse Centrale de la Réassurance, "Bilan des Catastrophes Naturelles". URL : https://geoportail.ccr.fr/portal/apps/sites/#/bilancatnat/pages/telechargements

[44] Wikipedia, "Variogram" (2021). URL : https://en.wikipedia.org/wiki/Variogram