Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 9 Septembre 2022

Par : JAMET Guillaume
Titre : Estimation déterministe de la charge ultime pour le risque de retrait-gonflement des argiles en France métropolitaine

Confidentialité : Oui 2 ans

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

Membres présents du jury de l'Institut des Actuaires :
GUILLEMOT David
SENGDY Davy
Signature :

Entreprise :
GENERALI France
Signature :

Membres présents du jury de l'EURIA :
BOIVIN Daniel

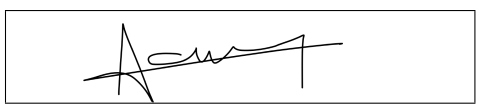
Directeur de mémoire en entreprise :
BRETTSCHNEIDER Marco
Signature :

Invité :

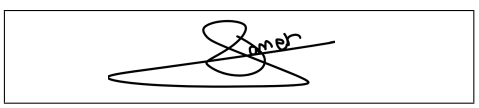
Signature :

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

Signature du responsable entreprise :



Signature du candidat :



EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Résumé

L'évaluation des risques naturels est un enjeu majeur et croissant pour les assureurs. Avec pour cause le dérèglement climatique, les catastrophes naturelles seront amenées à se produire plus fréquemment et avec plus d'intensité. Depuis 2016, la France subit une recrudescence des épisodes de sécheresse à tel point que cet aléa semble devenir certain.

Ce mémoire propose une démarche particulière afin d'évaluer les risques encourus par le phénomène de retrait-gonflement des argiles sur le portefeuille multirisques habitation (MRH) de Generali.

Après avoir contextualisé et définit précisément l'origine du risque et ses conséquences, nous avons construit plusieurs indices de sécheresse caractérisant l'intensité de l'évènement observé. En plus de ces indices, des critères météorologiques ont été établis pour caractériser l'éligibilité du phénomène au régime d'indemnisation des catastrophes naturelles (Cat Nat).

Le croisement de ces informations météorologiques avec les indices d'exposition des bâtiments à l'aléa ainsi que l'historique des demandes de reconnaissance communales permettent d'identifier avec un certain degré de précision les communes impactées et sinistrées.

Compte tenu de l'exposition de notre portefeuille sur le territoire ainsi que les données risques qu'il contient, nous avons construit un modèle permettant d'estimer le nombre de sinistres au sein des communes identifiées comme reconnues au dispositif Cat Nat.

La multiplication de ces deux modèles avec le coût moyen du risque permet d'obtenir une estimation robuste de la charge à l'ultime sur ce péril.

Mots clefs: Retrait-gonflement des argiles, RGA, Cat Nat, Catastrophes naturelles, Sécheresse, IBNR, Charge ultime, Modèles linéaires généralisés, GLM

Abstract

The evaluation of natural risks is a major and growing challenge for insurers. With climate change, natural disasters will occur more frequently and with greater intensity. Since 2016, France has experienced an increase in drought episodes to the point where this hazard seems to become certain.

This thesis proposes a specific approach to evaluate the risks incurred by the phenomenon of clay shrinkage and swelling on the multi-risk housing portfolio of Generali.

After having contextualized and precisely defined the origin of the risk and its consequences, we have built several drought indices characterizing the intensity of the observed phenomenon. In addition to these indices, meteorological criteria were established to characterize the eligibility of the phenomenon to the Cat Nat system.

This meteorological information combined with building exposure indices as well as the history of the communal recognitions make it possible to identify with a certain degree of precision the impacted and damaged municipality.

Multiplying these two models and the average cost of risk yields a robust estimate of the ultimate charge of this peril.

Keywords: Subsidence risk, Natural disaster, Drought, IBNR, Ultimate cost, Generalized Linear Model, GLM

Note de synthèse

Contexte et problématique

Lors de l'épisode estival de 2003, la Caisse Centrale de Réassurance (CCR) a évalué le coût global des sinistres liés au retrait-gonflement des argiles (RGA) à plus d'un milliard d'euros pour la France métropolitaine. Si cet évènement de grande ampleur a été exceptionnel, il pourrait être amené à se reproduire une fois tous les 3 ans d'ici 2050 en prenant le scénario le plus pessimiste du Groupe d'experts Intergouvernementale sur l'Evolution du Climat (GIEC). Le réchauffement climatique, l'inflation ainsi que l'augmentation des valeurs assurées accentuent la nécessité des assureurs à évaluer les risques encourus.

Les sinistres liés au RGA ont la particularité d'être caractérisés par une cinétique lente. Les effets dommageables à l'évènement naturel sont dans la plupart des cas connus bien après ce dernier, ce qui fait du RGA une branche longue en termes d'écoulement des sinistres. Ainsi, l'estimation de la charge ultime sur ce péril est un besoin crucial pour pouvoir déduire le montant associé aux sinistres survenus mais pas encore déclarés (IBNyR) .

Cependant, le caractère imprévisible et volatile de la sécheresse ainsi que les changements de réglementation intempestifs vis-à-vis des critères de reconnaissance au dispositif Cat Nat rendent les méthodes de provisionnement usuelles inapplicables.

Ce mémoire d'actuariat a pour objectif de déterminer la charge ultime liée au risque de subsidence sur le portefeuille de GENERALI.

Présentation de la démarche

Pour ce faire, nous avons choisi de comparer les résultats de deux méthodes pour estimer le nombre de sinistres liés au RGA :

- La première méthode repose sur un modèle linéaire généralisé (GLM) de type Poisson. Elle a comme objectif d'estimer directement le nombre de sinistres en fonction des conditions météorologiques observées, de l'exposition de notre portefeuille à l'aléa ainsi que des données risques de nos contrats.
- La seconde méthode se décompose en deux sous-parties :
  - Un modèle de classification dont l'objectif est de déterminer les communes et saisons qui vont être reconnues en l'état de catastrophe naturelle sécheresse. Pour cela le modèle se base uniquement sur des données provenant de l'*open data*.
  - Un modèle GLM de type Poisson qui, compte-tenu de cette reconnaissance, détermine le nombre de sinistres à survenir.

Avec le peu de sinistres observés, nous avons dû limiter le modèle de sévérité à l'utilisation d'un coût moyen. La charge ultime prédite correspondra donc à la multiplication des sinistres prédits par ces approches avec ce coût moyen.

Assemblage des données

Avant de construire nos modèles, nous avons du créer plusieurs indices de sécheresse afin de quantifier l'intensité de l'aléa. A partir des données météorologiques observées ou modélisées par le centre de prévisions européen à moyen terme (ECMWF), nous avons avons pu construire une collection de cartes d'indices à partir de mars 1950 jusqu'à la fin de l'année 2021. Ces indices standardisés sont relatifs aux déficits de précipitations, aux précipitations nettes, aux anomalies de températures ou encore aux déficits hydriques des sols. La figure 1 donne un aperçu des précipitations nettes standardisées au cours des saisons de l'année 2018.

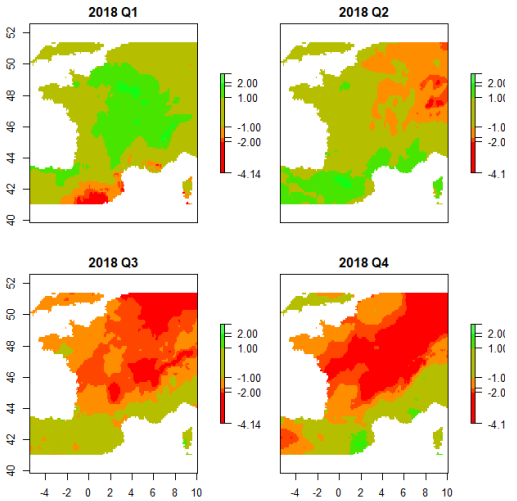
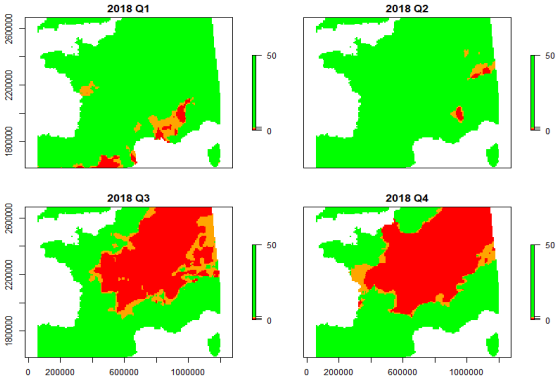
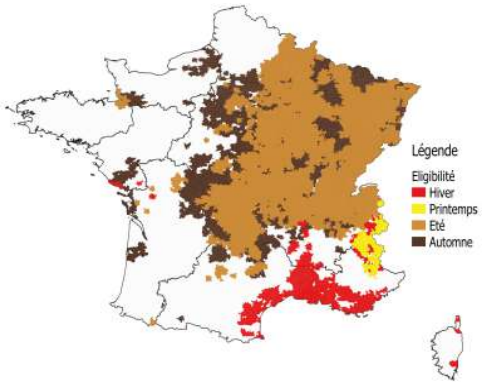


FIGURE 1 – Indice $ESPEI_3$ sur l'année 2018

Comme l'indemnisation des sinistres est conditionnée à la publication d'un arrêté Cat Nat sécheresse favorable au journal officiel, nous avons récréé les critères utilisés par la commission interministérielle à partir de l'humidité des sols modélisée par l'ECMWF. Cette approximation quantifie l'éligibilité potentielle des communes au régime d'indemnisation des catastrophes naturelles.



(a) Reproduction des critères météorologique (0-28cm)



(b) Critère météorologique uniforme

FIGURE 2 – Comparaison des critères météorologiques en 2018

Après avoir collecté ces informations météorologiques, nous avons recueilli plusieurs

indices relatifs à la présence d'argile dans les sols. Nous avons utilisé la carte d'exposition du BRGM pour obtenir la part de la surface communale ainsi que la superficie représentée dans chacune des zones d'aléa. La carte de l'*European Soil Data Centre* (ESDAC) nous permet d'obtenir une information complémentaire en retenant la concentration moyenne en argile pour chacune des communes. Enfin, nous avons utilisé les indicateurs de vulnérabilité du Service des Données et Études Statistiques (SDES) qui superpose la carte d'exposition à l'aléa avec la localisation des maisons individuelles en France métropolitaine.

Aux indices météorologiques et géologiques se sont également ajoutées d'autres variables comme le nombre de demandes de reconnaissance antérieures ou encore le nombre de jours écoulés depuis la dernière demande connue. Nous verrons que ces variables jouent un rôle déterminant dans nos modèles.

Mise en application : 1re méthode

Avant de procéder à la modélisation de la subsidence, nous avons dû faire plusieurs hypothèses. Avec la faible volumétrie de sinistres dans notre base ainsi que les taux de clotûre relativement bas pour les années récentes, nous avons fait le choix d'intégrer les sinistres ouverts dans la modélisation de la fréquence.

Étape	Variable	Pré simplification			Post simplification		
		AIC	BIC	Déviance	AIC	BIC	Déviance
Étape 1	Critère météorologique (0-28cm)	-3 107	- 2390	-3147	-3099	-3056	-3103
Étape 2	Nombre de demandes	-1882	-1509	-1903	-1888	-1845	-1893

TABLE 1 – Impact du nombre de demandes antérieures et des critères météorologiques sur les métriques

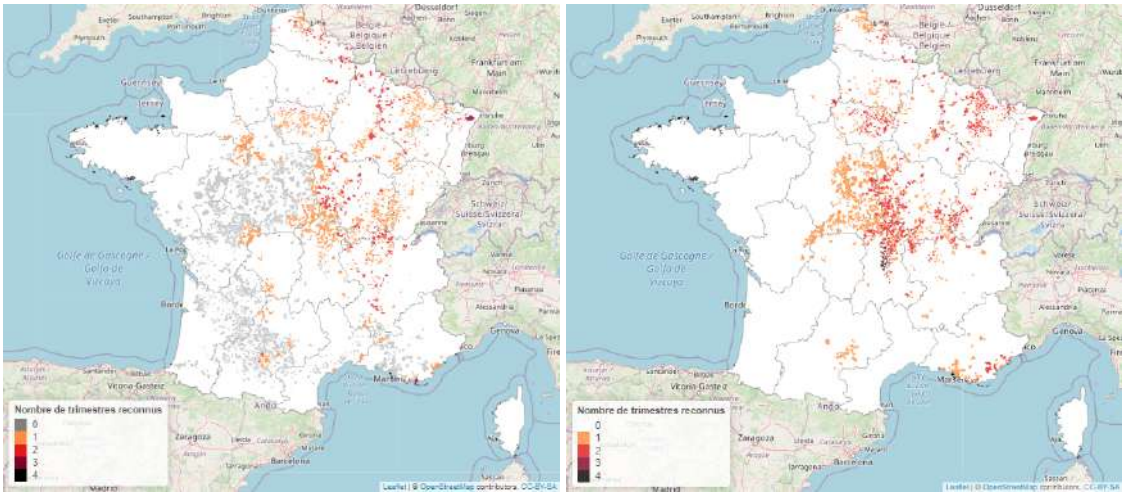
Au cours de la sélection de variables, ces attributs se sont révélés très discriminants. La reproduction des critères sur la couche superficielle du sol caractérise l'éligibilité potentielle de la commune au dispositif Cat Nat tandis que l'historique des demandes communales permet de déterminer, parmi les communes potentiellement éligibles, celles qui vont effectuer les démarches administratives.

Mise en application : 2e méthode

Le modèle de détection

Après avoir tenté, en vain, d'améliorer les performances de notre modèle avec différentes techniques de rééchantillonnage, nous avons choisi de restreindre l'apprentissage

de nos données sur une base aux critères de reconnaissance homogènes (2018 et 2019) et d'appliquer un seuil de classification maximisant la métrique de F1-score sur les années de test.



(a) Arrêtés observés - 2020 (b) Prévisions - 2020

FIGURE 3 – Comparaison entre observés et prédits sur l'année 2020

Même si le modèle de classification ne détecte pas toujours les bonnes communes, les résultats sont encourageants sur les années futures car nous observons une cohérence spatiale avec la variable cible observée.

Le modèle de fréquence post-détection

Après avoir identifié les communes qui vont être reconnues en l'état de catastrophe naturelle sécheresse, nous devons estimer le nombre de sinistres survenus sachant que la commune a été impactée.

Nous avons donc filtré nos données sur les communes et saisons reconnues en l'état de catastrophe naturelle sécheresse et nous avons construit un modèle comptage de type Poisson. Dès lors, la valeur prise par les indices de sécheresse se résume à leurs queues de distribution. L'unique variable météorologique intégrée à notre modèle concerne les précipitations nettes standardisées.

Comparaison des résultats

Après avoir vérifié les corrélations de nos variables, la stabilité des coefficients ainsi que leurs significativités, nous avons souhaité comparer simultanément la première et la seconde approche. Pour cela, nous avons appliqué le modèle de fréquence post-détection

à l'ensemble de nos données. Puis nous avons multiplié les prévisions de ce dernier par les résultats de notre classification. Nous avons également fait le choix de multiplier la fréquence prédite par la probabilité que la commune soit concernée par un arrêté Cat Nat. Cette probabilité correspond aux prédictions du modèle de détection avant d'avoir appliqué le seuil de classification.

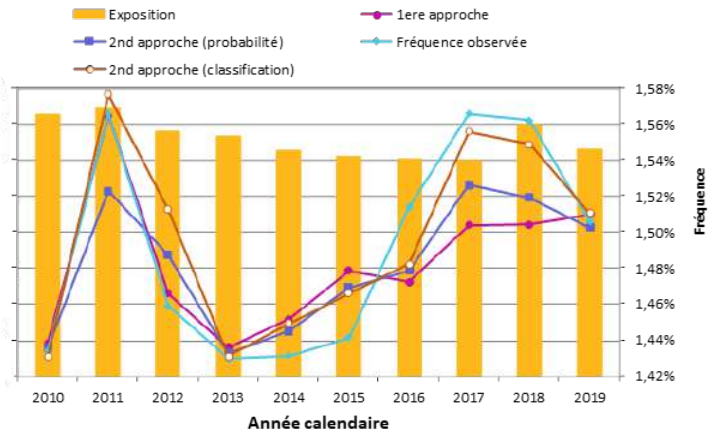


FIGURE 4 – Fréquence moyenne sur la base holdout

En comparant les résultats de nos prédictions sur la base de validation, nous avons conclu que la seconde approche, utilisant la classification du modèle de détection, était la meilleure en moyenne mais aussi la plus prudente. Les premières années, la 1re approche est relativement proche de la fréquence observée puis l'écart entre les deux courbes s'accentue. Finalement, la première approche sous-prédit la fréquence au cours des récents épisodes de sécheresse tandis que la 2e approche utilisant la classification ajuste mieux la fréquence observée.

Ensuite, nous avons multiplié les sinistres prédits par ces approches avec le coût moyen des sinistres clos sur la période 2011-2015. L'application de ce coût moyen a augmenté l'écart avec les données observées pour les années 2016 et 2017. A partir de 2018, les taux de clôture se sont révélés très faibles. Par conséquent, les prévisions de nos modèles à l'ultime anticipent la réévaluation des dossiers en-cours.

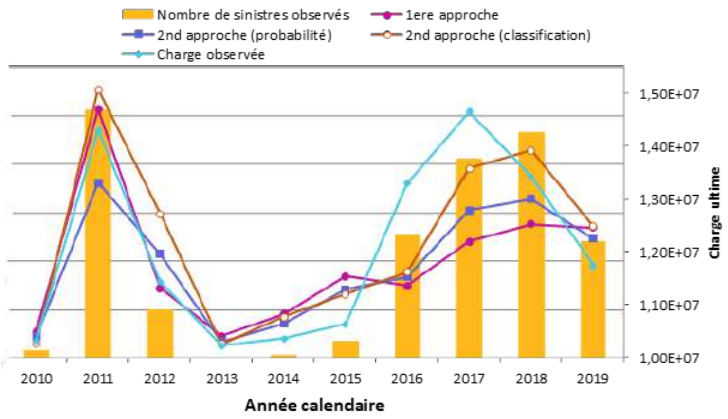


FIGURE 5 – Charge totale sur la base holdout

Limites de la modélisation

La principale limite de notre modélisation est l'interprétabilité du modèle de détection. Comme les forêts aléatoires sont des méthodes d'apprentissage non paramétriques, il est difficile de comprendre d'où proviennent les erreurs faites par le modèle ainsi que l'ensemble des décisions réalisées par celui-ci.

Lors de notre étude nous n'avons pas pris en considération le nombre de non reconnaissances successives. Lorsqu'une commune est sinistrée mais pas éligible à la reconnaissance, les dégâts subis par la sécheresse s'aggravent au cours du temps. Comme il est difficile pour les experts mandatés de dater les dégâts constatés, le coût des sinistres risque d'être plus conséquent lors de la prochaine reconnaissance.

Enfin, l'utilisation d'un coût moyen est une limite importante sur l'estimation de la charge ultime, certains dégâts comme les reprises en sous-oeuvre sont bien plus coûteux que d'autres réparations et nous aurions aimé discriminer le montant des sinistres.

Ouverture

Au sein de nos modèles, les quelques variables de risques sélectionnées sont peu discriminantes. Or nous avons connaissance de certains facteurs aggravant comme la présence d'arbres à proximité de l'habitation, le niveau de pentification sur la parcelle ou encore la forme géométrique du bâtiment. L'utilisation de l'imagerie satellite apparaît comme le principal levier d'actions pour récolter ces indicateurs.

Enfin, nous aurions pu comparer nos résultats avec d'autres méthodes comme les modèles de régression à inflation de zéro (ZIP ou ZINB). Ces modèles offrent une alternative à notre schéma de modélisation car ils combinent la régression logistique ainsi que le modèle de comptage en un unique modèle.

Synthesis note

Context and issues

During the summer episode of 2003, the french reinsurer called *Caisse Centrale de Réassurance* (CCR) estimated the overall cost of losses related to subsidence at more than one billion euros for metropolitan France. Although this large-scale event was exceptional, it could happen again once every three years by 2050, according to the most pessimistic scenario of the Intergovernmental Panel on Climate Change (IPCC). Global warming, inflation and rising insured values increase the need for insurers to assess the risks involved.

Claims related to subsidence are characterised by slow kinetics. The damaging effects of the natural event are in most cases known well after the event, the ultimate view of the losses incurred will only be known after a few years.

Thus, estimating the ultimate cost on this peril is a crucial need to be able to deduct the amount associated with claims that have occurred but not yet been reported (IBNyR).

However, the unpredictable and volatile nature of drought, as well as the frequent changes in regulations regarding the recognition criteria under the Cat Nat scheme, make the usual provisioning methods inapplicable.

The objective of this actuarial report is to determine the ultimate cost of subsidence risk on the GENERALI portfolio.

Approach

To do this, we choose to compare the results of two methods for estimating the number of subsidence-related claims :

- The first method is based on a generalized linear model (GLM) with Poisson errors. Its objective is to directly estimate the number of claims as a function of observed weather conditions, the exposure of our portfolio to the hazard and the risk data of our contracts.
- The second method is divided into two sub-sections :
  - A classification model whose objective is to determine the municipalities and seasons that will be recognised as being in a state of natural disaster due to drought. The model is based on open data.
  - A GLM model with Poisson errors which, given this recognition, determines the number of claims to occur.

With few observed claims, we had to limit the severity model to an average cost. The predicted ultimate cost will therefore be the multiplication of the number of claims predicted by these approaches with this average cost.

Database

Before building our models, we had to create several drought indices to quantify the intensity of the hazard. Using observed or modelled meteorological data from the European Centre for Medium-Range Forecasting (ECMWF), we were able to construct a collection of index maps from March 1950 to the end of 2021. These standardized indice relate to precipitation deficits, net precipitation, temperature anomalies and soil moisture deficits. Figure 6 gives an overview of the standardized precipitation evapotranspiration index during the seasons of 2018.

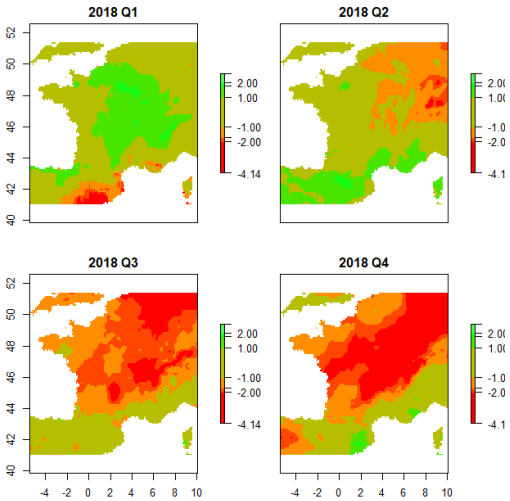
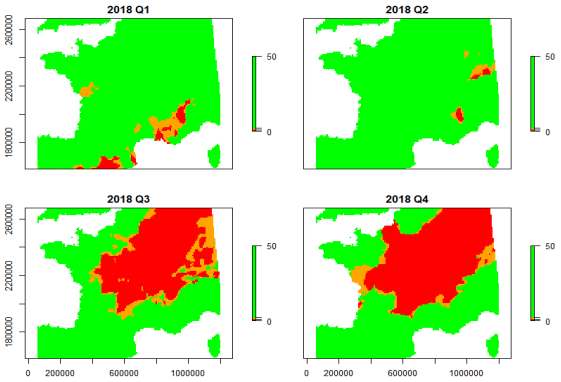
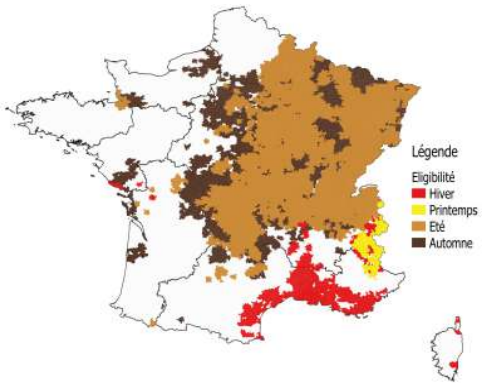


FIGURE 6 – $ESPEI_3$ Index in 2018

As the compensation of claims is conditioned by the publication of a favourable drought Cat Nat decree in the so called *journal officiel* (JO), we have recreated the criteria used by the interministerial commission on the basis of the soil moisture modelled by the ECMWF. This approximation quantifies the potential eligibility of municipalities for the natural disaster compensation scheme.



(a) Replication of meteorological criteria (0-28cm)



(b) Uniform Criteria

FIGURE 7 – Comparison between meteorological criteria in 2018

After having collected this meteorological information, we collected several measures relating to the presence of clay in the soils. We used the exposure map to obtain the share of the community area as well as the area represented in each of the hazard zones. The

map of the European Soil Data Centre (ESDAC) allows us to obtain additional information by retaining the average clay concentration for each community. Finally, we used the vulnerability indicators of the *Service des Données et Études Statistiques* (SDES) which superimposes the map of exposure to the hazard with the location of individual houses in metropolitan France.

In addition to the meteorological and geological index, other variables were also added, such as the number of previous requests for recognition or the number of days since the last known request. We will see that these variables play a determining role in our models.

Implementation : 1^{st} method

Before proceeding with the subsidence modelling, we had to make several assumptions. With the low volume of claims in our database as well as the relatively low closure rates for recent years, we made the choice to integrate open claims in the frequency modelling.

Step	Features	Pré simplification			Post simplification		
		AIC	BIC	Deviance	AIC	BIC	Deviance
Step 1	Wheather Criteria (0-28cm)	-3 107	- 2390	-3147	-3099	-3056	-3103
Step 2	Number of previous applications	-1882	-1509	-1903	-1888	-1845	-1893

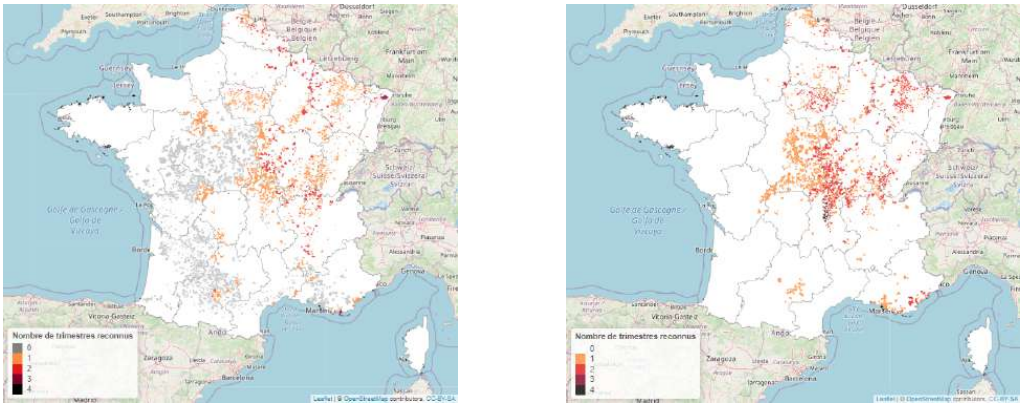
TABLE 2 – Impact of the number of previous applications and weather criteria on the metrics

During the feature selection, these attributes proved to be very discriminating : the reproduction of criteria on the surface layer of the soil characterises the potential eligibility of the municipality for the Cat Nat system, while the history of communal requests makes it possible to determine, among the potentially eligible municipality, those which will carry out the administrative procedures.

Implementation : 2^{nd} method

The detection model

After unsuccessful attempts to improve the performance of our model with different resampling techniques, we chose to restrict the learning of our data to a database with homogeneous recognition criteria (2018 and 2019) and to apply a classification threshold maximising the F1-score metric over the test years.



(a) Observed decree - 2020 (b) Predicted decree - 2020

FIGURE 8 – Comparison between observed and predicted for the year 2020

Even if the classification model does not always detect the right municipality, the results are encouraging for future years as we observe a spatial consistency with the observed target feature.

The post-detection frequency model

After having identified the municipality that are going to be recognised in the state of natural drought disaster, we have to estimate the number of claims that have occurred knowing that the community has been impacted.

We have therefore filtered our data on the municipality and seasons recognised as being in a state of natural disaster due to drought and we have built a counting model with poisson errors. From then, the value taken by the drought index is reduced to their distribution tails. The only meteorological feature integrated in our model concerns the standardized precipitation evapotranspiration index.

Comparison of results

After checking the correlations of our features, the stability of the parameters as well as their significances, we wished to compare simultaneously the first and the second approach. To do this, we applied the post-detection frequency model to our data set. Then we multiplied the predictions by the results of our classification. We also chose to multiply the predicted frequency by the probability that the municipality is concerned by a Cat Nat decree. This probability corresponds to the predictions of the detection model before applying the classification threshold.

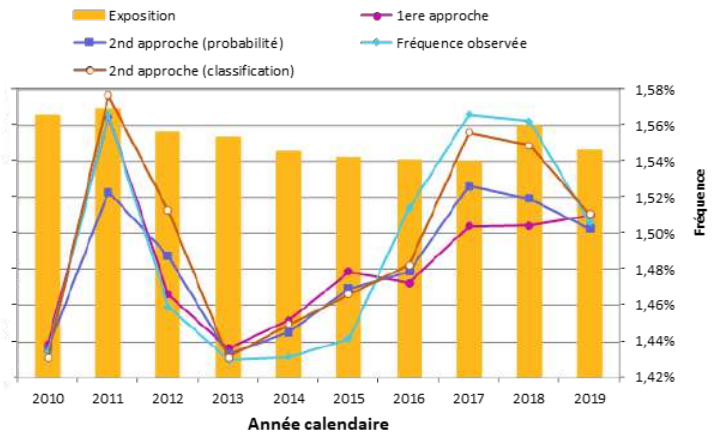


FIGURE 9 – Average frequency on the holdout data set

By comparing the results of our predictions with the holdout data set, we concluded that the second approach, using the classification of the detection model, was the best on average but also the most prudent. In the first few years, the first approach is relatively close to the observed frequency and then the gap between the two curves increases. Finally, the first approach underpredicts the frequency during recent droughts, while the second approach using classification better fits the observed frequency.

Next, we multiplied the claims predicted by these approaches with the average cost of closed claims over the period 2011-2015. Applying this average cost increased the variance with the observed data for the years 2016 and 2017. From 2018, the closure rates were found very low. As a result, our models' ultimate forecasts anticipate the revaluation of outstanding cases.

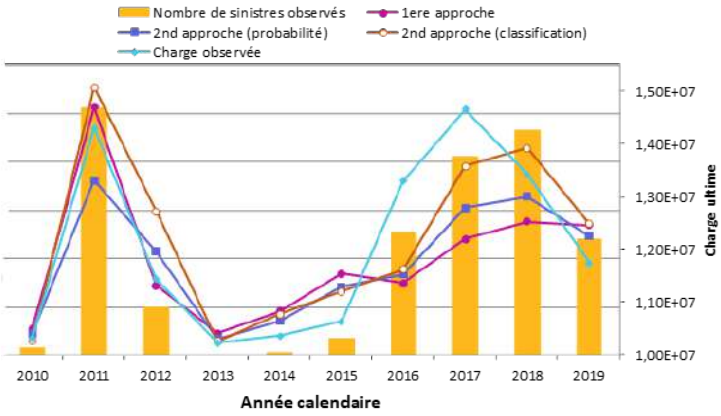


FIGURE 10 – Total cost of claims on the holdout data set

Limitations of the modeling

The main limitation of our modelling is the interpretability of the detection model. As random forests are non-parametric learning methods, it is difficult to understand where the errors made by the model come from and the set of decisions made by the model.

During our study we did not take into account the number of successive non-recognitions. When a municipality is affected but not eligible for recognition, the drought damage worsens over time. As it is difficult for the experts commissioned to date the damage observed, the cost of the claims is likely to be more substantial at the next recognition.

Finally, the use of an average cost is an huge limitation on the estimation of the ultimate cost, some damages such as house foundations refund are much more expensive than other repairs and we would have liked to discriminate the amount of the claims.

Opening

In our models, the few selected risk features are not very discriminating. However, we are aware of certain aggravating factors such as the presence of trees near the house, the field slope or the geometric shape of the building. The use of satellite imagery appears to be the main lever for collecting these indicators.

Finally, we could have compared our results with other methods such as the Zero Inflated Poisson regression models (ZIP) . These models offer an alternative to our modelling scheme as they combine the logistic regression as well as the counting model into a single model.

Remerciements

Je tiens à remercier en particulier mon tuteur d'alternance, Marco BRETTSCHEIDER, qui m'a suivi pendant plus d'un an en entreprise tant sur certaines missions que pour mon mémoire de fin d'étude. Sa disponibilité et ses précieux conseils ont été indispensables à la bonne réalisation de ce mémoire.

Je tiens à exprimer ma gratitude envers ma tutrice de l'institut des actuaires et collaboratrice, Sonia GUELOU, pour son implication sur ce projet. Je la remercie également pour ses précieux conseils sur la rédaction ainsi que pour la relecture de ce mémoire.

Je tiens à remercier mon tuteur académique, Pierre AILLIOT, qui m'a suivi durant cette alternance, ce mémoire mais également lors d'autres projets académiques. Ses conseils ainsi que les formations dispensées par Mr AILLIOT m'ont été d'une grande utilité pour la réalisation de ce mémoire. De la même manière je remercie Franck VERMET, directeur de l'EURIA, ainsi que le corps d'enseignants de l'EURIA pour la qualité de la formation.

Je tiens également à remercier ma manager, Amandine GOMEZ, pour son encadrement et son esprit critique qui m'a permis tant sur les missions que pour le mémoire à améliorer mes compétences ainsi que ma connaissance actuarielle.

J'adresse également mes remerciements à Yassine LAGHZALI pour ses connaissances, ses conseils sur la modélisation du péril et l'utilisation des logiciels dédiés. Par la même occasion, je tiens à remercier Maxime GATEAU qui m'a aidé en amont de la modélisation sur divers aspects techniques.

In fine, je tiens à remercier celles et ceux qui ont participé de près comme de loin à ce mémoire.

Avertissement

Pour des raisons de confidentialité, certains résultats ont été modifiés. Les ordres de grandeurs sont restés identiques de sorte à ne pas modifier les conclusions originales.

Table des matières

Introduction

Le risque de retrait-gonflement des sols argileux (RGA) est un péril couvert en France métropolitaine via une extension de garantie obligatoire sur les contrats multirisques-habitations (MRH) et fait partie du régime d'indemnisation des catastrophes naturelles (Cat Nat) depuis 1989.

Depuis son intégration au régime Cat Nat, les critères de reconnaissance en l'état de catastrophe naturelle sécheresse, qui conditionne l'indemnisation des assurés, ont évolués. Ces changements de réglementation associés à la cinétique lente de la sécheresse et son interaction complexe entre les conditions hydrométéorologiques, la nature du sol ainsi que le bâti fait de ce phénomène physique, encore aujourd'hui, un risque peu modélisé et appréhendé par les assureurs.

Pourtant, au cours des prochaines décennies, les catastrophes naturelles seront amenées à s'intensifier tant en intensité qu'en fréquence avec pour cause le dérèglement climatique. Selon la FFA, le coût cumulé du risque RGA pour la période 2020-2050 est estimé à 43 milliards d'euros contre 13.8 milliards d'euros pour la période 1989-2019. Soit un coût du risque environ trois fois plus élevé avec le scénario le plus pessimiste du GIEC. La contribution du changement climatique serait responsable de cette hausse à hauteur de 25% selon une étude réalisée par la CCR, le reste étant dû à l'inflation, à l'augmentation des biens assurés ainsi qu'à la concentration des populations et bâtiments sur les zones à risques.

Aujourd'hui le RGA représente, à lui seul, plus d'un tiers de la sinistralité Cat Nat [1]. Sur les vingts évènements naturels les plus coûteux, onze concernent la sécheresse ce qui place ce risque à la seconde place des dépenses Cat Nat derrière les inondations. L'évaluation des risques naturels représente donc un enjeu majeur et croissant pour les assureurs.

L'objectif de ce mémoire d'actuariat est d'estimer la charge ultime liée au RGA en France métropolitaine pour l'exercice courant afin d'obtenir une estimation plus robuste des sinistres survenus mais pas encore déclarés (IBNyR).

Afin de proposer une méthodologie pertinente et viable au fil du temps, il est nécessaire de comprendre l'origine du risque, ses conséquences ainsi que les facteurs aggravants et déclencheurs. La première partie de ce mémoire définit le cadre réglementaire et l'ensemble des informations utiles à la compréhension globale du sujet.

Après avoir contextualisé l'étude et défini l'approche de modélisation, nous présenterons les concepts théoriques utilisés lors de la modélisation du péril. Puis, nous présenterons et analyserons les différentes données utilisées au cours de l'étude.

1. Source : [CC, 2022]

Ensuite, nous parcourrons les résultats des différentes approches utilisées jusqu'à l'estimation de la charge ultime. Enfin, nous conclurons l'étude en sélectionnant l'approche la plus pertinente, nous exposerons également les limites ainsi que les pistes d'améliorations.

Chapitre 1

Le risque de subsidence en France métropolitaine

1.1 La sécheresse parmi les catastrophes naturelles

Au cours de cette section, nous évoquerons les différentes typologies de risques catastrophiques et en particulier celui du risque de retrait-gonflement des argiles (RGA) en France métropolitaine. Après avoir décrit le phénomène physique, l'exposition du territoire français au RGA ainsi que ses conséquences sur les bâtiments, nous nous attarderons sur un aspect essentiel de la réglementation : celui des critères de reconnaissance en l'état de catastrophe naturelle sécheresse. Sujets à d'innombrables critiques et contestations, les critères de reconnaissance impactent de manière directe le montant à la charge des assureurs et réassureurs mais également la modélisation du péril. Enfin, nous parcourrons certaines études sur le sujet afin de proposer une méthodologie pertinente pour estimer la charge ultime sur ce péril.

1.1.1 Définition des risques catastrophiques

Nous parlons de risque catastrophique lorsque les effets dommageables, aussi bien matériels que humains, résultent d'un phénomène brutal, durable ou intense qu'il soit d'origine naturelle ou causé par l'activité de l'homme. Il existe ainsi de nombreux risques catastrophiques aux origines ainsi qu'aux fréquences diverses et variées. Le tableau 1.1 donne un aperçu de ces risques et de leurs typologies.

Origine	Catégorie	Risques
Naturelles	Climatiques	Tempêtes, sécheresse, feux de forêts, cyclones
	Géologiques	Séismes, mouvement de terrain, éruptions volcaniques
	Biologiques	Pandémie, épizooties, panzooties
	Hydrologique	Inondation, coulée de boue
Anthropiques	Technologiques	Rupture d'un barrage, accidents nucléaires ou industriels ...
	Cyber-risques	Cyber-malveillance ou Cyber-terrorisme
	Guerres et émeutes	Révoltes populaires, conflits armés ...

TABLE 1.1 – Typologie des risques catastrophiques

Le graphique 1.1 montre l'évolution du nombre de catastrophes naturelles mondiales ainsi que le coût cumulé des évènements (en dizaine de millions de dollars) par année de survenance depuis 1950.

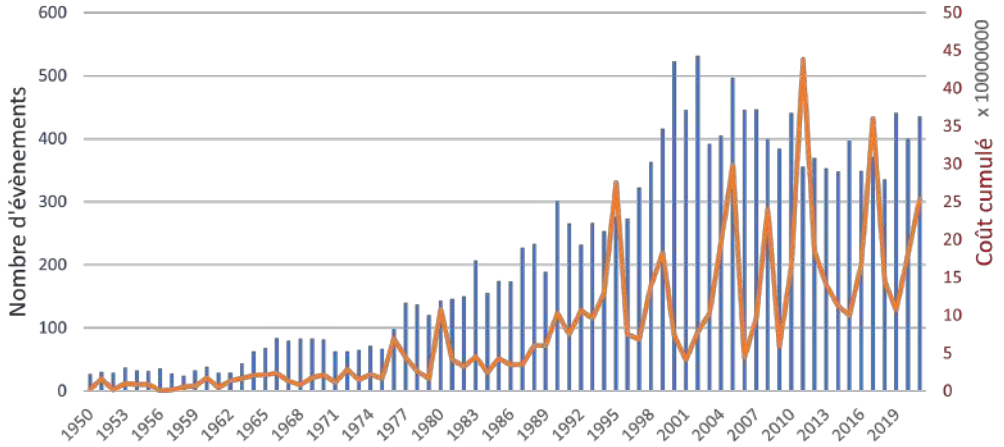


FIGURE 1.1 – Évolution du nombre de catastrophes naturelles dans le monde depuis 1950 et leur coût cumulé - SOURCE : EM-DAT

Avec le réchauffement climatique, la fréquence d'apparition des catastrophes naturelles (Cat Nat) a fortement augmenté. La concentration des populations sur des zones à risques ainsi que l'augmentation des biens assurés ont fortement contribué à l'augmentation des coûts engendrés par les Cat Nat. Bien que la sécheresse ne représente qu'une faible proportion des catastrophes mondiales d'origine naturelle (4.72% des catastrophes naturelles mondiales depuis 1950 selon l'EM-DAT), ces dernières années ont été marquées par des épisodes d'une rare intensité en France et dans le monde.

Pour citer quelques exemples :

- En 2021, la Californie a enregistré une sécheresse record : le manque de précipitations et de neige au cours de l'hiver associé aux températures anormalement élevés au cours de l'été (50°C au mois de juillet) ont eu des conséquences inédites sur l'agriculture. A cette sécheresse, s'est ajouté le deuxième feu de forêt le plus désastreux pour la Californie nommé « Dixie Fire » avec plus de 180000 hectares de terrains détruits.
- Avec plus de 46° le 11 mai 2022, l'Inde connaît une sécheresse centennale ce qui pose des problématiques sur la gestion des ressources en eau.

Plus généralement, l'augmentation des températures dû au réchauffement de la planète et la répétition des épisodes de sécheresse viennent perturber le cycle de l'eau. Outre les conséquences directes sur les récoltes, la gestion des ressources en eau et l'aggravation des feux de forêts, l'élévation des températures provoquera une augmentation de la fréquence et de l'intensité des inondations. Dans les régions les plus humides, l'augmentation de la température va accentuer le phénomène d'évaporation à la surface du sol et d'évapotranspiration des plantes.

Or, plus une masse d'air est chaude plus sa capacité de rétention en vapeur d'eau est élevée. Lorsque cette masse d'air chargée en vapeur d'eau s'élève et se refroidit, il s'en suit des pluies diluviennes. En revanche, dans les régions où l'eau se fait plus rare, celle-ci s'évapore mais ne suffira pas à créer des nuages et de la pluie entraînant alors une aridification des terrains.

La sécheresse n'est pas sans conséquence sur l'être humain, l'épisode caniculaire de 2003 a causé le décès d'environ 13500 personnes. Le graphique 1.2 montre l'évolution du nombre de décès journalier rapporté à la moyenne des quatre années précédentes. Nous y remarquons une brève sur-mortalité en comparaison avec la crise de la COVID-19 [1].

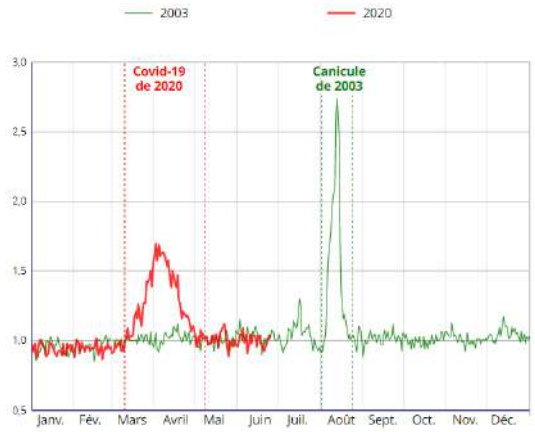


FIGURE 1.2 – Évolution de la mortalité journalière - SOURCE : INSEE

1. Source : https://www.insee.fr/fr/statistiques/4764693

A terme, certaines régions humides du globe seront inhabitables car le corps humain ne pourra plus se refroidir via la transpiration ce qui nous amène à nous questionner sur les futurs mouvements de population de réfugiés climatiques.

1.1.2 Définition et typologie des différentes sécheresses

Comme nous l'avons évoqué, la sécheresse est un terme général regroupant plusieurs typologies de sécheresse aux origines ainsi qu'aux conséquences diverses et variées. De manière générale, nous pouvons distinguer quatre grandes typologies de sécheresse :

- **La sécheresse météorologique** est caractérisée par un déficit de précipitations et mesurée à plusieurs horizons temporels allant de 1 à 12 mois.
- **La sécheresse agricole**, caractérisée par un déficit hydrique des sols, nuit au développement de la végétation.
- **La sécheresse hydrologique** est caractérisée par un niveau des cours d'eau, lacs et/ou nappes phréatiques anormalement bas.
- **La sécheresse géotechnique**, caractérisée par un déficit hydrique des sols argileux, cause de nombreux dégâts sur les bâtiments.

1.1.3 Les indices de sécheresse

De la même façon, il existe de nombreux indices de sécheresse permettant d'apprécier l'intensité d'un phénomène selon sa typologie. Ils ont tous pour point commun d'être standardisés (i.e centré et réduit) ce qui permet d'apprécier la déviance de l'indice par rapport à sa moyenne historique. Le processus de standardisation permet également de rendre comparable ces indices dans l'espace et le temps.

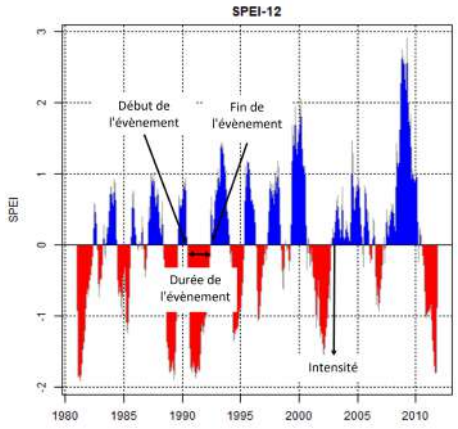


FIGURE 1.3 – Exemple fictif d'un indice standardisé

A la lecture d'un indice de sécheresse, nous pouvons :
- dater le début et la fin d'un évènement à partir du moment ou l'indice devient négatif ou positif. Par conséquent, nous pouvons déduire la durée de l' évènement.
- déterminer la sévérité (minimum ou maximum absolu)
- déterminer la magnitude (valeur absolue de la somme des valeurs de l'indice pendant l'évènement).

La liste non exhaustive qui suit dresse un état des lieux des indices de sécheresse les plus fréquemment utilisés. La méthodologie de calcul ainsi que la reproduction de ces indices seront détaillés à la page 62 à l'exception des indices de Palmer car ces derniers sont peu commentés dans la littérature.

Les indices météorologiques

Le **SPI (*Standardized Precipitation Index*)** est un indice de sécheresse largement utilisé pour sa simplicité et son efficacité. Il se base uniquement sur les données de précipitations totales mensuelles. Il compare, sur une période choisie (1 à 36 mois) et une localisation donnée, la valeur des précipitations par rapport à son historique. Il peut être interprété comme le nombre d'écart-type par lequel l'anomalie de précipitations s'écarte de sa moyenne long terme. Son principal défaut est sa simplicité car il ne tient pas compte des températures et donc du niveau d'évapotranspiration.

Le **SPEI (*Standardized Precipitation Evapotranspiration Index*)** est une extension du SPI qui contrairement à son homologue, tient compte de l'évapotranspiration pour quantifier la sécheresse. Pour le calculer, nous avons besoin des précipitations mensuelles totales ainsi que des moyennes mensuelles de l'évapotranspiration potentielle. En l'absence de données, l'évapotranspiration potentielle peut être approchée via la formule fermée de Thornthwaite utilisant les températures mensuelles.

Les indices agricoles et géotechniques

Le **SSWI (*Standardized Soil Wetness Index*)** est la mesure standardisé du « Soil Wetness Index » (SWI). Le SWI est l'indice qui est utilisé par la commission interministérielle pour émettre son avis sur le caractère anormal de la sécheresse. Sur la base de critères que nous évoquerons plus tard, il conditionne l'éligibilité au dispositif Cat Nat. Le SWI mesure le niveau d'humidité contenue dans le sol par rapport à son niveau optimal. Il est compris entre 0 (sol sec) et 1 (sol saturé en eau). Le SSWI mesure donc l'écart normalisé de cet indice par rapport à sa moyenne historique.

Les indices de Palmer

Les indices de Palmer regroupent respectivement :
- ***Palmer Drought Severity Index* (PDSI)**, est un indice de sécheresses à long terme largement utilisé aux États-Unis par l'Administration Américaine pour les Océans et l'Atmosphère (NOAA). Pour calculer cet indice, nous avons

besoin des précipitations totales mensuelles, des températures moyennes mensuelles ainsi que la capacité de rétention des sols en eau. La technique de standardisation de cet indice ainsi que certaines règles de calcul ont été jugées trop arbitraire par la communauté scientifique.

- ***Palmer Hydrological Drought Index* (PHDI)** est une extension du PDSI prenant en considération la sécheresse long terme affectant les réserves en eau des sols, le débit des cours d'eau et les eaux souterraines. L'impact de la sécheresse sur les réserves est plus lente et un retour à la normale est également plus long. Cet indice a donc un temps de réponse moins rapides aux évolutions climatiques.

Dans leurs formules de calcul, certaines constantes empiriques ont été choisies pour l'étude de départ dans le Kansas et l'Iowa. Ces constantes ne sont pas susceptibles d'être identiques pour d'autres régions du monde. C'est pourquoi nous utiliserons la version calibrée des indices de Palmer , [WELLS *et al.*, 2004], afin d'obtenir de meilleure comparaison spatiale ainsi que pour adapter ces indices au territoire français.

1.2 Le RGA, ses origines et conséquences

Comme nous avons pu le voir précédemment, il existe plusieurs typologies de sécheresse aux répercutions variées. Dans ce mémoire, nous nous concentrerons exclusivement sur le risque de retrait-gonflement des argiles (RGA) et ses conséquences sur le bâtiment.

En France, seulement 30% des surfaces agricoles et moins de 20% des agriculteurs sont couverts par l'assurance récolte via les contrats multirisques climatiques. Face à ces constations, l'assurance récolte évolue à compter de 2023, grâce à la loi du 2 mars 2022 et l'ordonnance du 29 juillet 2022, notamment en imposant une couverture des risques partagée entre l'assuré via la franchise, l'assureur et l'Etat pour les évènements exceptionnels dont les seuils d'intervention dépendent de la typologie de culture. La limitation de l'indemnisation pour les agriculteurs non assurés devrait permettre un accroissement des surfaces agricoles assurées contre les risques climatiques.

Aujourd'hui, la majeure partie des conséquences agricoles de la sécheresse sont indemnisées par le « Fonds National de Gestion des Risques en Agriculture » (FNGRA) via le régime des calamités agricoles. Les contrats multirisques climatiques en assurance agricole peuvent également couvrir les pertes d'exploitation liées aux épisodes de sécheresse et de forte chaleur. Les conséquences de la sécheresse pour les assureurs sont essentiellement portées par le risque de retrait-gonflement des argiles.

Cette section présente l'origine du phénomène, la prédisposition du territoire et les critères de vulnérabilité qui risque de déclencher la survenance ou d'aggraver les conséquences matériels du RGA.

1.2.1 Description du phénomène physique

Le RGA fait partie de la typologie des sécheresses géotechniques, il est donc mesuré par l'indice d'humidité des sols standardisés (SSWI) qui compare l'état de la réserve hydrique avec son historique.

Le risque de retrait-gonflement des argiles ou mouvements de terrains différentiels consécutifs à la sécheresse et à la réhydratation des sols comme son nom l'indique est la conséquence d'un changement d'amplitude des sols argileux sur le bâtiment. Ce péril résulte de l'interaction entre trois composantes essentielles :

- Des conditions météorologiques défavorables affectant le bilan hydrique des sols.
- Une prédisposition au phénomène, le niveau d'argile dans le sol.
- Un bâtiment vulnérable exposé.

L'argile contenu dans le sol s'organise sous la forme de couches superposées les unes aux autres, nous parlons alors de structure minéralogique en feuillet. L'espace entre les différentes couches d'argiles laisse place à l'eau infiltrée (eau interstitielle) ce qui confère à l'argile une modification de ses propriétés.

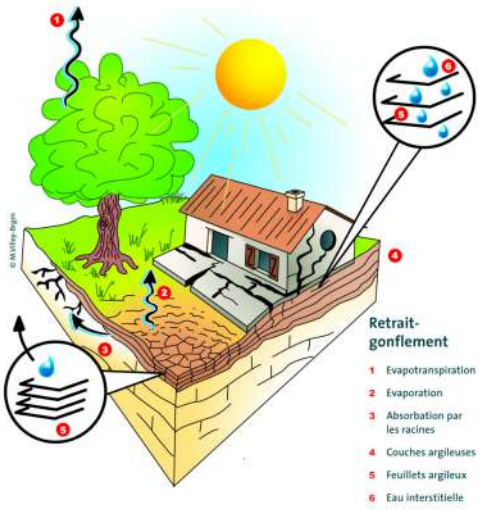


FIGURE 1.4 – Description du phénomène physique

Au contact de l'eau, les feuillets argileux vont gonfler et devenir malléables, on parlera alors de gonflement des argiles. A l'inverse lors d'épisodes de forte chaleur favorisant l'évaporation de l'eau et/ou lors de déficit de précipitations, les sols argileux vont se rétracter et devenir cassants, nous parlerons alors de retrait des argiles. C'est la succession de ces épisodes de sécheresse et de réhydratation qui vont perturber le bilan hydrique des sols causant à la fois une modification de la consistance de l'argile ainsi qu'une variation du volume des sols fragilisant la structure du bâtiment.

L'amplitude de ce phénomène dépend de la composition minéralogique des sols argileux. Certains minéraux possèdent un potentiel de déformation plus élevé que d'autres soit une prédisposition au RGA plus forte. C'est le cas pour la smectite, la vermiculite ainsi que la montmorillonite tandis que l'illite et la kaolonites sont des minéraux moins sensibles aux variations de volume.

In fine, tous les éléments affectant l'infiltration ou l'évapotranspiration de l'eau vont

venir perturber le bilan hydrique du sol. Par exemple, en zone urbaine l'imperméabilisation des sols limite le phénomène car l'eau peine à s'infiltrer comme à s'évaporer tandis que la présence de végétation à proximité des bâtiments va affecter le bilan hydrique au travers du processus d'évapotranspiration.

1.2.2 La susceptibilité du territoire au RGA

La cartographie de la susceptibilité du territoire au RGA est un programme de cartographie qui a été initié en 1997. Ce programme a été en partie financé par le Fonds de Prévention des Risques Naturels Majeurs (FRPNM), le ministère de la recherche, la Caisse Centrale de Réassurance (CCR) ainsi que le Centre Européen de Prévention des Risques (CEPR). L'analyse géologique des sols a permis d'identifier plus de 2000 formations argileuses sur le territoire métropolitain.

L'objectif de cette carte est de délimiter les zones à priori sujettes au RGA en France métropolitaine et de les hiérarchiser selon un degré d'aléa croissant en 4 zones prédéfinis :

- **Non renseigné :** Au sein de cette zone considérée sans aléa (zone blanche), il n'y a pas à priori de couches argileuses sub-affleurantes. Toutefois, il n'est pas exclue qu'il puisse exister des lentilles argileuses non identifiées sur les cartes géologiques et pouvant causer des désordres aux habitations.
- **Faible :** Au sein de cette zone, nous estimons que seul une sécheresse de forte intensité peut causer des dégâts. De plus seul, une faible proportion de bâtiments sera concerné par les désordres avec en priorité les bâtiments les plus vulnérables.
- **Moyen :** Cette zone est considérée comme une zone intermédiaire entre la zone faible et la zone forte.
- **Fort :** Au sein de cette zone, nous estimons que la probabilité qu'un bâtiment soit sinistré est la plus forte. De plus, l'amplitude du phénomène au sein de cette zone est aussi considérée comme la plus élevée.

Le degré de susceptibilité (nul, faible, moyen ou fort) est fonction de deux facteurs de prédisposition principaux : la nature du sol et le contexte hydrogéologique. En ce qui concerne la nature du sol, la probabilité de survenance du phénomène au sein des formations argileuses est la somme :

- Du comportement géotechnique : les propriétés mécaniques.
- De la composition minéralogique : la proportion de minéraux argileux favorables et la proportion de matériaux argileux.
- De la nature lithologique : la profondeur et l'épaisseur de la formation argileuse.

La présence de nappes phréatiques ou de ruissellement sous la surface et à faible profondeur sont également des facteurs de prédisposition affectant le bilan hydrique des sols.

La figure 1.5 représente la carte de susceptibilité des sols au RGA pour la France métropolitaine. Les zones d'aléa non renseigné, faible, moyen et fort représentent respectivement 37%, 42%, 19% et 2% du territoire métropolitain. Toutefois, nous verrons

ultérieurement que certaines zones d'aléa fort ne sont pourtant pas les plus sinistrogènes, ce qui introduira la carte d'exposition du territoire au RGA et non plus la susceptibilité au phénomène.

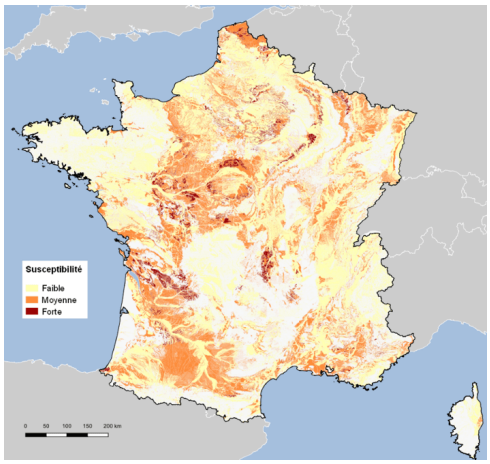


FIGURE 1.5 – Cartographie de la susceptibilité au RGA

1.2.3 Les conséquences sur le bâtiment

Les variations d'amplitude des sols argileux ne sont pas sans conséquences sur le bâti. En effet, l'hétérogénéité des mouvements entre deux points de la structure va conduire à une déformation pouvant entraîner la fissuration voire la rupture de la structure. La réponse du bâtiment dépend de ses possibilités de déformation ainsi que de ses points de faiblesses.

Si la majeure partie de ces désordres engendre des travaux d'embellissement ou de finition consistant à enduire les fissures et repeindre la façade, dans de rares cas ces fissures peuvent mettre en péril la structure du bâtiment ce qui nécessite une intervention plus complexe et plus coûteuse. Il s'agit de la reprise en sous-oeuvre qui consiste à reprendre les fondations de l'habitation et dont le coût peut atteindre plusieurs centaines de milliers d'euros.

FIGURE 1.6 – Fissuration (à gauche) et reprise en sous-oeuvre (à droite)

D'autres désordres peuvent se manifester comme la distorsion des portes et fenêtres,

la dislocation/affaissement des dallages, voir dans de rares cas une rupture de canalisation ce qui va perturber d'avantage le bilan hydrique des sols.

La construction sinistrée type correspond à une habitation individuelle de plain-pied reposant sur des fondations inadaptées, à la géométrie complexe et en présence d'arbres à proximité immédiate de la structure. Les maisons construites sur des terrains en pente dont les fondations présentent des différences d'ancrage d'un point à un autre de la structure, sont encore plus sensibles.

1.2.4 Les mesures de prévention et de remédiation

Si les origines et conséquences du RGA sont désormais clairement identifiées, les mesures préventives ainsi que de remédiation se sont révélés insuffisantes ou tardives.

Les préconisations de construction

Le Bureau de Recherche Géologique et Minière (BRGM) préconise des mesures de construction pour pallier aux conséquences de la sécheresse :

- approfondir les fondations pour qu'elles soient ancrées dans un terrain peu sensible aux variations saisonnières d'humidité (en deçà des feuillets argileux).
- homogénéiser ces profondeurs d'ancrage pour éviter les dissymétries sur les terrains en pente.
- réaliser un trottoir étanche autour de la maison pour limiter l'évaporation à proximité immédiate des façades.
- maîtriser les eaux de ruissellement et les eaux pluviales pour éviter les infiltrations au pied du mur.
- ne pas planter d'arbres trop près de la maison.

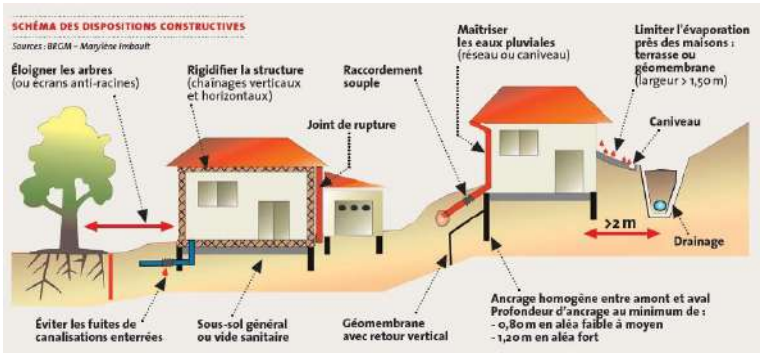


FIGURE 1.7 – Mesures de construction préventive *Source : BRGM*

La loi Elan, une réglementation tardive

En application de l'article 68 de la loi ELAN du 23 novembre 2018, une sous section consacrée à la prévention du RGA a été ajoutée au Code de la construction et de l'habitat.

Le décret n° 2019-495 du 22 mai 2019 impose la réalisation de deux études de sol dans les zones d'exposition moyenne ou forte au retrait-gonflement des argiles :

- à la vente d'un terrain non bâti mais constructible, le vendeur doit remettre une étude géotechnique qui sera annexée à l'acte de vente ou à la promesse de vente. Toutefois si les dispositions d'urbanisme ne permettent pas de construire de maison individuelle sur ce terrain alors l'étude géotechnique n'est plus une obligation.
- au moment de la construction de la maison : l'acheteur doit faire réaliser une étude géotechnique à destination du constructeur. Si cette étude géotechnique révèle un risque de mouvement de terrain différentiel consécutif à la sécheresse et à la réhydratation des sols, le constructeur doit en suivre les recommandations ou respecter les techniques particulières de construction définies par voie réglementaire.

L'arrêté ministériel du 22 juillet 2020 définit le contenu des études géotechniques à réaliser dans les zones exposées à l'aléa moyen ou fort :

- caractérisation du comportement des sols d'assise vis-à-vis du phénomène de retrait-gonflement.
- reconnaissance de la nature géologique et des caractéristiques géométriques des terrains d'assises.
- vérification de l'adéquation du mode de fondation prévue par le constructeur avec les caractéristiques et le comportement géotechnique des terrains d'assises.
- vérification de l'adéquation des dispositions constructives prévues par le constructeur avec les caractéristiques intrinsèques du terrain et son environnement immédiat.

Il a fallu attendre 2018 pour qu'un bond soit fait en matière de prévention. En rendant obligatoire la réalisation d'une étude géotechnique et en imposant aux constructeurs des mesures de constructions préventives, le nombre de nouveaux bâtiments vulnérable va fortement décroître. Cette loi, certes tardive mais nécessaire, aurait dû voir le jour bien avant étant donnée que le risque ainsi que les facteurs aggravant était déjà connus à l'époque.

Les Plans de Prévention des Risques Naturels (PPRN)

Les plans de préventions des risques naturels prévisibles (PPRN) ont été institué par la loi du 2 février 1995 et sont définis aux articles L562-1 du Code de l'environnement. Les PPRN constituent le principale levier d'actions pour la prévention des catastrophes naturelles, leurs objectifs est de réduire la vulnérabilité des personnes et des biens. Pour cela les PPRN ont la possibilité, au sein des zones exposées, d'interdire les nouvelles constructions notamment résidentiels, commerciales ou encore industriels. Ils ont également la possibilité de prescrire les conditions dans lesquels ces bâtiments peuvent être construits (bâtiments adaptés, réhabilitation des terrains) ou d'imposer des travaux de prévention pour les habitations préexistantes.

D'autres mesures peuvent également être retenues pour les zones qui ne sont pas di-

rectement exposées si les aménagements prévus constituent un facteur aggravant ou sont susceptibles de provoquer de nouveaux évènements. La réalisation des travaux nécessaires peut être rendue obligatoire dans un délai de 5 ans, à défaut, des sanctions pénales peuvent s'appliquer et l'assureur est en droit de refuser l'indemnisation de l'assuré même en cas de reconnaissance Cat Nat.

En ce qui concerne la sécheresse, les PPRN délimitent à l'échelle communale les zones exposées au RGA grâce à la carte d'exposition conçu par le BRGM et préconise des règles constructives obligatoires ou recommandées afin de réduire ne serait-ce que partiellement l'apparition des désordres. Il peut également imposer la réalisation d'une étude géotechnique à la vente d'un terrain constructible ou avant un projet de construction. Malheureusement, même pour les zones fortement exposées, les PPRN sécheresse ne prévoient pas actuellement d'inconstructibilité mais des prescriptions de bon sens engendrant un faible surcoût de construction.

La carte 1.8 montre la localisation et l'état des PPRN sécheresse en France, nous pouvons nous apercevoir que les PPRN sécheresse prescrits sont très peu nombreux au regard du territoire. Cette mesure de prévention, qui pourtant a un fort pouvoir préventif, n'a pas été utilisée à bonne escient par les administrations publiques au grand regret des assureurs et assurés.

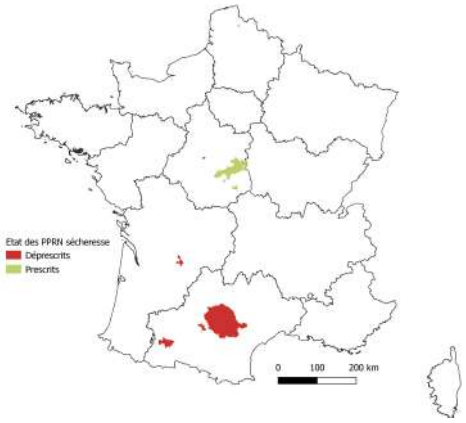


FIGURE 1.8 – Etat des PPRN sécheresse en France

Le projet MACH et MACH+

Le projet MACH pour MAison Confortée par Humidification, est une solution de remédiation alternative aux travaux de grands ampleurs comme la reprise en sous-oeuvre. Missionné par la direction Générale de la Prévention des Risques (DGPR), le Centre d'études sur les Risques, l'Environnement, la Mobilité et l'Aménagement (CEREMA) a développé la solution MACH qui consiste a humidifier le sol de fondation tout autour de

la maison en injectant de l'eau de pluie stockée afin de limiter les variations d'amplitude des sols en cas de sécheresse. La réhumidification des sols est composée de trois étapes :

- La collecte et le stockage d'eau de pluie.
- La mesure de la succion des sols et le déclenchement de la réhumidification par un seuil critique.
- L'acheminant de l'eau à différents points de la structure via un réseau hydraulique.

Pour le moment, le dispositif a été expérimenté uniquement sur une maison individuelle construite dans les années 1960 et sinistrée en 2015 mais non reconnue en l'état de catastrophe naturelle. Toutefois, les performances mesurées par des fissuromètres sont prometteuses. Les résultats ont montré à la fois une stabilisation des fissures mais aussi l'absence d'apparition de nouvelles fissures sachant que l'année 2018 a été particulièrement marquée par la sécheresse.

Avec un coût avoisinant les 15000 €, cette solution s'avère plus rentable au long terme étant donné que les travaux d'embellissement ont un coût moyen de 6300 € et les reprises en sous-oeuvre ont un coût moyen de 24800 € selon une étude d'analyse des rapports d'expertise initiée par la MRN en 2018 [2].

Cette solution va à l'avenir être appliquée à un échantillon plus important et prévoit également un système d'intelligence artificielle automatisé qui déclenchera spontanément la réhydratation en fonction du bilan hydrique des sols.

1.3 La subsidence dans le régime Cat Nat

1.3.1 Le régime Cat Nat

Après une année marquée par de fortes inondations touchant en particulier la ville de Saintes en Charente-Maritimes mais aussi la Seine-et-Marne, le régime d'indemnisation contre les catastrophes naturelles a été institué le 13 juillet 1982 afin de pallier au manque de couverture sur ces évènements naturels. A l'exception des contrats d'assurance pour les bateaux, l'assurance catastrophes naturelles devient une extension de garantie obligatoire pour tous les contrats d'assurance dommages comme la multirisque habitation ou encore l'assurance tous risques automobiles.

La liste non exhaustive qui suit donne un aperçu des catastrophes naturelles dont les dommages sont couverts par le régime Cat Nat :

- Les inondations : par remontées de nappes phréatiques, par submersion marine, par ruissellement ou crue torrentielle, ...
- Les mouvements de terrain
- Les séismes
- Les tsunamis

2. Source : https://www.mrn.asso.fr/rapport-mrn-secheresse-geotechnique/

- Les avalanches
- Les éruptions volcaniques
- Les cyclones, les ouragans sous conditions paramétriques sur la vitesse des vents et/ou des rafales
- La sécheresse

La sécheresse au sens du RGA n'a été incluse au régime Cat Nat qu'à partir de l'année 1989. A l'époque, en plus d'un déficit pluviométrique, l'année 1989 est l'une des plus chaudes enregistrée depuis les années 1950 avec 1.4°C au dessus de la normale de l'époque. L'effet de cette épisode de sécheresse, particulièrement long puisqu'il s'est étalé de 1988 à 1990, a eu un impact important sur le bâti avec 3869 communes touchées et 250 Millions d'euros de dégâts à l'échelle de la France métropolitaine.

Compte tenu de leur assurabilité, les tempêtes, les chutes de grêle et de neige font partie des périls exclus de la garantie Cat Nat mais peuvent être des garanties obligatoires comme c'est le cas pour la tempête ou optionnelles pour les deux autres périls. Nous parlons de risque assurable lorsqu'un assureur peut tarifier le risque en proposant une prime finie. Le coût élevé des catastrophes naturelles associé à une fréquence non négligeable sur des régions très exposées font de ces dernières des risques non assurables.

Définition des effets dommageables

Selon l'article L 125-1 du code des assurances relatif à l'indemnisation des victimes de catastrophes naturelles, il est considéré comme effet dommageable imputé aux catastrophes naturelles : « **Les dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel, lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises.** ». En revanche, sont exclus du champ d'application les désordres subis par les corps aériens, maritimes, lacustres et fluviaux. Les dommages causés aux récoltes et aux cultures font l'objet d'une indemnisation spécifique via le régime des calamités agricoles.

Cette définition des désordres liés aux catastrophes naturelles a évolué avec la loi n°2021-1837 récemment promulguée le 28 décembre 2021. En effet, cette dernière apporte des modifications ainsi que des ajouts aux textes de lois précédemment en vigueur. Elle vise à améliorer la transparence des décisions rendues par la commission interministérielle et par les experts, à faciliter les démarches de demandes de reconnaissance et adopte des spécificités en ce qui concerne le RGA. Désormais, les frais de relogements d'urgence des personnes sinistrées dont l'habitation principale est rendue impropre pour des mesures de sécurité ou de salubrité directement imputables à l'évènement climatique sont ajoutés aux effets dommageables des catastrophes naturelles. Il en va de même pour le remboursement des études géotechniques préalables ainsi que pour les frais d'architecte et de maîtrise d'oeuvre qui sont à la charge de l'assureur.

Loi du 28 décembre 2021

La redéfinition des effets imputables aux catastrophes naturelles par la loi du 28 décembre 2021 [Légifrance, 2021] s'est accompagnée d'autres modifications et retraitements. Pour n'en citer que quelques unes :

- Les actions dérivant d'un contrat d'assurance relatives aux dommages liés au RGA sont prescrites à partir de 5 ans à compter de l'évènement.
- La commission nationale va devoir rendre un rapport annuel incluant un avis sur la pertinence des critères de reconnaissance ainsi que sur les modalités et conditions pour mandater un expert certifié à l'étude des catastrophes naturelles.
- Les indemnisations des désordres liés au RGA devront couvrir les travaux permettant un arrêt total des désordres dès lors que les résultats de l'expertise montre une atteinte à la solidité de l'habitation ou une nature impropre à l'habitation (dans la limite de la valeur du bien assuré).
- Dans un délai de 6 mois, le Gouvernement doit remettre un rapport sur les moyens de renforcement des constructions. Il doit également émettre des propositions sur un système juridique et financier propre au RGA, viable à long terme et permettant l'indemnisation des propriétaires concernés par ces désordres. Cette ouverture évoque une possible sortie du régime Cat Nat et de la garantie décennale pour le RGA au motif que ce risque représente un risque certain dans certaines régions et non plus un aléa de nature imprévisible. De plus, il est possible que des mesures de construction préventive puissent empêcher la survenance du risque ce qui serait contraire à la définition des effets dommageables aux catastrophes naturelles par l'article L125-1.
- Les demandes de reconnaissance en l'état de catastrophe peuvent intervenir dans les 24 mois à compter de la date de début des évènements contre 18 mois auparavant.
- Le délai de publication de l'arrêté au journal officiel est abaissé de trois à deux mois à compter du dépôt des demandes communales.
- Les montants des franchises dépendront du niveau d'aléa et les modulations seront supprimées si aucun PPRN n'a été prescrits.

Cette réforme du régime d'indemnisation, en faveur des élus locaux et des assurés, va venir alourdir le bilan des assureurs. En effet, le rallongement du délai pour effectuer les démarches administratives est propice à l'augmentation des sinistres tardifs. De plus, les assureurs vont devoir remédier définitivement aux désordres laissant penser à une augmentation des reprises en sous-oeuvre et donc à une augmentation du coût moyen des sinistres sécheresse.

Le niveau des franchises

L'Etat impose un niveau de franchise obligatoire et non rachetable aux assureurs pour les dommages liés aux Cat Nat. Ces niveaux de franchise sont modulés à la hausse en fonction du nombre de reconnaissance Cat Nat de la commune au cours des 5 dernières

années si aucun PPRN n'a été prescrits. La sécheresse est un péril particulier puisqu'il n'est apprécié qu'en France métropolitaine et comme le coût des sinistres sécheresse est particulièrement élevé, le niveau de la franchise est spécifique sur ce péril. Le tableau 1.2 récapitule les niveaux de franchise applicables depuis 2001.

Biens à usage d'habitation et autres bien à usage non professionnels	Dommages directs	380 €	1520 € (sécheresse)
Biens à usage professionnel	Dommages directs	10% minimum à 1140 €	10% avec un minimum à 3050 € (sécheresse)
	Pertes d'exploitation	3 jours ouvrés minimum à 1140 €	

TABLE 1.2 – Niveau de franchise applicable

De plus, si plusieurs arrêtés ont été publiés pour la commune au cours des cinq dernières années sur le même péril alors la franchise peut être modulée :

- 1 à 2 reconnaissances : Franchise simple
- 3 reconnaissances : Franchise double
- 4 reconnaissances : Franchise triple
- 5 reconnaissances : Franchise quadruplée

Comme nous avons déjà pu l'évoquer, la récente réforme du régime d'indemnisation Cat Nat datant de décembre dernier prévoit un montant de franchise variant en fonction du niveau d'aléa de l'évènement ainsi que la suppression des modulations de franchise en l'absence de PPRN prescrits.

Le fonds barnier

L'extension de garantie Cat Nat est financée par une surprime dont le taux est définit par l'Etat et répartit uniformément sur le territoire par principe de solidarité. Autrement dit, il n'y a pas de modulation de la prime en fonction de l'exposition et de la typologie du péril auquel l'assuré est exposé. Le montant de cette surprime s'élève à 12% de la prime afférente aux garanties dommages du contrat de base pour les biens autres que les véhicules à moteur et 6% des primes vol et incendie (ou à défaut, 0,50% de la prime dommage) pour les véhicules terrestres à moteur.

Cette participation financière vient alimenter le Fonds de Prévention pour les Risques Naturels Majeurs (FPRNM) dit « Fonds Barnier » dont le budget est utilisé pour maintenir des mesures de préventions et d'informations (carte d'aléa, PPRN), les mesures de délocalisation (expropriation, acquisition à l'amiable), les mesures d'adaptation du

territoire ainsi que pour financer d'autres études et travaux. La gestion financière et comptable de ce fond est gérée par la Caisse Centrale de Réassurance (CCR).

En France, l'Etat garantit une réassurance illimitée à la CCR c'est pourquoi la majeure partie des assureurs opérant en France se réassure auprès de la CCR pour se couvrir des risques naturelles. En règle général, les traités de réassurance pour les risques naturelles reposent sur un traité en Quote-Part (QP) à 50% suivi d'un Stop Loss (XL) qui est appliqué sur le reste à charge de l'assureur. Le schéma suivant détaille la structure d'un tel traité ainsi que les montants à charge pour les différentes parties.

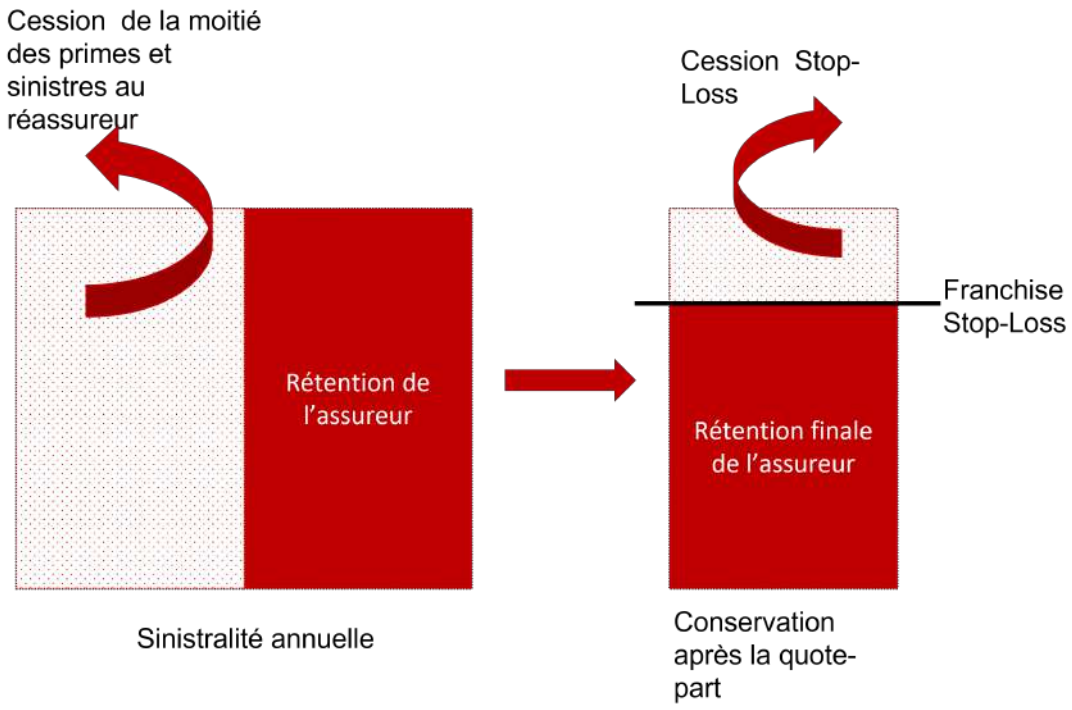


FIGURE 1.9 – Schéma de réassurance Cat Nat *Source : CCR*

1.3.2 La procédure administrative

Pour qu'un assuré soit indemnisé au titre des désordres liés au RGA, il est nécessaire que la commune soit concernée par la publication d'un arrêté Cat Nat au journal officiel comme cela doit être le cas pour toutes les catastrophes naturelles. Les délais d'instruction des demandes communales ainsi que le délai pour effectuer la demande de reconnaissance ont été modifié par la loi du 28 décembre 2021 en faveur des assurés. La figure 1.10 récapitule les différentes étapes de la procédure ordinaire ainsi que l'acheminement du dossier.

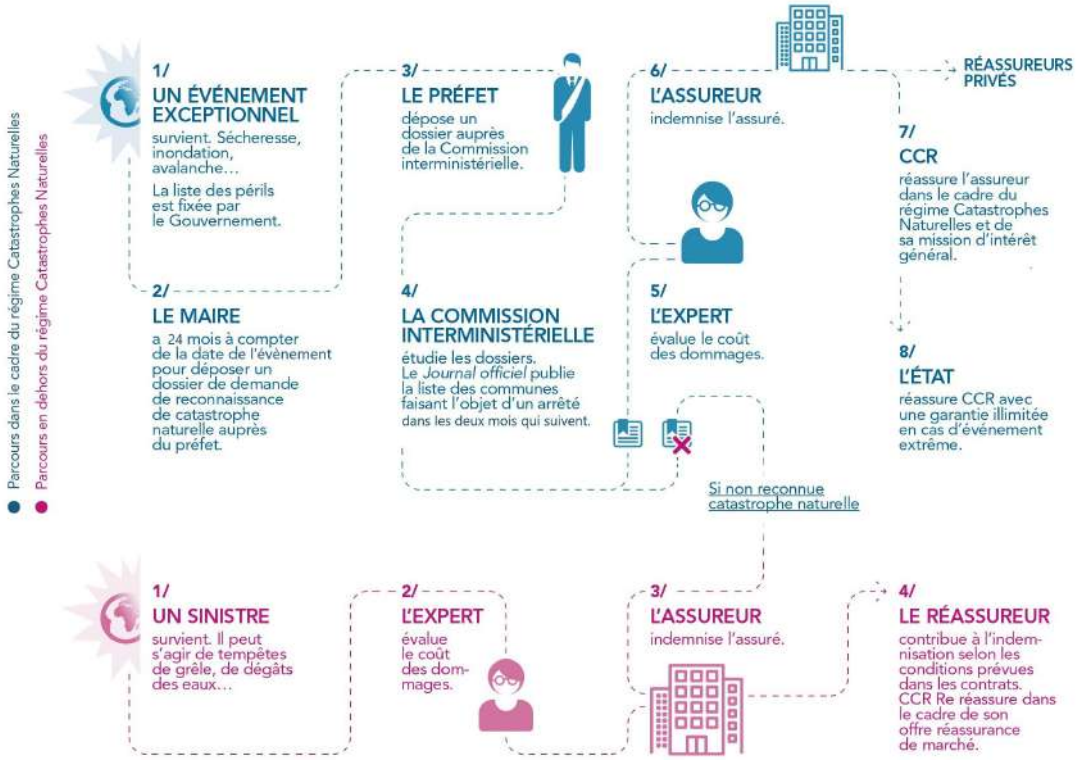


FIGURE 1.10 – Procédure ordinaire de demande de reconnaissance *Source : [CCR, 2022]*

En premier lieu, lorsqu'un assuré est sinistré celui-ci doit déclarer le sinistre auprès de son assureur et solliciter le maire de sa commune pour qu'il puisse faire la demande de reconnaissance en l'état de catastrophe naturelle. A ce moment, l'état du dossier est déclaré « ouvert » dans les systèmes de gestion de données des assureurs car pour le moment, aucun arrêté n'a été publié.

Le maire a ensuite 24 mois à compter du début de l'évènement pour formuler sa demande auprès de la préfecture de son département. Si ce délai est aussi long, c'est à cause de la cinétique lente de la sécheresse dont les dégâts peuvent survenir bien après l'épisode en lui-même. Le dossier envoyé à la préfecture doit contenir le formulaire CERFA de demande ainsi que la carte d'exposition sur la commune complétée par la localisation des sinistres ainsi que la date de début et fin de l'évènement.

Afin de ne pas avoir à formuler plusieurs demandes au cours d'une année et pour éviter le risque de ne pas être reconnue pour une période non sollicitée, les élus sont invités à indiquer l'année civile dans son entièreté pour l'étude du dossier. Le graphique 1.11 montre l'évolution des temps d'instructions des demandes de reconnaissance au cours du temps. Nous pouvons constater l'amélioration de ces temps d'instructions au fil du

temps.

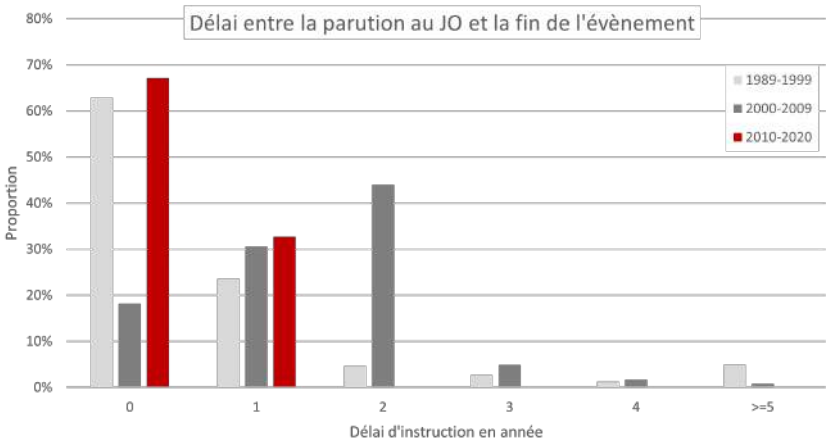


FIGURE 1.11 – Evolution des délais d'instruction *Source : GASPAR*

Après réception du dossier par la préfecture, celle-ci collecte auprès de Météo France les informations et rapports techniques supplémentaires qu'elle inclue au dossier comme les périodes de retour du niveau de l'humidité des sols pour une instruction sur la sécheresse.

Puis, la préfecture envoie au ministère de l'intérieur un dossier par commune qui sera ensuite analysé par la commission interministérielle. Celle-ci émet un avis favorable ou défavorable en fonction de l'intensité du phénomène ou des critères de reconnaissance. L'arrêté sera publié au journal officiel quelques jours plus tard en indiquant l'avis pris par la commission. Dès lors, les assurés sinistrés n'ont plus que 10 jours à compter de la publication de l'arrêté pour avertir leur assureur si cela n'a pas été déjà fait.

En cas d'avis défavorable, les sinistres « ouverts » doivent être classés « sans-suite » et en cas d'avis favorable le dossier sera clôturé une fois que l'indemnisation a eu lieu. La clotûre du dossier peut être assez longue suivant les cas, c'est pourquoi il y a peu de sinistres clos pour les années récentes.

Le graphique 1.12 montre l'évolution du nombre de communes reconnues en l'état de catastrophe naturelle sécheresse depuis 1989.

Les années 2003 et 2011 ont été particulièrement marquées par la sécheresse tout comme les quatre dernières années enregistrées. Comme le RGA est entré dans le régime Cat Nat en 1989, les sécheresses survenues avant cette date ont été reconnues à ce moment. En plus de cela, la seule mise en place du critère géotechnique a conduit à un taux de reconnaissance supérieur à 90%. C'est la raison pour laquelle nous apercevons autant de communes reconnues à cette date.

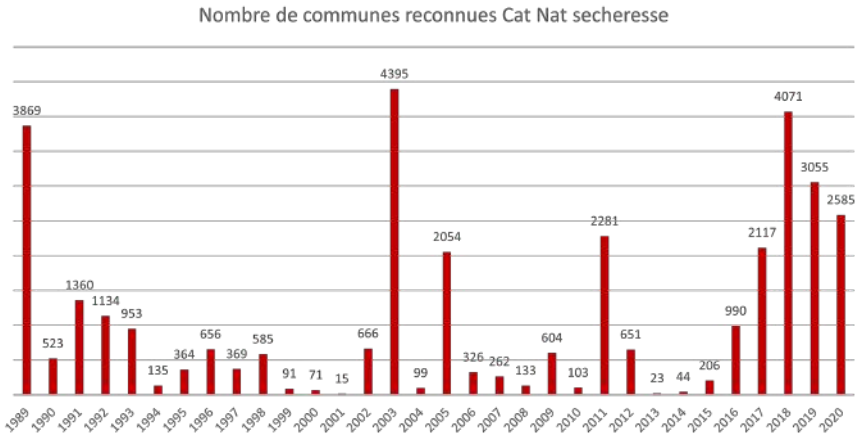


FIGURE 1.12 – Évolution du nombre de reconnaissance Cat Nat sécheresse *Source : GASPAR*

1.3.3 L'évolution des critères de reconnaissance

Pour que la commission interministérielle donne son avis favorable en ce qui concerne la demande de reconnaissance en l'état de catastrophe naturelle sécheresse, certaines conditions géotechniques et météorologiques doivent être respectées.

Or depuis l'intégration du RGA au régime Cat Nat en 1989, ces conditions définies par des critères techniques ont évoluées au fil de l'amélioration scientifique sur la connaissance du phénomène et au gré des épisodes de sécheresse aux conditions d'apparitions disparates.

Jusqu'en 1999, l'éligibilité d'une commune à la reconnaissance Cat Nat se fondait uniquement sur la présence d'argile sur son territoire (critère géotechnique) et sur la présentation d'un rapport de la station météorologique la plus proche afin de justifier le caractère « anormal » de la sécheresse subie. Le critère géotechnique d'une commune est validé lorsque la surface en argile avéré représente plus de 3% de la surface communale. En ce qui concerne le rapport météorologique, celui-ci repose sur l'analyse des déficits de précipitations pour caractériser l'intensité de la sécheresse. Ces critères ont conduit à un taux de reconnaissance sur les demandes supérieures à 90% à l'époque mais ne reposait sur aucun critère scientifique précis.

A partir de l'année 2000, il s'est ajouté au critère géotechnique un critère reposant sur une définition « agricole » de la sécheresse, comprise comme un déficit de l'humidité des sols superficiels et mesurant l'état de la réserve hydrique du sol par rapport à son niveau optimal. Dès lors, le critère géotechnique bien que présent s'est retrouvé quelque peu délaissé voir dans de rare cas (49 cas) outrepassé car peu de communes ne respectent pas ce premier critère. Le critère géotechnique est d'autant plus imperfectible quant à la superficie des communes. En effet, la probabilité de ne pas respecter le critère décroît

avec la superficie de la commune. L'ajout d'un critère hydrique a constitué un indéniable progrès, puisqu'il permet de prendre en compte la réserve en eau du sol, paramètre influençant directement le RGA.

Malgré cela, le critère hydrique initialement évalué pour la période hivernale s'est révélé insuffisant. En effet, l'épisode de 2003 s'est caractérisé par une sécheresse estivale et non hivernale. Si le critère « 2000 » avait été appliqué seulement 200 des 8000 communes demanderesses, environ, auraient été reconnues [3]. A la suite de l'épisode intense de 2003, un critère météorologique estival a été instauré puis assoupli deux fois au cours de l'année 2005.

Au gré des épisodes de sécheresse, de nouveaux critères sont apparus comme en 2011 où la sécheresse a été printanière et plus récemment en 2018 après une refonte complète des critères de reconnaissance. Le tableau 1.3 récapitule les critères adoptés et ainsi que les assouplissements opérés depuis 1989.

TABLE 1.3 – Évolution des critères dans le temps

Période	Critères retenus
2000 (Création du critère hivernal)	Critère géotechnique + choc hivernale : réserve hydrique inférieur à 80% de la normale sur au moins une décade du premier trimestre de l'année (trimestre de fin de recharge) + Calculé sur 4 trimestres consécutifs, l'indice d'humidité des sols doit être inférieur à la normale, le dernier trimestre indique la fin de l'épisode de sécheresse.
2004 (Création du critère estival)	Critère 2000 + rapport entre la moyenne hydrique du troisième trimestre et la moyenne hydrique normale inférieur à 20% + nombre de décades pendant lequel le réservoir hydrique est égal à zéro compris entre le 1er et le 2e rang sur la période 1989-2003
01/2005 (Assouplissement)	Critère 2000 + rapport de la moyenne de la réserve hydrique du 3e trimestre sur la moyenne hydrique normale inférieur à 21% + nombre de décades pendant lequel le réservoir hydrique est égal à zéro compris entre le 1er et le 3e rang sur la période 1989-2003)
06/2005 (Assouplissement)	Critère de janvier 2005 ou critère alternatif : la durée de retour de la moyenne des réserves en eau du sol du troisième trimestre doit être supérieure à 25 ans.
2011 (Création du critère printanier)	Critère du 06/2005 + durée de retour de la moyenne des SWI des 9 décades d'avril à juin supérieur à 25ans.

3. Source : http://www2.senat.fr/rap/r09-039/r09-039.html

2018 (Révision des critères)	Pour chaque saisons, la période de retour d'au moins une grille SWI de la commune doit être supérieur ou égale à 25 sur une période glissante de 50 ans.

En raison du manque de lisibilité des anciens critères envers les élus locaux et les sinistrés ainsi que de l'augmentation des recours gracieux et contentieux sur les décisions prises par la commission, de nouveaux critères plus pertinents, plus facile à exposer et permettant de réduire les délais d'instruction ont vu le jour courant 2018 et restent applicables aujourd'hui. Depuis 2018, deux critères sont pris en compte :

- **Le critère géotechnique :** maintenu depuis 1989, la surface communale en argile avéré (faible, moyen ou fort) doit excéder 3%. Ce critère n'est pas stricte dans le sens où une commune peut tout de même être reconnue après avoir fourni une étude de sols montrant la présence d'argile à un niveau local.
- **Le critère météorologique :** il se base sur l'humidité des sols moyens et il est évalué pour chaque saison avec un critère commun. La période de retour de l'indice doit être supérieur ou égale à 25 ans.

Mise en oeuvre du critère météorologique

L'analyse du critère météorologique repose sur un indice de sécheresse, le « Soil Wetness Index », qui n'est autre qu'un output du modèle hydrométéorologique développé par Météo-France et le Climsec. Le modèle Safran-Isba-Modcou (SIM) modélise le bilan hydrique des sols superficiels à partir d'une multitude de données climatiques observées comme la température de l'air, le niveau de précipitation, le niveau des rayonnement solaire, la vitesse des vents, le niveau de la pression atmosphérique, l'évapotranspiration, l'infiltration, le ruissellement et bien d'autres encore. Ces données hydrométéorologiques permettent une modélisation complexe des interactions entre l'atmosphère et le sol où des hypothèses de texture uniforme sur le territoire ont été prises :

- Une végétation de surface de type gazon.
- Une composition des sols propices aux RGA, 58% d'argile et 12% de sable. [4]

Le SWI évalue la réserve en eau des sols sur une profondeur de 2,35 mètres par rapport à sa réserve optimale. Par construction, plus l'indice est proche de 1 plus le sol est saturé en eau, à l'inverse une valeur proche de 0 indique un sol anormalement sec. Les données du SWI uniforme sont établies sur un maillage géographique fixe (SAFRAN) et numéroté. Ce maillage contient 8981 mailles de $64km^2$ ($8km \times 8km$) recouvrant l'ensemble de la France métropolitaine.

L'indice d'humidité des sols fournis pour un mois donné correspond à la moyenne des indices journaliers de ce mois et des deux mois qui le précède afin de mieux appréhender la cinétique lente de la sécheresse qui peut s'étaler sur plusieurs mois. Au cours d'une

4. Source : [CLIMSEC, 2011]

année donnée, ce sont donc 12 indicateurs qui sont ainsi créés.

Ensuite, pour juger du caractère anormal de la sécheresse, l'indicateur d'un mois donné est comparé avec les 50 derniers indicateurs du même mois. L'application de cette période glissante de 50 ans permet la prise en compte de l'évolution du climat. Prenons l'exemple du mois de janvier 2020, celui-ci correspond à la moyenne des SWI journaliers des mois de novembre 2019, décembre 2019 et janvier 2020. L'indicateur du mois de janvier 2020 est ensuite comparé avec les indicateurs du mois de janvier des années allant de 1970 à 2020 soit 50 ans [5].

Une période de retour supérieure ou égale à 25 ans sur un historique de 50 ans correspond à un indicateur figurant soit au premier ou au second rang des indicateurs les plus faibles jamais enregistrés sur cet historique. Lorsqu' il y a égalité entre deux indices pour un rang r alors l'indicateur le plus ancien conserve sa place au rang r tandis que l'indicateur le plus récent est placé au rang $r + 1$.

Comme les critères météorologiques sont appréciés par saison, pour qu'un trimestre soit éligible il suffit qu'un des indicateurs du trimestre le soit. Prenons l'exemple de la saison hivernale, dans ce cas précis une maille est éligible sur la saison si au moins un des indicateurs des mois de janvier, février ou mars est éligible (i.e période de retour supérieur ou égale à 25 ans).

Enfin, l'éligibilité d'une commune est régie par le principe de « bord bénéfique ». Chaque commune est intersectée par un nombre de mailles qui varie en fonction de sa localisation et de la superficie de la commune. Pour qu'une commune soit éligible selon le critère météorologique pour une saison donnée, il suffit qu'au moins l'une des mailles qui l'intersecte soit éligible sur cette saison.

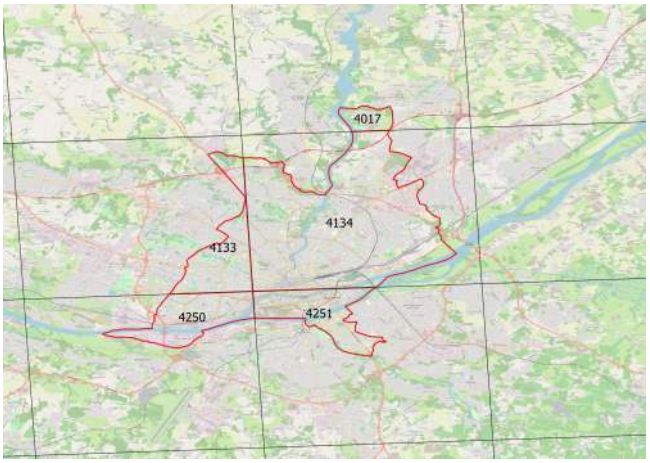


FIGURE 1.13 – Illustration du maillage SAFRAN sur la ville de Nantes

5. La circulaire n° INTE1911312C décrit les modalités de calcul des critères météorologiques mis en place à partir de 2018 :[Légifrance, 2019]

La figure 1.13 donne une superposition des polygones délimitant la commune de Nantes avec le maillage Safran. Pour que cette commune soit éligible sur la saison hivernale de l'année N, il suffit que l'une de ces 5 grilles présentent une période de retour supérieur ou égale à 25 ans sur les mois de janvier de l'année N ou novembre et décembre de l'année N-1.

1.4 Estimation de la charge à l'ultime

Dans cette section, nous commencerons par introduire l'intérêt des assureurs à modéliser le risque de subsidence. Ensuite, nous ferons un état des lieux de quelques études existantes sur le sujet. Ceci nous permettra de proposer une méthodologie pertinente pour la modélisation de ce péril et répondant aux faiblesses rencontrées au cours des études annexes.

1.4.1 Intérêts pour l'estimation de la charge ultime

Les assureurs ont un besoin cruel d'évaluer le montant des risques encourus par les catastrophes naturelles auxquels ils sont exposés.

Contrairement aux modèles Cat Nat qui, par l'objet de simulations/scénarios, permettent de probabiliser les pertes ce qui a une grande utilité pour l'estimation du SCR Cat en Non-Vie. Un modèle déterministe permet d'estimer les pertes réelles encourus par l'assureur à différents instants de l'année en fonction des conditions météorologiques observées. L'anticipation de cette charge ultime permet une meilleure gestion des risques sur ce péril que ce soit pour définir une stratégie tarifaire, réorienter la politique de souscription vers les zones moins exposées (développement commerciale en zone blanche, blocage à la souscriptions sur les zones hotspots) ou encore la surveillance du risque et l'optimisation des traités de réassurance.

L'évaluation au plus juste du coût de ce risque est très utile pour le provisionnement car la sécheresse est une branche longue et une grande partie des sinistres sont déclarés tardivement.

1.4.2 RGA et provisionnement

L'inversion du cycle de production caractérise le secteur des assurances, l'assuré paie la prime commerciale souvent à terme d'avance et l'assureur n'indemnise ce dernier qu'en cas de sinistre dont la date de survenance peut être inconnue lors de la clôture des comptes. C'est la raison pour laquelle les assureurs ont le besoin de constituer des provisions techniques afin de faire face à leurs engagements. Ces provisions sont enregistrées au passif du bilan des assureurs et en représente une large partie ce qui impacte donc directement le résultat de l'assureur.

Les **provisions pour sinistres à payer** (PSAP) couvrent les sinistres déjà survenus qu'ils soient déclarés ou non. D'après l'article R 331-6, la PSAP représente la valeur

estimative des dépenses en principal et en frais, tant internes qu'externes (frais d'expertise, frais judicaires), nécessaires au règlement de tous les sinistres survenus (connus ou inconnus) et non payés. Elle est déterminée net de recours, brut de réassurance et doit être suffisante pour le règlement intégrale des engagements envers les assurés. En contrepartie, la provision pour sinistres réassurés apparait à l'actif du bilan. Les PSAP regroupent deux provisions principales :

- La provision pour les sinistres survenus et déclarés mais dont le paiement n'est pas terminé. La méthode dossier-dossier est à la main des gestionnaires de sinistres qui évaluent au cas par cas le montant restant à payer sur chaque sinistre non clos.
- La provision « ***Incurred But Not Reported*** » (IBNR) qui couvre les sinistres survenus mais non déclarés à l'assureur. Les IBNR peuvent se décomposer en deux sous-parties :
  - La provision pour les sinistres survenus mais pas encore déclarés « ***Incurred But Not yet Reported*** »(IBNyR). Du fait de la cinétique lente de la sécheresse, les sinistres liés à l'évènement peuvent survenir après celui-ci par conséquent les sinistres ne sont déclarés que tardivement. Pour l'exercice en cours, il y a donc très peu de sinistres RGA survenus et déclarés. Le plus souvent, il faut attendre 2 ans pour que les demandes de reconnaissances soient reconnues ou non et que l'assureur procède à la clôture des dossiers.
  - La provision pour les sinistres survenus déclarés mais pas assez provisionnés. « ***Incurred But Not enough Reported*** »(IBNeR), cette provision permet de couvrir l'insuffisance potentielle de la provision initialement placée à la date de clôture des états financiers.

De part sa nature catastrophique, les sinistres imputables au RGA sont rares et volatiles puisqu'ils sont conditionnés à l'intensité de l'aléa, à l'ampleur du phénomène. Le manque de données associé à la cinétique lente de la sécheresse qui en fait une branche longue en termes d'écoulement des sinistres ainsi que les changements de réglementation qui conditionne l'indemnisation des assurés rendent les méthodes statistiques déterministes ou stochastiques de provisionnement inapplicables.

Estimation de la charge ultime

Pour le calcul des IBNR sécheresse, Generali appliquait jusqu'alors un proxy reposant sur les coûts et les fréquences moyennes observées. La charge ultime peut être décomposée de la manière suivante :

$$\text{Charge ultime}_N = \text{Charge}_{\text{sans-suite},N} + \text{Charge}_{\text{ouverts},N} + \text{Charge}_{\text{clos},N} + \text{IBNR}_N$$

avec :

- La charge des sinistres ouverts qui correspond au nombre de sinistres ouverts de l'année multiplié par le coût moyen d'un sinistre sécheresse clos.

- Le montant des IBNR pour les communes sans arrêtés publiés correspond à la multiplication entre le nombre de maisons dans ces communes, la fréquence moyenne issue de la CCR et le coût moyen départemental d'un sinistre sécheresse clos.

La modélisation de la subsidence pour l'estimation de la charge ultime permet d'avoir une estimation plus robuste de la charge des IBNR et plus ajusté au portefeuille que ce que nous fournit le proxy précédent dont les hypothèses sont largement discutables.

1.4.3 Présentation des résultats d'études préexistantes

Ces dernières années, de nombreuses recherches et études ont été réalisées sur le sujet de la sécheresse et de ses conséquences. La Caisse Centrale de Réassurance (CCR) [6], l' association Mission Risques Naturels (MRN) et le Bureau de Recherches Géologiques et Minières (BRGM) sont les acteurs les plus actifs. Cette section présente les résultats et limites d'une étude sur le sujet.

Predicting drought and subsidence risk in France

En juillet 2021, un article scientifique a été publié par Arthur CHARPENTIER, Molly JAMES et Hani ALI [7] sur la modélisation de ce péril et détaille la méthodologie, la structure des données ainsi que la présentation des résultats de l'étude sur le portefeuille du réassureur Willis Re. Les paragraphes qui suivent dressent un résumé synthétique de la publication.

De manière instinctive, l'estimation de la charge sécheresse repose sur une modélisation en trois étapes :

- Prédire les communes concernées par la publication d'un arrêté au journal officiel (modèle de classification binaire).
- Sachant que la commune est concernée par un arrêté, il faut prédire la fréquence de sinistres (modèle de comptage de type poisson ou binomiale négatif).
- Pour chaque sinistre, il faut étudier la sévérité (modèle de régression gamma ou log-normale).

Dans leur publication, plusieurs méthodes ont été comparées pour l'estimation de la charge ultime. L'une d'entre elle repose sur une modélisation fréquence/sévérité via la théorie des modèles linéaires généralisés. En utilisant un modèle de régression à zéro-inflation (ZIP ou ZINB) [8], le modèle binaire et le modèle de comptage peuvent être réunis en un seul et unique modèle qui tient compte de la sur-proportion de sinistre nul dans le jeu de données.

6. Dans le dernier rapport scientifique publié par la CCR [CCR, 2021], [ECOTO *et al.*, 2021] présentent une modélisation du risque de subsidence via les *Super Learner*

7. Source : [CHARPENTIER *et al.*, 2021]

8. La référence suivante introduit les modèles à zéros inflation : [RAKOTOMALALA, 2012]

Données

La base de modélisation utilisée repose sur les données de risques et de sinistres d'un portefeuille MRH sur la période 2001 à 2018. L'exposition, le nombre de sinistres ainsi que leurs montants ont été agrégés par année à la maille INSEE. A cette base, ont été ajoutés des indicateurs d'intensité extrême de la sécheresse qui sont les suivants : *Extreme Standardized Precipitation Index (ESPI), Extreme Standardized Soil Temperature Index (ESSTI), Extreme Standardized Soil Wetness Index (ESSWI)*. La mention extrême signifie que l'indicateur retenu pour la commune α et l'année t correspond à sa valeur la plus défavorable. Ainsi, la valeur extrême de l'indice d'humidité des sols standardisés revient à appliquer la transformation suivante :

$$\text{ESSWI}_{\alpha,t} = min(\text{SSWI}_{\alpha,t,m}), m \in [1:12] \tag{1.1}$$

avec :

- α, une commune métropolitaine
- t, l'année de référence
- m, les mois de l'année

De plus il a également été ajouté la concentration des sols en argile sur la couche 0-20cm à partir d'une carte publiée par l'*European Soil Data Centre* (ESDAC). L'indice d'argile retenu pour une commune correspond à la concentration d'argile maximale des mailles qui l'intersecte. Un autre indicateur binaire a été ajouté au jeu de données, ce dernier vaut 1 si la commune a déjà fait une demande de reconnaisse par le passé et 0 sinon.

Résultats

Les graphiques et cartes de la figure 1.14a sont les résultats des prévisions de la fréquence sur l'année 2018 avec les différentes méthodes.

Sur la partie fréquence, le modèle souffre d'une mauvaise détection des communes respectant les critères d'éligibilité. Par conséquent la fréquence de sinistre est sur-estimé sur les zones non reconnues en l'état de catastrophe. Naturellement ces mauvaises prédictions se répercutent sur les coûts.

Nous pouvons également remarquer de très fortes ressemblances entre les prévisions des coûts en 2017 et en 2018 avec

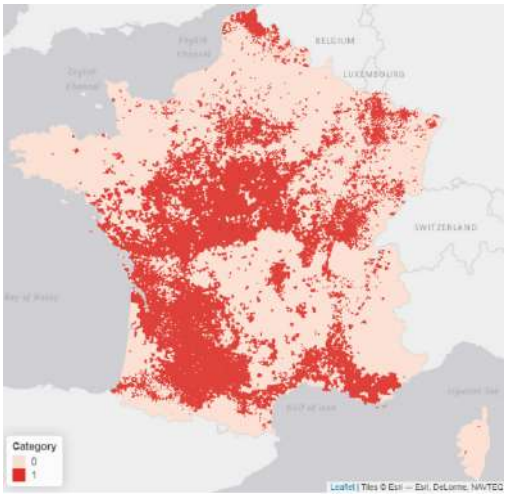
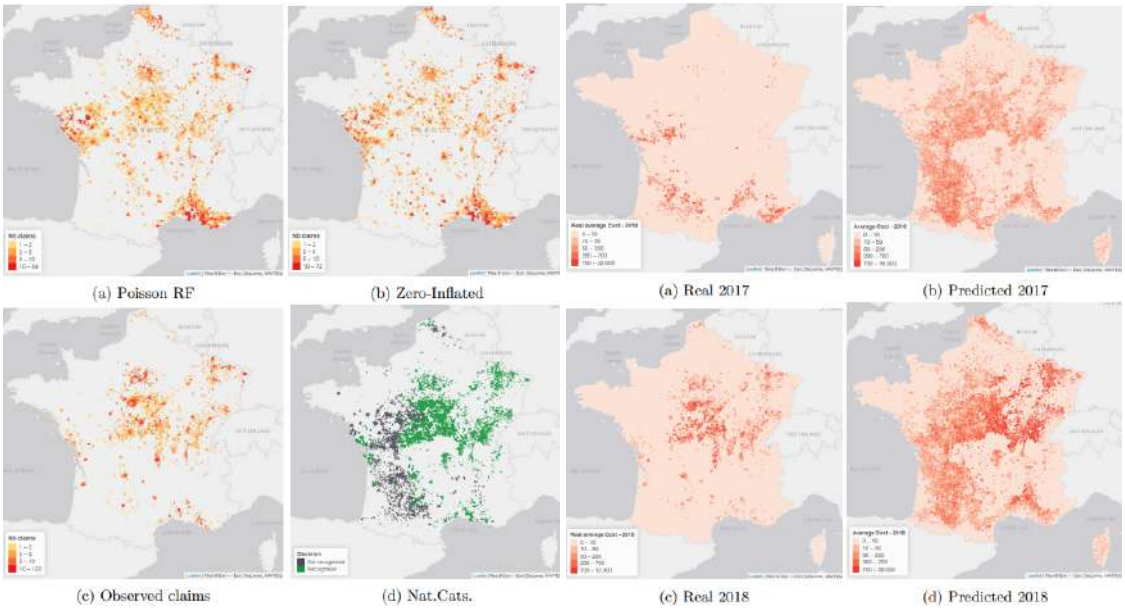


FIGURE 1.15 – Communes reconnues par le passé *Source : Arthur CHARPENTIER, Molly JAMES, Ani HALI*



(a) Prédictions de la fréquence (b) Prédictions de la charge de sinistre

FIGURE 1.14 – Résultats de l'étude - *Source : Arthur CHARPENTIER, Molly JAMES, Ani HALI*

la cartographie des communes ayant déjà réalisées une demande de reconnaissance dans le passé.

1.4.4 Aperçu du processus de modélisation

Étant donnée les faiblesses des modèles à zéro-inflation en ce qui concerne la prévision d'un nombre de sinistres nul lié à l'absence d'apparition du phénomène. Nous avons choisi d'adopter une modélisation qui s'articulera en trois étapes :

- Un modèle de classification dont l'objectif est d'identifier les communes et périodes reconnues Cat Nat.
- Un modèle fréquence qui, compte-tenu de la reconnaissance, prédit le nombre de sinistres dans le portefeuille.
- Un modèle sévérité pour prédire la charge associée.

Les résultats de cette méthodologie seront comparés avec une approche fréquence-sévérité plus traditionnelle afin de quantifier le gain engendré par l'utilisation d'un modèle de classification. La figure 1.16 résume synthétiquement la méthodologie proposée pour

cette étude.

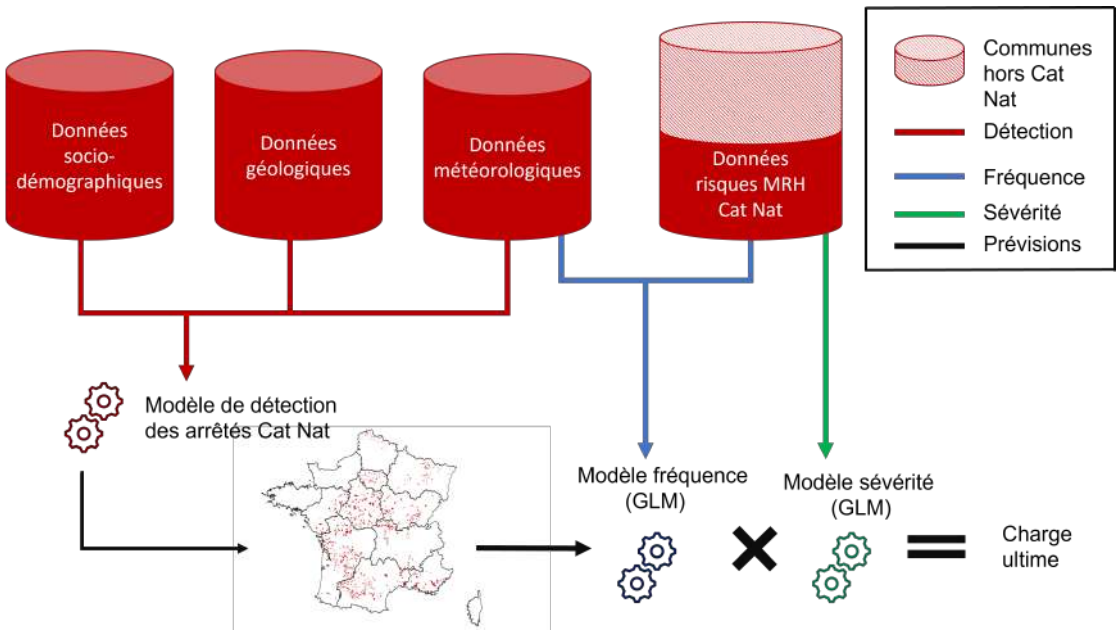


FIGURE 1.16 – Schéma de modélisation

Le modèle de reconnaissance

L'objectif du modèle de détection des arrêtés Cat Nat sécheresse est de résorber les faiblesses rencontrées par les modèles à zéro-inflation. Il permet d'éliminer la partie structurelle d'un nombre de sinistres nul. Ces derniers dépendent au cours du temps des critères de reconnaissance appliqués à l'époque.

Le modèle de détection repose uniquement sur des données provenant de l'*open data*. Comme le caractère anormal d'une sécheresse est apprécié par saison, nous avons naturellement fait le choix de construire une base à la maille INSEE, année et trimestre afin de pouvoir comparer dans l'espace et le temps les communes et saisons reconnues et prédites.

Les données utilisées contiennent :

- Des données météorologiques : les indices de sécheresses et la reproduction des critères de reconnaissance pour la version non uniforme du SWI.
- Des données géologiques, il s'agit de la part de la surface communale pour chaque couche d'exposition ou encore la concentration moyenne en argile.
- Des données socio-démographiques : la densité de population, le nombre de maisons individuelles pour chaque couche d'exposition à l'aléa ainsi que le nombre de demandes de reconnaissance passées.
- La variable cible binaire qui vaut 1 si la commune α est concernée par la publication d'un arrêté Cat Nat sécheresse sur l'année t ainsi que le trimestre s et 0 sinon.

Comme les données du SWI uniforme pour l'année t ne sont publiées par Météo France qu'une fois que la commission interministérielle a statué sur la période correspondante, le modèle de détection sera calibré en utilisant les données non-uniformes fournit par un centre de prévisions météorologiques européen.

D'un point de vue formel, un modèle de détection revient à prédire l'occurrence d'au moins un sinistre au sein d'une commune donnée ainsi que son éligibilité à la reconnaissance. Notons les évènements et variables aléatoires suivantes :

- $A_{\alpha,t,s} \in \{0,1\}$, la variable aléatoire désignant si la commune α est reconnue en l'état de catastrophe pour l'année t et la saison s.
- $N_{\alpha,t,s}$, la variable aléatoire désignant le nombre de sinistres lié au RGA pour la commune α au titre de l'année t et de la saison s.
- $E_{\alpha,t,s}$, l'évènement suivant : "La commune α est éligible vis-à-vis des critères de reconnaissance pour l'année t et la saison s."

Au cours d'une année donnée, nous cherchons à modéliser pour chaque commune α :

$$P(A_{\alpha,t,s} = 1) = P(N_{\alpha,t,s} > 0 \cap E_{\alpha,t,s})$$

Ensuite, un seuil sera appliqué à cette probabilité afin de déterminer l'appartenance à l'une des deux classes.

Modèle fréquence post-détection

Après avoir éliminé la part structurelle des zéro sinistres identifiés grâce au modèle de détection des arrêtés Cat Nat sécheresse, l'étape suivante consiste à modéliser la fréquence de sinistre sachant qu'un arrêté Cat Nat a été publié pour la commune.

Pour cela, nous allons construire un modèle de comptage de type Poisson à partir des données risques filtrés sur les communes et périodes reconnues en l'état de catastrophe. De plus, nous ajouterons les indices de sécheresse aux données risques. Ceci nous permettra d'estimer $E(N_{\alpha,t,s}|A_{\alpha,t,s} = 1)$.

La fonction de masse de la variable aléatoire $N_{\alpha,t,s}$ est modifiée de la manière suivante :

$$P(N_{\alpha,t,s} = k) = \begin{cases} P(A_{\alpha,t,s} = 0) + P(A_{\alpha,t,s} = 1) \cdot P(N_{\alpha,t,s} = 0|A_{\alpha,t,s} = 1) & \text{si } k = 0 \\ P(N_{\alpha,t,s} = k|A_{\alpha,t,s} = 1) \cdot P(A_{\alpha,t,s} = 1) & \text{si } k > 0 \end{cases}$$

En effet, en théorie nous avons $P(N_{\alpha,t,s} > 0|A_{\alpha,t,s} = 0) = 0$. La classification aura alors pour conséquence de modifier les probabilités $P(A_{\alpha,t,s} = 1)$ par les valeurs 0 ou 1.

Les résultats de cette modélisation de la fréquence en deux étapes seront comparés avec les résultats d'un modèle linéaire généralisé de type Poisson dont l'objectif est d'estimer directement le nombre de sinistres. Ce second modèle reposera sur les données risques augmentés des données utilisées pour le modèle de classification et ne tiendra pas compte de la sur-proportion de zéro sinistres dans notre base.

Chapitre 2

Méthodes et théorie

Au cours de ce chapitre, nous présenterons les différentes méthodes et concepts utilisés pour la réalisation de notre étude. Pour commencer, nous détaillerons plusieurs techniques de rééchantillonnage de données utiles pour rééquilibrer la base de notre modèle de détection.

Ensuite, nous définirons le principe de fonctionnement des forêts aléatoires et nous introduirons la théorie des modèles linéaires généralisés. A cette occasion, nous présenterons plusieurs métriques et indicateurs de comparaison adaptés aux différentes problématiques.

Enfin, nous passerons en revue la méthodologie de sélection de variable sur les différents modèles.

2.1 Les méthodes de rééchantillonnage

Au sein de la base de données utilisée pour le modèle de détection des arrêtés Cat Nat sécheresse, notre variable cible présente un déséquilibre des classes important. Avec seulement 1.2% de notre base contenant un arrêté favorable, l'apprentissage des méthodes supervisés devient plus difficile et certains algorithmes peinent à parvenir à une classification correcte. Dans cette section, nous allons présenter les différentes méthodes de rééchantillonnage utilisées afin de rééquilibrer notre jeu de données. L'impact de ces méthodes sur les métriques sera présenté dans la section résultat à la page 90.

2.1.1 Sur/sous-échantillonnage

Il existe dans la littérature de nombreuses méthodes de rééchantillonnage plus ou moins complexes. Avant d'introduire certaines méthodes plus sophistiquées, nous introduirons les méthodes les plus simplistes.

Sur-échantillonnage

Cette méthode consiste simplement à dupliquer les lignes de la classe minoritaire jusqu'à ce que la base soit parfaitement rééquilibrée. Cela permet à la classe minoritaire d'être mieux représentée lors de l'apprentissage. Pour ce faire, il suffit de tirer aléatoirement et avec remise autant d'individus minoritaires qu'il en faut pour atteindre l'équilibre des classes.

Avec cette méthode, nous ne perdons pas l'information contenue dans notre base puisqu'aucun individu n'est retiré. En revanche, la duplication des individus minoritaires ne renforce pas la frontière de décisions des méthodes d'apprentissage et le classifieur peut rencontrer les mêmes difficultés pour séparer les classes.

Sous-échantillonage

Le sous-échantillonage quant à lui ne duplique pas l'information mais en retire une partie ce qui a pour conséquence d'entraîner une perte de l'information contenue dans notre jeu de données au bénéfice d'un gain de temps lors de l'apprentissage de nos modèles. Plutôt que de dupliquer la classe minoritaire, il retire autant d'individus de la classe majoritaire qu'il en faudrait pour atteindre l'équilibre des classes. Cette méthode réduit drastiquement la taille de notre base puisque le nombre d'arrêté favorable est faible.

2.1.2 *Synthetic Minority Over-sampling TEchnique* (SMOTE)

La méthode SMOTE est une méthode alternative au sur-échantillonage. Plutôt que de dupliquer l'information des individus de la classe minoritaire, de nouveaux individus synthétiques sont créés. La création de ces nouveaux individus va venir renforcer la région/frontière de décision séparant les deux classes.

Pour cela, les individus synthétiques sont créés le long des segments liant un individu de la classe minoritaire avec ses k plus proches voisins. Ainsi, si nous souhaitons tripler l'effectif de la classe minoritaire (+200%) alors il suffit de créer 2 individus synthétiques par individus minoritaires. La figure 2.1 montre un exemple simpliste de création d'individus synthétique avec la méthode SMOTE.

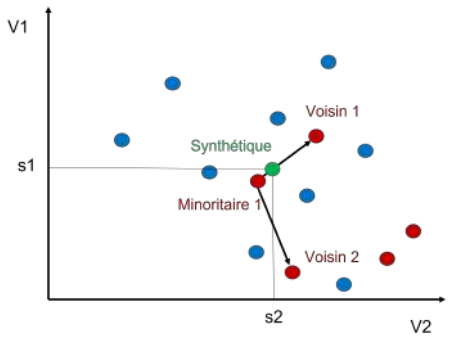


FIGURE 2.1 – Création d'un individu synthétique

Comme la méthode repose sur les distances euclidiennes entre deux points, il est préférable de centré et réduire notre jeu de données au préalable afin de résorber les problèmes d'échelle. De plus, l'algorithme de base ne traite que les variables numériques et un retraitement doit être effectué pour les variables discrètes. Enfin, si la création des individus synthétiques doit renforcer l'apprentissage, il peut avoir l'effet inverse lorsque les individus de la classe minoritaire sont éloignés ou lorsque que le nombre de plus proches voisins est mal choisi. Cela aura pour effet de bruiter les données comme nous le montre la figure 2.2.

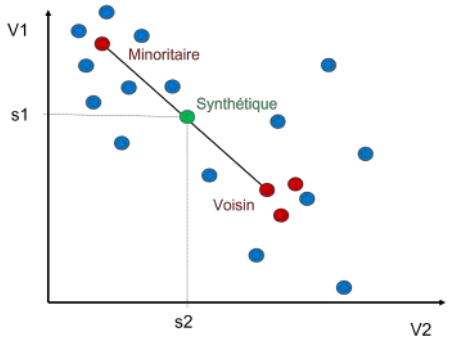


FIGURE 2.2 – Bruitage des données par l'algorithme SMOTE

2.1.3 *Random Over-Sampling Examples* (ROSE)

L'algorithme ROSE propose une manière différente de construire ces individus synthétiques dans le voisinnage de points observés.

Considérons un ensemble d'apprentissage T_n contenant n individus. Chacune des lignes de l'ensemble d'apprentissage contient la paire (x_i, y_i) avec :

- $x_i \in \mathcal{R}^p$, les p variables explicatives correspondant à la réalisation d'un vecteur aléatoire défini dans $\mathcal{R}^p$ de densité $f(x)$ inconnue.
- $y_i \in \{0, 1\}$, la variable réponse dont la modalité minoritaire est égale à 1.

Pour créer s individus de la classe minoritaire, la procédure ROSE consiste à répéter s fois les deux étapes suivantes :

- Tirer aléatoirement un individu avec remise parmi la classe minoritaire.
- Créer l'individu synthétique x^{new} à partir de $K_H(., x_i)$ une distribution de probabilité centré en x_i et de matrice de covariance H.

Le principe de ROSE est de créer les individus synthétiques dans le voisinnage des observations de la classe minoritaire. La forme de ce voisinnage est déterminée par les contours de K tandis que la largeur est fonction de la matrice de covariance H, les individus sont générés à partir d'une estimation de la densité conditionnelle $f(x|y = 1)$. La figure 2.3 donne un bref aperçu de la méthode à partir de données fictives.

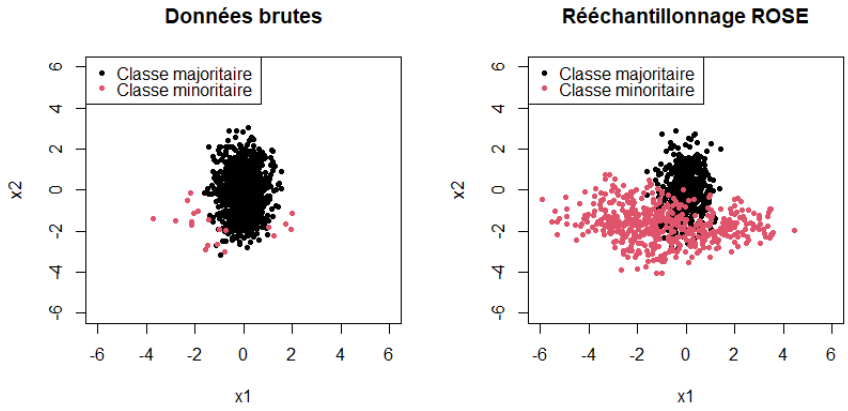


FIGURE 2.3 – Exemple fictif de rééchantillonnage avec ROSE

2.1.4 Conclusion

Les méthodes présentées ci-dessus ainsi que les résultats associés reposeront sur un rééquilibrage parfait du jeu de données c'est à dire avec autant de communes et trimestre reconnus que non reconnus. Cependant, il est possible d'ajuster le niveau de rééquilibrage de la base en créant moins d'individus synthétiques, en retirant moins d'informations ou encore en dupliquant moins d'individus.

Les méthodes de rééchantillonnage ont été implémentées via l'utilisation du package *caret* sur le langage de programmation *R*. Ce dernier nous permet d'appliquer ces méthodes sur les folds d'apprentissage de la validation croisée choisie et de récolter les métriques sur les folds de test qui n'ont pas été rééquilibrés.

2.2 Méthodes d'apprentissage supervisés

Cette section présente les différents méthodes d'apprentissage utilisées au cours de notre étude. A savoir :

- Les forêts aléatoires comme algorithme de classification.
- Les modèles linéaires généralisés pour estimer la fréquence et le coût des sinistres.

2.2.1 Les arbres de décision en classification

Les arbres de décision ont été inventés en 1984 par Breiman et col. et sont communément appelés CART pour *classification and regression tree*. Les arbres de classification reposent sur un partitionnement récursif du jeu de données à partir d'arbre de décision binaire. En considérant notre jeu de données comme la réalisation du vecteur aléatoire (X, Y), $X \in \mathbb{R}^p$, $Y \in \{0, 1\}$ avec p le nombre de variables explicatives, l'objectif est de classer les individus i dans l'une des deux classes $\{0, 1\}$ sur la base de critère à appliquer aux caractéristiques x_i de l'individu.

La première étape consiste à séparer l'espace X en deux sous-parties distinctes en choisissant une coupure de la forme $\{X^j \leq s\} \cup \{X^j > s\}$ ou $\{X^j = s\} \cup \{X^j \; != s\}$ si la variable est factorielle. Le choix de la coupure optimale s est déterminer en minimisant la fonction de coût suivante :

$$C(s, j) = \sum_{k=1}^{K} \hat{p}^k_{n_{1,-(j,s)}}(1 - \hat{p}^k_{n_{1,-(j,s)}}) + \sum_{k=1}^{K} \hat{p}^k_{n_{1,+(j,s)}}(1 - \hat{p}^k_{n_{1,+(j,s)}}) \tag{2.1}$$

où :

- $\hat{p}^k_{n_{1,-(j,s)}}$ est la proportion d'observations de la classe $k \in \{0, 1\}$ dans l'ensemble $n_{1,-(j,s)}$.
- $n_{1,-(j,s)} = \{i \in [1 : n], \mid x^j_i \leq s\}$ et $n_{1,+(j,s)} = \{i \in [1 : n], \mid x^j_i > s\}$

Cette fonction de coût n'est autre que l'indice d'impureté Gini, il représente la probabilité total qu'un individu choisit aléatoirement soit mal classé. cet indice est donc à minimiser.

Le processus est ainsi réitéré jusqu'à ce qu'un critère d'arrêt soit vérifier. Par exemple, on ne peut partitionner un noeud contenant trop peu d'observations. Le nombre d'observations minimal dans un noeud pour continuer la partition est un hyper-paramètre choisi par l'utilisateur (*min node size*). Une fois que le critère d'arrêt est vérifié, les noeuds correspondant aux derniers partitionnements sont appelés les feuilles de l'arbre de décision.

L'étape final consiste à élaguer l'arbre c'est à dire à rechercher le sous-arbre avec le meilleur compromis biais/variance issue de l'arbre maximale. En effet, l'arbre maximal a le biais le plus faible mais il est sujet au sur-apprentissage et risque de ne pas pouvoir se généraliser aux données de test ce qui en fait une faiblesse.

In fine une fois l'arbre élagué, les individus que l'on souhaite classer vont reparcourir l'arbre et ses décisions jusqu'à se retrouver dans l'une des feuilles. Dès lors, nous pouvons attribuer à chacun de ses individus la probabilité d'appartenir à l'une de classes $k \in \{0, 1\}$. Si l'on ne souhaite pas construire d'arbre de probabilité, les classes sont attribuées en utilisant le seuil de classification de 50%.

Les arbres de classification sont appréciés pour leur interprétation et lisibilité grâce à leur représentation graphique. De plus, contrairement aux méthodes de classification paramétriques comme la régression logistique, il est possible d'intégrer des prédicteurs corrélés dont la liaison avec la variable cible n'est pas forcément linéaire. Il a également l'avantage de ne pas être sensible à l'échelle des variables explicatives contrairement à d'autres algorithmes de machine learning.

En revanche, les performances sont fortement liés aux données d'apprentissage ce qui rend la méthode CART difficilement généralisable c'est pourquoi les arbres de classification sont utilisés comme une base pour d'autres méthodes plus sophistiquées comme les forêts aléatoires.

2.2.2 Les forêts aléatoires

Nous l'avons évoqué précédemment, malgré l'élagage des arbres, la méthode CART se généralise difficilement. Les forêts aléatoires répondent à ce problème de généralisation en introduisant de l'aléa dans le choix des individus utilisés pour l'apprentissage (bagging) ainsi que dans la sélection des variables utilisées.

Les hypers-paramètres des forêts aléatoires sont donc les suivantes :
- $max.depth$, la profondeur maximale des arbres qui composent la forêt.
- $ntrees$, le nombre d'arbres contenu dans la forêt aléatoire. C'est aussi le nombre d'échantillons *bootsrap* tirés aléatoirement dans le jeu de données (1 échantillon par arbres).
- $min.node.size$, le nombre d'instance minimale dans un noeud pour pouvoir effectuer une coupure. Ce paramètre guide également la profondeur des arbres car plus il est grand, moins les arbres sont profonds et inversement.
- $mtry$, le nombre de prédicteurs tirés aléatoirement parmi les p contenus dans le jeu de données pour la construction des arbres (1 tirage aléatoire par arbre).

Les forêts aléatoires constituent un panel d'arbres de classification binaire décorrélés les uns aux autres. A chaque tirage d'un échantillon *bootstrap* d'individus, un arbre de classification est construit en effectuant les coupures sur un ensemble restreint de $mtry$ variables parmi les p disponibles.

Les prévisions de la forêt pour un individu correspond ni plus ni moins à un vote majoritaire parmi les prévisions de chaque arbre composant la forêt en cas de classification.

$$\hat{\phi}(.) = \text{argmax}_k \left(\text{card}\{b : \hat{\phi}_b(.) = k\}\right)$$

où :
- $\hat{\phi}(.)$, la prévision de la forêt.
- $\hat{\phi}_b(.)$ la prévision de l'arbre b.
- $k \in \{0, 1\}$, la classe de la variable cible.

La probabilité d'appartenir à la classe $k \in \{0, 1\}$ correspond à la moyenne arithmétique des proportions d'individus de cette classe dans la feuille terminale de chaque arbre.

$$\hat{\phi}(.) = \frac{1}{B} \sum_{b=1}^{B} p_{b,k}$$

où :
- B correspond au nombre d'échantillon bootstrap.
- $p_{b,k}$, la proportion d'individu de la classe k dans la feuille terminale de l'arbre b.

2.3 Les Modèles Linéaires Généralisés

Dans les sections qui suivent, nous détaillerons la théorie utilisée pour les modèles linéaires généralisés (GLM). Ces méthodes de régression sont couramment utilisées en tarification pour l'estimation de la prime pure via une approche fréquence-sévérité. Il existe d'autres variantes du GLM avec des particularités intéressantes pour l'estimation de la charge ultime sur le péril sécheresse.

Pour les sections qui vont suivre, nous désignerons dans les équations les lettres majuscules comme étant des variables aléatoires et les lettres minuscules comme étant la réalisation de variables aléatoires à l'exception de $X \in \mathcal{M}_{n \times p}$ la matrice contenant les valeurs prisent par les p variables explicatives sur les n individus.

2.3.1 Le modèle linéaire gaussien

Avant de détailler les modèles GLM, nous allons revenir brièvement sur le modèle linéaire gaussien dont voici l'équation :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \epsilon_i \tag{2.2}$$

avec :
- $Y_i \in \mathbb{R}$ la variable cible.
- $x_{i,j}$ la valeur de la variable explicative numéro j pour le contrat i.
- $(\epsilon_i, \ldots, \epsilon_n)$ sont variables iid tel que $E[\epsilon_i] = 0$ et $Var(\epsilon_i) = \sigma^2 < \infty, \sigma \in \mathbb{R}$.

Nous pouvons réécrire les n équations sous la forme matricielle $Y = X\beta + \epsilon$ avec :
- $Y = (Y_1, \ldots, Y_n)^t \in R^n$
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t \in R^n$
- $\beta = (\beta_0, \ldots, \beta_p)^t \in R^{p+1}$

et :

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

Sous les hypothèses de distribution gaussienne des résidus $\epsilon \sim \mathcal{N}(\vec{0}, \sigma^2 I_n)$ il vient $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.

Estimation des paramètres

Les paramètres $\beta = (\beta_1, \ldots, \beta_n)$ du modèle linéaire gaussien sont estimés par la méthode des moindres carrés ordinaires. Nous cherchons les paramètres $(\hat{b}_0, \ldots, \hat{b}_p)$ minimisant l'écart quadratique entre la réalisation de notre variable cible y_i et la régression linéaire. Autrement dit, nous cherchons $\hat{b} = argmin_{b \in \mathbf{R}^{p+1}}(F(b))$ avec $F(b) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i,1} - \ldots - b_p x_{i,p})^2 = ||y - Xb||^2$. Après l'étude des points critiques ou en raisonnant en termes de projection orthogonale, nous pouvons déduire que $\hat{\beta} = (X'X)^{-1}X'Y$.

2.3.2 La théorie des GLM

Les modèles GLM ont été introduits par John Nelder et Robbert Wedderburn en 1972, elles font parties des méthodes de régression paramétriques couramment utilisés en assurance IARD pour la tarification et appréciées pour leur interprétabilité.

Pour les modèles linéaires généralisés, le support de loi de notre variable aléatoire cible Y peut ne pas être l'ensemble des réels comme cela était le cas précédemment avec le modèle linéaire gaussien.

Par exemple :
- $Y_i \in \{0, 1\}$ lorsque l'on souhaite faire de la classification comme prédire si la commune i est concernée par un arrêté Cat Nat sécheresse.
- $Y_i \in \mathbb{N}$ si l'on souhaite prédire le nombre de sinistres sécheresse.
- $Y_i \in \mathbb{R}^+$ lorsque l'on souhaite prédire le montant des sinistres.

Pour répondre à ce besoin, les GLM reposent sur 3 composantes essentielles :
- $(Y_1, \ldots, Y_n)$ est une suite de variable aléatoire *iid* dont chaque élément peut être modélisé par une loi appartenant à la famille exponentielle dont les paramètres dépendent des variables explicatives.
- Il existe une fonction de lien que nous noterons g permettant de lier l'espérance de la loi de Y_i avec les variables explicatives $xi_{i,.}$.
- $(X_1, \ldots, X_p)$ les valeurs déterministes prisent par ces p variables explicatives.

In fine, nous pouvons écrire l'équation générale du modèle GLM de cette façon :

$$g\left(E[Y_i|X_i = x_i]\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_p x_{i,p} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \tag{2.3}$$

Par défaut le logiciel de tarification Emblem, que nous utiliserons pour construire nos modèles, repose uniquement sur des variables qualitatives. Les variables quantitatives ont donc été discrétisées en variables factorielles ordonnées ou non. L'encodage des variables par le logiciel se fait sous la forme d'indicatrice aussi appelé *dummy* variables. Par conséquent, nous pouvons adapter l'équation générale précédente dans un cadre multi-factorielle et sans interactions :

$$\begin{aligned} g\left(E[Y_i|X_i = x_i]\right) &= \beta_0 + \sum_{j=1}^{q_1} \beta_j^1 \mathbf{1}_{\{x_{i,1}=j\}} + \ldots + \sum_{k=1}^{q_p} \beta_k^p \mathbf{1}_{\{x_{i,1}=k\}} \\ g\left(E[Y_i|X_i = x_i]\right) &= \beta_0 + \sum_{k=1}^{p} \sum_{j=1}^{q_k} \beta_j^k \mathbf{1}_{\{x_{i,k}=j\}} \end{aligned} \tag{2.4}$$

où :
- Y_i la variable aléatoire cible du i^{e} individu.
- $x_i \in \mathbb{R}^p$ les p caractéristiques de l'individu i.
- q_k le nombre de modalité de la k^{e} variable explicative.
- β_j^k le coefficient de régression linéaire associé à la j^{e} modalité de la k^{e} variable explicative.

La famille de lois exponentielles

Comme nous l'avons évoqué précédemment, la loi de la variable cible Y doit appartenir à la famille des lois exponentielles. Par définition, Y fait partie de cette famille de loi si nous pouvons écrire sa densité ou sa fonction de masse de la manière suivante :

$$f_{\theta,\phi}(y) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{2.5}$$

avec :
- $a(.)$, $b(.)$ et $c(.)$ des fonctions.
- θ le paramètre naturelle de la loi exponentielle qui est lié aux coefficients et donc aux valeurs prises par les variables explicatives.
- ϕ le paramètre de dispersion.

Le tableau ci-dessous récapitule les familles de lois ainsi que la valeur des paramètres ϕ et θ correspondants.

Loi	$\boldsymbol{\theta}$	$\boldsymbol{\phi}$	$\mathbf{a}(\phi)$	$\mathbf{b}(\theta)$	$\mathbf{c}(\mathbf{y},\phi)$
Gaussienne $\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	ϕ	$\theta^2/2$	-0,5(y^2/ϕ + $\ln(2\pi\phi)$)
Poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	1	1	$\exp(\theta)=\lambda$	-ln(y !)
Gamma $\Gamma(\alpha,\gamma)$	$-1/\gamma$	α^{-1}	ϕ	-ln(-θ)	$(1/\phi$ - 1)ln(y) - ln($\Gamma(1/\phi)$)
Binomiale Négative(r,p)	ln(p)	-	1	-$r$ln(p)	-
Binomiale (n,π)	$\ln(\frac{p}{1-p})$	1	1	$n\ln(1+\exp(\theta))$	$\ln(C_n^y)$

TABLE 2.1 – Tableau de lois appartenant à la famille exponentielle

Les fonctions de liens

Les fonctions de liens permettent de lier l'espérance de la loi de notre variable cible Y avec les variables explicatives déterministes. Si Y appartient à la famille de lois exponentielles alors nous pouvons montrer que :

$$E[Y_i] = b'(\theta_i) = \mu_i = g^{-1}(X_i\beta) \tag{2.6}$$
$$V[Y_i] = b''(\theta_i)\phi = b''(b'^{-1}(\mu_i))\phi = \phi V(\mu_i) \tag{2.7}$$

Nous pouvons remarquer que θ_i est fonction des paramètres du modèle. De manière générale, nous utilisons les fonctions de liens canoniques g tel que $g(E[Y_i]) = g(\mu_i) = \theta_i$ avec θ_i le paramètre naturel de la famille exponentielle. Par conséquent, nous pouvons en déduire que $g(.) = b'(.)^{-1}$.

Le tableau 2.2 récapitule les lois usuelles ainsi que les fonctions de lien canoniques associées. Bien que la fonction de lien canonique soit la fonction inverse pour la loi Gamma en pratique nous utilisons la fonction de lien logarithmique afin de se rapporter à un modèle multiplicatif.

En effet, si $g(x) = ln(x)$, alors il vient :

$$E[Y_i|X_i = x_i] = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j})$$

Dans ce cas, nous pouvons interpréter $\exp(\beta_j) - 1$ comme une majoration ou une minoration de la fréquence par rapport à la modalité de référence.

Support de loi	**Distribution**	**Fonction de lien canonique**
$Y_i \in \mathbb{R}$	Gaussien	g(x) = x
$Y_i \in \{0,1\}$	Bernouilli	$g(x) = ln(\frac{x}{1-x})$
$Y_i \in \mathbb{R}^+$	Gamma	$g(x) = \frac{1}{x}$
$Y_i \in \mathbb{N}$	Poisson	$g(x) = ln(x)$
$Y_i \in \mathbb{R}^+$	Log-normale	$g(x) = ln(x)$

TABLE 2.2 – Les fonctions de liens canoniques usuelles

Le V de Cramer

L'utilisation des modèles linéaires généralisés repose sur un jeu de prédicteurs non corrélés les uns par rapport aux autres. La présence de multicolinéarité a pour conséquence d'augmenter la variance des coefficients estimés et donc de les rendre instables voir non interprétables. En effet, certains coefficients pourront être non significatifs à tord et l'ajout ou le retrait d'une variable corrélée aura un impact sur la significativité des autres coefficients. Dans notre étude, les variables quantitatives ont été discrétisées de sorte à ce que l'ensemble du jeu de données soit factorielle. Le V de Cramer se base sur la statistique du test du χ^2 permettant de tester l'indépendance de deux variables.

Le V de Cramer pour les variables X_1 et X_2 est défini de la manière suivante :

$$V_{X_1X_2} = \sqrt{\frac{\chi^2_{X_1X_2}}{n \cdot min(l-1; c-1)}} \tag{2.8}$$

où :
- $n = \sum_{i,j} n_{i,j}$ correspond au nombre total d'observations.
- $n_{i,j}$ correspond au nombre d'observations où l'on retrouve la i^{e} modalité de la variable X_1 et la j^{e} modalité de la variable X_2.
- l, le nombre de modalités de X_1 et c le nombre de modalités de X_2.

La statistique du test du $\chi^2_{X_1X_2}$ est défini par :

$$\chi^2_{X_1X_2} = \sum_{i,j} \frac{(n_{i,j} - \frac{n_{i.} \cdot n_{.j}}{n})^2}{\frac{n_{i.} \cdot n_{.j}}{n}} \tag{2.9}$$

où :
- $n_{i.} = \sum_j n_{i,j}$ correspond au nombre d'individus présents sur la i^{e} modalité de la variable X_1.
- $n_{.j} = \sum_i n_{i,j}$ correspond au nombre d'individus présents sur la j^{e} modalité de la variable X_2.

Les valeurs prises par le V de Cramer se situent entre 0 et 1, plus celles-ci sont proches de 1 plus la relation entre X_1 et X_2 est forte. En nous référant au logiciel de tarification

Emblem, nous écarterons les variables pour lesquelles le V de Cramer est supérieur ou égale à 0,7.

Estimation des paramètres du modèle

L'estimation des paramètres se fait en maximisant la vraisemblance. Comme notre variable cible Y_i appartient à la famille de loi exponentielle nous pouvons écrire la vraisemblance associée à cette famille de lois.

$$L(y_1, \ldots, y_n, \theta_1, \ldots, \theta_n, \phi) = \prod_{i=1}^{n} f_{\theta_i,\phi}(y_i)$$
$$\ln(L(y_1, \ldots, y_n, \theta_1, \ldots, \theta_n, \phi)) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \tag{2.10}$$

Comme maximiser la vraisemblance revient à maximiser la log-vraisemblance $\mathcal{L}$, nous recherchons le maximum global de cette fonction concave. En prenant les notations suivantes :

- $\mu_i = E[Y_i] = b'(\theta_i)$
- $Var(Y_i) = b''(\theta_i)$
- $\eta_i = g(\mu_i) = X_i\beta$, le prédicteur linéaire théorique et X_i est la i^{e} ligne de la matrice des régresseurs.

Il vient pour j donnée :

$$\frac{\delta\mathcal{L}(y_1, \ldots, y_n, \theta_1, \ldots, \theta_n, \phi)}{\delta\beta_j} = 0 \leftrightarrow \sum_i \frac{\delta\mathcal{L}(y_i, \theta_i, \phi)}{\delta\beta_j} = 0$$

Pour i et j donnée,

$$\frac{\delta\mathcal{L}(y_i, \theta_i(\beta_j), \phi)}{\delta\beta_j} = \frac{\delta\mathcal{L}(y_i, \theta_i(\beta_j), \phi)}{\delta\theta_i(\beta_j)} \times \frac{\delta\theta_i(\beta_j)}{\delta\mu_i} \times \frac{\delta\mu_i}{\delta\eta_i} \times \frac{\delta\eta_i}{\delta\beta_j}$$

Or :

$$\frac{\delta\mathcal{L}(y_i, \theta_i)}{\delta\theta_i} = \frac{y_i - \mu_i}{a(\phi)}$$
$$\frac{\delta\theta_i}{\delta\mu_i} = \frac{1}{b''(\theta_i)}$$
$$\frac{\delta\eta_i}{\beta_j} = \frac{X_i\beta}{\delta\beta_j} = X_{i,j}$$
$$\frac{\delta\mu_i}{\delta\eta_i} = \frac{1}{g'(\mu_i)}$$

In fine pour $j = 1, \ldots, p$, nous obtenons les équations suivantes aussi appelées équations du score :

$$\sum_i \frac{\delta \mathcal{L}_i}{\delta \beta_j} = \sum_i \frac{\delta \mu_i}{\delta \eta_i} \times \frac{y_i - \mu_i}{a(\phi) b''(\theta_i)} X_{i,j} \tag{2.11}$$

Ces équations ne sont pas solvables analytiquement c'est pourquoi nous estimons les coefficients à l'aide d'algorithme numérique.

La méthode de ***Newton-Raphson*** est une méthode itérative pour trouver la racine d'une fonction réelle $f(x)$ sur les points de l'itération. Nous commençons par choisir un point de départ x_0 puis nous considérons la fonction $f(x)$ comme égale à l'équation de la tangeante en ce point $f(x) \simeq f(x_0) + f'(x_0)(x - x_0)$. Dès lors il est aisé de déterminer la racine de l'équation de la tangeante qui sera le point de départ de l'itération suivante. Ainsi, nous répétons le processus suivant jusqu'à convergence : $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$.

En appliquant cette méthode pour déterminer les racines des équations du score, le processus itératif devient :

$$\beta^{(k+1)} = \beta^{(k)} - Jacob_S\left(\beta^{(k)}\right)^{-1} S\left(\beta^{(k)}\right)$$

avec :

- k le pas de l'itération
- $Jacob_S\left(\beta^{(k)}\right)$, la matrice jacobienne des équations S évaluée en $\beta^{(k)}$.
- $d_k = -Jacob_S\left(\beta^{(k)}\right)^{-1} S\left(\beta^{(k)}\right)$ la direction de la descente du gradient à l'itération k pour l'équation du score S.

Les différents résidus

Les résidus sont utiles pour représenter la qualité d'apprentissage du modèle GLM à nos données et identifier les valeurs anormales nécessitant des investigations supplémentaires. Les résidus les plus utilisés sont les résidus de Pearson et les résidus de déviance formulé de la manière suivante :

$$r_p = \frac{y - \hat{y}}{\sqrt{V(\hat{y})}}$$

$$r_d = sign(y - \hat{y})\sqrt{d}$$

où :

- r_p sont les résidus de pearson.
- y et $\hat{y}$ sont respectivement la variable cible observée et prédite par le modèle
- d la contribution de l'individu à la déviance.

Les résidus de déviance se distingue des résidus de Pearson par leur symétrie ce qui les rend plus facile à interpréter. Un bon modèle est un modèle qui prédit correctement la fréquence de sinistre. Par conséquent les résidus doivent être centrés et le moins dispersés

possibles. A partir de ces deux résidus, nous pouvons exprimer leurs valeurs standardisés qui ont l'avantage de compenser la corrélation entre les valeurs observées et les valeurs prédites.

$$r_{sp} = \frac{r_p}{\sqrt{(1-h)}}$$
$$r_{sd} = \frac{r_d}{\sqrt{(1-h)}}$$

où :

- r_p et r_d sont respectivement les résidus de Pearson et les résidus de déviance.
- h est la diagonale de la matrice de projection de la variable réponse sur le sous-espace vectoriel engendré par l'ensemble de variables explicatives $h = diag(H) = diag\left(X(X'X)^{-1}X'\right)$.

Sur Emblem, les résidus sont représentés graphiquement selon une transformation des prédictions afin de stabiliser la variance des résidus en fonction de la distribution théorique de la variable cible. Les transformations appliquées aux prévisions sont les suivantes :

- L'axe des abscisses représente $2\sqrt{\hat{y}}$ pour une distribution de Poisson.
- L'axe des abscisses représente $2\ln(\hat{y})$ pour une distribution de Gamma.

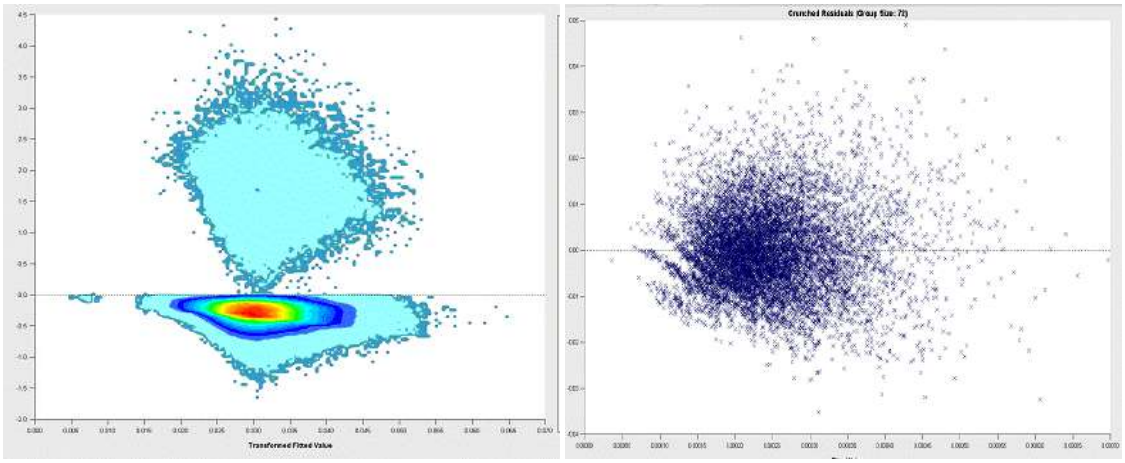
Les résidus « crunched » consiste à calculer les résidus pour un groupuscule d'individus plutôt que pour chacun d'entre eux. Nous les utiliserons pour les modèles fréquences car la plupart des profils de risque n'ont pas de sinistres. Ceci a pour conséquence de créer plusieurs clusters de résidus correspondant aux différentes valeurs prises par la variable cible.

$$r_c = \frac{\sum_i y_i - \hat{y}_i}{\sqrt{V(\hat{y}_i)}}$$

où :

- r_c correspond à la valeur du résidu pour un regroupement c.
- $i \in \{1, , n\}$ avec n le nombre d'individus dans le regroupement.

Les regroupements sont formés en ordonnant les prédictions de sorte à ce que les individus d'un groupe ait des prédictions homogènes. La figure 2.4 compare les résidus de déviance standardisés avec les « crunched » résidus.



(a) Résidus de déviance standardisés (b) Résidus « crunched »

FIGURE 2.4 – Comparaison entre les différents résidus

2.4 La sélection de variable

Cette section présente les méthodes de sélection de variable utilisées pour le modèle de détection ainsi que pour les modèles linéaires généralisés.

En ce qui concerne le modèle de détection, la méthode d'apprentissage est non-paramétrique. La sélection de variable permet de trouver un sous-ensemble de prédicteurs offrant des performances équivalentes au modèle complet. Ceci permet de réduire la volumétrie de nos données et d'omettre les variables non pertinentes de notre modèle.

Pour les modèles linéaires généralisés, le nombre de coefficients à estimer dépend du nombre de variables ainsi que du nombre de modalités, la sélection de variables permet d'ajouter les prédicteurs les plus discriminants au modèle.

2.4.1 Les méthodes *backwards* et *forwards*

Les méthodes *backwards* et *forwards* font partie des procédures fréquemment utilisées pour la sélection de variables des modèles linéaires généralisés mais également pour d'autres types de modèle. Les sous-sections qui suivent détaillent la méthodologie appliquée pour les modèles de fréquence et de sévérité.

La méthode *forward*

La sélection de variables *forward* consiste à ajouter une à une les variables, qui ne sont pas déjà incluses, au modèle de référence. Cela permet de quantifier l'impact de l'ajout de la variable explicative dans le modèle au travers des critères de comparaison

comme l'AIC, le BIC ou la déviance.

Dans cette étude, nous itérons la sélection *forward* pas à pas en commençant par un modèle vierge. Lors de chaque étape, la variable permettant de diminuer le plus l'AIC vis-à-vis du modèle précédent est ajouté au modèle. Nous regardons dans un premier temps la métrique AIC car elle est moins pénalisante envers le nombre de coefficients à estimer. En revanche il est indispensable que cette variable une fois simplifiée (regroupement de modalité) diminue les métriques d'AIC, de BIC et de déviance. Si ce n'est pas le cas alors la variable n'est pas sélectionnée et nous pouvons recommencer l'opération sur la variable suivante.

Une fois que le modèle est à priori stable, c'est à dire lorsqu'il n'y a plus de variables faisant baisser l'AIC avec la sélection forward, nous tentons d'ajouter les variables restantes et de les simplifier. Si post-simplification, la variable améliore les métriques alors elle est incluse au modèle sinon elle est rejetée. A la fin de ce processus, le modèle est définitivement stable et il n'y a plus de variable à ajouter.

La sélection *backward*

La sélection *backward* est une méthode qui consiste à enlever une à une les variables d'un modèle de référence. Cette méthode permet de quantifier le retrait de la variable du modèle sur les métriques. Si le modèle s'améliore lorsque nous retirons la variable alors cette dernière n'est pas pertinente et peut être retirée.

Une fois que nous avons le modèle stable, nous y appliquons une sélection *backward*. Comme les variables contenues dans le modèle stable ont été simplifiées, nous retirerons celles faisant diminuer le BIC car ce critère est le plus pénalisant vis-à-vis du nombre de paramètre à estimer.

Des tolérances peuvent s'appliquer lorsque la baisse du critère BIC est très faible et que l'AIC ainsi que la déviance se détériore en retirant la variable du modèle.

2.4.2 *Recursive feature elimination* (RFE)

L'algorithme RFE (*Recursive feature elimination*) fait partie des méthodes dites « enveloppe ». Elles se distinguent des méthodes « filtres » car elle utilise l'algorithme d'apprentissage pour trouver le sous-ensemble de variables explicatives optimal. Le principe de fonctionnement de l'algorithme RFE repose sur l'élimination *backward* des variables explicatives les moins importantes au sens de Gini et cela de manière récursive. Cette méthode de sélection de variables a été appliquée pour le modèle de détection des arrêtés Cat Nat.

Définissons $S = (S_0, S_1, \ldots, S_n)$ tel que $S_0 > S_1 > S_2 > ... > S_n$, une séquence décrivant le nombre de variables explicatives de chaque sous-ensemble à tester. Par défi-

nition $S_0 = p$ le nombre total de variables explicatives que contient notre jeu de donnée. L'algorithme RFE est détaillé de la manière suivante :

Algorithme RFE

- Pour chaque itération de la validation croisée :
  1. Partition des données en base d'apprentissage et de test.
  2. Apprentissage du modèle avec l'ensemble des variables explicatives (modèle complet).
  3. Calcul des métriques sur la base test.
  4. Calcul de l'importance des variables et attribution des rangs pour les attributs.
  5. Pour $i \in \{1, \ldots, n\}$:
     - Garder les S_i plus importantes variables explicatives.
     - Apprentissage du modèle en utilisant les S_i attributs restants.
     - Calculer les performances sur la base de validation de la partition.
     - Recalculer les rangs des attributs par importance. (Optionnel)
- Calculer la performance moyenne sur les itérations de la validation croisée pour chaque séquence S_i.
- Déterminer le nombre approprié d'attributs du modèle final.
- Estimer la liste finale des prédicteurs.
- Ajuster le modèle avec les variables sélectionnées sur les données choisies.

La taille du sous-ensemble choisi correspond par défaut à celle maximisant la métrique de performance. Toutefois nous pouvons appliquer une certaine tolérance afin de retenir un sous-ensemble plus petit et aussi performant.

$$S_{tol} = \inf \left\{ S_i \mid \alpha \geq \frac{AUC^{PR}_{optimal} - AUC^{PR}_{S_i}}{AUC^{PR}_{optimal}} \right\}$$

où :

- AUC^{PR} désigne l'aire sous la courbe précision-rappel.
- $\alpha \in [0 : 1]$, le paramètre de tolérance qui doit être proche de 0 afin de ne pas trop réduire l'écart entre les performances optimales et celles retenues.

L'algorithme laisse le choix à l'utilisateur de recalculer ou non les rangs d'importance des variables à chaque itération *backward*. Toutefois, les performances seraient moindres si l'on recalcule les rangs des attributs à chaque itération pour les forêts aléatoires [1]. Par conséquent, les rangs ne seront calculés qu'une seule fois par itérations de la validation croisée.

Comme les rangs initiaux ne sont pas identiques pour chacune des itérations de la validation croisée, les S_{tol} variables optimales diffèrent d'une itération à l'autre. Dans

1. Application of Breiman's Forêt Aléatoire to Modeling Structure-Activity Relationships of Pharmaceutical Molecules, Vladimir Svetnik,Andy Liaw,Christopher Tong,Ting Wang

ce cas, nous calculons l'importance moyenne et nous retenons les S_{tol} variables les plus importantes.

Cet algorithme a été modifié afin d'y appliquer une implémentation plus rapide des forêts aléatoires en utilisant la fonction *ranger* sur le langage *R*. De plus afin de tenir compte du déséquilibre de classe dans nos données, les métriques basées sur la matrice de confusion ont été compilées avec un seuil de classification arbitraire de 20%.

2.5 Métriques de performance et critères de comparaison

En fonction de notre problématique, de la structure de nos données ainsi que de la typologie d'algorithme utilisé, nous avons besoin d'introduire diverses métriques de performances et critères de comparaison afin de choisir le modèle offrant la meilleure capacité de généralisation et les meilleures performances.

Cette section présente l'ensemble des métriques et critères utilisés lors de la détection des arrêtés Cat Nat, la modélisation de la fréquence et de la sévérité.

2.5.1 Matrice de confusion, précision, rappel et F_1 score

Avant de présenter les métriques utilisées pour évaluer les performances du modèle de détection, il est nécessaire d'introduire la matrice de confusion.

La matrice de confusion est une matrice 2×2 contenant les effectifs croisés entre la variable cible originale $y \in \{0, 1\}$ et la classification faite par notre modèle $\hat{y} \in \{0, 1\}$. En règle générale, les colonnes de cette matrice sont utilisées pour représenter la variable observée tandis que les lignes représentent la variable prédite.

Notons $C \in M_{2\times 2}$ la matrice de confusion défini comme :

$$C = \begin{pmatrix} TN & FN \\ FP & TP \end{pmatrix}$$

avec :

- TN (*True Négative*) : les vrais négatifs correspondent à l'absence d'arrêtés correctement prédit $y_i = \hat{y}_i = 0$. La modalité majoritaire est correctement prédite par le modèle.
- FN (*False Negative*) : le modèle de classification n'a pas détecté l'arrêté Cat Nat donc $y_i = 1,\ \hat{y}_i = 0$, la prédiction du classifieur est incorrecte.
- FP (*False Positive*) : le classifieur prédit un arrêté à tord, $y_i = 0,\ \hat{y}_i = 1$.
- TP (*True Positive*) : le classifieur a bien détecté l'arrêté, $y_i = \ \hat{y}_i = 1$.

Dans notre cas, nous utilisons les forêts aléatoires pour effectuer la classification. Cette méthode retourne pour chaque individu i, la probabilité p_i qu'un arrêté soit pu-

blié. Par défaut, le seuil de classification correspond à 50% de sorte que $\hat{y}_i = \mathbf{1}_{p_i \geq 50\%}$. Cependant, ce choix n'est pas toujours optimal et sera discuté ultérieurement.

Taux d'erreur

Le taux d'erreur est défini comme la proportion de prédiction incorrecte de la classification soit :

$$\text{erreur} = \frac{FP + FN}{FP + TP + FN + TN}$$

Cette métrique n'est pas envisageable lorsque le jeu de données est déséquilibré car si le modèle ne prédit aucun arrêté alors le taux d'erreur est aussi faible que la proportion d'individus de la classe minoritaire soit 1,2% dans notre cas.

Nous avons donc besoin d'une métrique de performance qui tienne compte du déséquilibre de classe et offre un compromis entre les faux positifs et les faux négatifs. Dans l'idéal, quitte à ce que le classifieur se trompe, nous aimerions qu'il y ait autant de faux négatifs que de faux positifs. Dans cas précis, le nombre total d'arrêtés prédits correspondra au nombre total d'arrêtés observés dans notre base.

Ensuite nous assurer que ces effets compensatoires ont lieu dans une zone géographique proche de nos observations.

Rappel, précision et F_1score

Le rappel correspond au rapport entre le nombre de prédictions positives correctes (TP) et le nombre d'observations positives (TP+FN) soit :

$$\text{Rappel} = \frac{TP}{TP + FN}$$

La valeur du rappel est comprise entre 0 et 1 et s'interprète comme la proportion de classification correcte parmi la classe minoritaire observée.

La précision correspond au rapport entre le nombre de prédictions positives correctes (TP) et le nombre de prédictions positives (TP+FP) soit :

$$\text{Précision} = \frac{TP}{TP + FP}$$

La valeur de la précision est comprise entre 0 et 1 et s'interprète comme la proportion de classification correcte parmi la classe minoritaire prédite.

A elles seules, le rappel et la précision ne permettent pas d'apprécier la qualité du modèle car un modèle ne prédisant que des arrêtés aura un niveau de rappel égale à 1 mais la précision sera proche de 0 et inversement.

Le niveau de ces deux métriques ont des tendances opposées en fonction du seuil de classification, c'est pourquoi nous introduisons le F_β score. Il correspond à la moyenne harmonique entre le rappel et la précision lorsque $\beta = 1$.

$$\begin{aligned} \text{F}_\beta\text{-score} &= (1+\beta^2) \cdot \frac{\text{précision} \cdot \text{rappel}}{(\beta^2 \cdot \text{précision}) + \text{rappel}} \\ &= \frac{TP}{TP + \frac{1}{1+\beta^2}(\beta^2 FN + FP)} \end{aligned}$$

- Lorsque $\beta \geq 1$, nous accordons davantage de poids aux faux négatifs donc au niveau du rappel.
- Lorsque $\beta \leq 1$, nous accordons davantage de poids aux faux positifs donc au niveau de la précision.
- Lorsque $\beta = 1$, nous accordons autant d'importance aux faux positifs qu'aux faux négatifs.

Dans notre problématique, nous accordons autant d'importance aux arrêtés prédits à tord que les arrêtés non détectés. L'objectif est également de prédire si possible le même nombre d'arrêtés $TP + FP \simeq TP + FN$. Lorsque $\beta = 1$, nous pouvons définir le F_1 score :

$$\begin{aligned} F_1\text{score} &= 2 \cdot \frac{\text{rappel} \cdot \text{précision}}{\text{précision} + \text{rappel}} \\ &= \frac{TP}{TP + \frac{1}{2}(FN + FP)} \end{aligned}$$

Avec une valeur de 50% le modèle fait deux erreurs (FP ou FN) pour une prédiction positive correcte.

La figure 2.5 donne un aperçu des courbes de rappel, de précision et de F_1 score. A la gauche de l'optimum, le modèle prédit bien les vrais positifs mais se trompe sur les négatifs. Autrement dit, le niveau de rappel est élevé alors que la précision est faible donc le modèle prédit trop d'arrêtés. A l'inverse, à la droite de l'optimum, le modèle prédit mieux les négatifs mais détecte moins bien les positifs donc le modèle prédit trop peu d'arrêté.

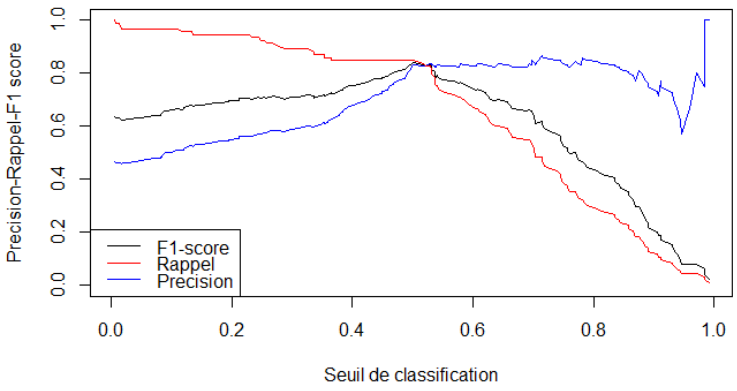


FIGURE 2.5 – Illustration des courbes de F1 score

La métrique de F_1 score permet de comparer les performances des différents classifieurs mais il est également important de comparer ces modèles de classification avec un modèle non informatif. Nous définissons un modèle aléatoire lorsque les individus de la classe majoritaire ont la même distribution de probabilité que les individus de la classe minoritaires.

Supposons que notre jeu de données contient une proportion $p = \frac{TP+FN}{TP+TN+FP+FN}$ d'individus de la classe minoritaire. En fonction de notre seuil de classification noté s, nous obtenons par le classifieur un taux de prédiction positives $q(s) = \frac{TP+FP}{TP+FP+FN+TN}$. Alors nous pouvons réécrire les notations précédentes en fonction de p et de q :

- $TN = (1-p) \cdot (1-q)$
- $TP = p \cdot q$
- $FN = p \cdot (1-q)$
- $FP = (1-p) \cdot q$

Par conséquent :

$$\begin{aligned} F1score &= \frac{TP}{TP + \frac{1}{2}(FN+FP)} \\ &= \frac{2 \cdot p \cdot q}{p+q} \end{aligned}$$

où :

- $p \in [0-1]$ est une constante
- $q = f(s)$ est fonction de notre seuil de classification.

Dans un modèle non-informatif uniforme, les prédictions du modèle peuvent être simulées par un tirage uniforme. Dans ce cas la précision du modèle aléatoire vaut p car :

$$\text{Précision} = \frac{TP}{TP+FP} = \frac{p \cdot q}{p \cdot q + (1-p) \cdot q} = p$$

.

Comme p est constant, maximiser le F_1 score revient à maximiser le rappel et celui atteint son maximum en 1 pour le seuil de classification 0. La valeur maximale du score F_1 pour un modèle aléatoire vaut donc :

$$F_1 score = \frac{2 \cdot p}{p + 1}$$

Cette valeur peut alors être utilisée comme baseline afin de comparer notre modèle avec l'aléa.

Courbe précision-rappel et AUC

L'inconvénient majeur du F_1 score est qu'il se base sur la matrice de confusion, il est donc sensible au seuil de classification.

Afin de remédier à cette problématique, nous introduisons cette fois-ci une mesure indépendante du seuil de classification mais qui reste sensible aux déséquilibres de classes. L'aire sous la courbe précision/rappel (AUC-PR) consiste de part son nom à retenir l'aire sous la courbe précision/rappel comme mesure de qualité du modèle.

La figure 2.6 compare le modèle parfait avec le modèle aléatoire pour un exemple fictif avec une proportion d'individus de la classe minoritaire égale à 30%. L'AUC-PR d'un modèle parfait est toujours égale à 1 tandis que l'AUC-PR d'un modèle aléatoire est constant au taux de déséquilibre p dans notre base. Nous pouvons également en déduire un ratio de performance qui correspond à :

$$\text{Ratio de performance} = \frac{AUC_{PR}}{p}$$

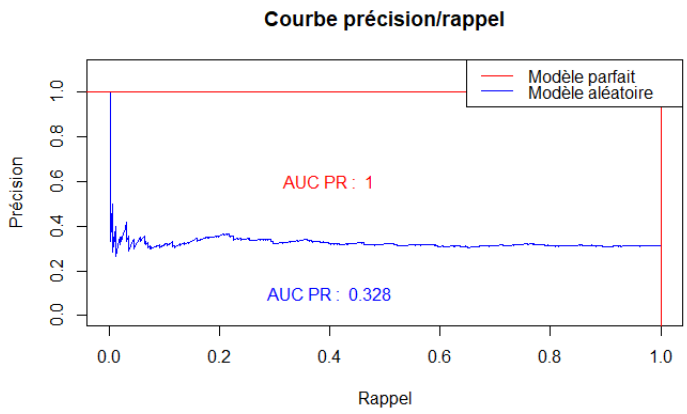


FIGURE 2.6 – Exemple de courbe précision-rappel

2.5.2 Critère d'information bayésien & déviance

Pour prédire le nombre de sinistres et les coûts associés, nous nous sommes basés sur une approche commune en utilisant les modèles linéaires généralisés (GLM). Contrairement aux forêts aléatoires, les GLM sont des modèles paramétriques. A chaque variables et modalités, un coefficient doit être estimé c'est pourquoi nous devons introduire de nouvelles métriques permettant de trouver un compromis entre efficacité et complexité du modèle. C'est le cas de l'AIC pour « *Akaike information Criterion* » et du BIC « *Bayesian Information Criterion* » :

$$AIC = 2k - 2\ln(L(\hat{\theta})) \tag{2.12}$$

$$BIC = k\ln(N) - 2\ln(L(\hat{\theta})) \tag{2.13}$$

avec :
- k : le nombre de paramètres du modèle.
- L : la vraisemblance du modèle avec l'ensemble des paramètres $\hat{\theta}$ estimés par un algorithme numérique.
- N : le nombre d'observations de l'échantillon.

Comme nous le montre ces équations, les deux critères offrent un compromis entre une grande vraisemblance et un petit nombre de paramètres à estimer. L'objectif est donc d'avoir une valeur de l'AIC et du BIC la plus faible possible. Le BIC se distingue de l'AIC par une pénalisation plus forte sur le nombre de paramètre à estimer. C'est la raison pour laquelle, nous préférerons l'AIC lors de la sélection *forward* (avant la simplification des variables), puis le BIC lors de la sélection *backward* (post-simplification des variables).

La déviance

La déviance est un critère de qualité globale du modèle qui se définit comme un écart entre la log-vraisemblance du modèle et la log vraisemblance du modèle saturé.

$$D = 2(l - l_{\text{saturé}}) \tag{2.14}$$

où :
- $l_{\text{saturé}}$ correspond à la log-vraisemblance du modèle saturé ou parfait.
- $l = \ln(L(\hat{\theta}))$ correspond à la log-vraisemblance du modèle de référence avec l'ensemble de paramètres $\hat{\theta}$ estimés.

La déviance est donc un indicateur à minimiser puisque plus nous sommes proche de la vraisemblance du modèle saturé, meilleur est le modèle.

2.5.3 Courbe de gain et l'indice de Gini

La courbe de gain ou courbe *lift* cumulé est une aide visuel pour apprécier la performance prédictive d'un modèle. Il consiste à tracer le cumul de sinistre observé en fonction

du cumul d'exposition trié par ordre décroissant de nos prévisions $\hat{y}$. La figure 2.7 montre un exemple fictif de courbe de gain.

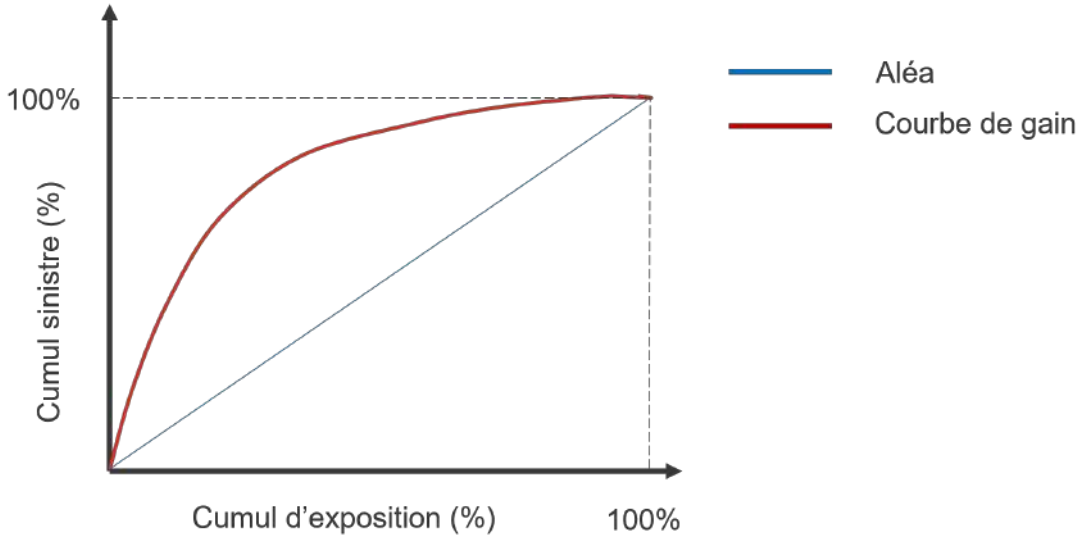


FIGURE 2.7 – Courbe de gain

Ce type de graphique permet de comparer les modèles GLM entre eux et vérifier l'absence de sur ou sous-apprentissage du modèle au travers d'une métrique qui est l'indice de Gini défini comme :

$$\text{Gini} = 2 \cdot AUC - 1$$

Si la valeur de l'indice de Gini est semblable sur la base d'apprentissage et de validation alors le modèle ne sur-apprend pas ni ne sous-apprend.

Chapitre 3

Construction et analyse des bases de données

Au cours de ce chapitre, nous présenterons et analyserons les différentes sources de données utilisées pour nos modèles. Nous commencerons par introduire les attributs relatifs à la présence d'argile dans le sol ainsi que plusieurs indicateurs d'exposition du bâti à l'aléa.

Après avoir détaillé la méthodologie de calcul des indices de sécheresse, nous les comparerons avec l'éligibilité des communes vis-à-vis des critères de reconnaissance. De la même façon, nous analyserons les disparités entre les données d'humidité des sols uniforme et non uniforme.

Enfin, nous présenterons la base des contrats contenant les données de risques ainsi que la base sinistre qui contient les informations sur l'état du dossier et le montant des sinistres.

3.1 Les données géotechniques

Depuis son intégration au régime Cat Nat, le RGA intègre un critère géotechnique reposant sur la part de sa surface communale en argile avéré. Au cours des dernières décennies, les cartographies répertoriant les lentilles argileuses sur le territoire national ont évolué. Au cours de cette section, nous présenterons à la fois l'évolution des cartographies réglementaires ainsi qu'une cartographie de l'« *European Soil Data Centre* » (ESDAC) portant sur la concentration des sols superficiels en argile. Ceci permettra d'introduire des indices de vulnérabilité et d'exposition du bâti à l'aléa.

3.1.1 La carte d'aléa du BRGM

La carte d'aléa ou susceptibilité est une cartographie du territoire métropolitain initiée dans les années 1990 à l'échelle départementale pour les départements les plus affectés. Le plan d'étude a été étendu à l'ensemble du territoire métropolitain après la sécheresse

estivale survenue au cours de l'année 2003 qui a causé de nombreux dégâts matériels.

Cette première carte hiérarchise le territoire en 4 zones d'aléa (sans aléa, aléa faible, moyen et fort) au format 1/50 000. Cette graduation est fonction de trois critères : la nature lithologique, la composition minéralogique et le comportement géotechnique des sols, c'est à dire leurs propriétés mécaniques.

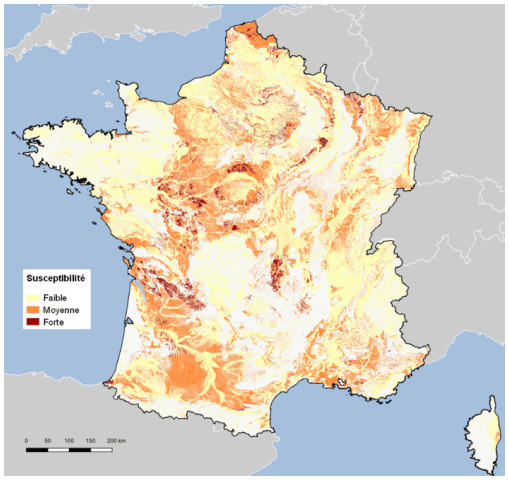


FIGURE 3.1 – Carte de susceptibilité *Source - BRGM, Géorisques*

Les zones d'aléa non renseignés, faibles, moyens et forts représentent respectivement 37%, 42%, 19% et 2% du territoire.

Cependant, cette carte n'exclue pas la présence de lentilles argileuses à une échelle plus locale autrement dit certains sinistres peuvent se produire sur une zone dépourvue a priori d'aléa. Ceci nous conduit à introduire la seconde version de cette cartographie qui pallie à ce biais.

3.1.2 La carte d'exposition du BRGM

La carte d'exposition au RGA, disponible depuis 2020, vient en remplacement de la carte d'aléa qui identifiait la susceptibilité des terrains au phénomène.

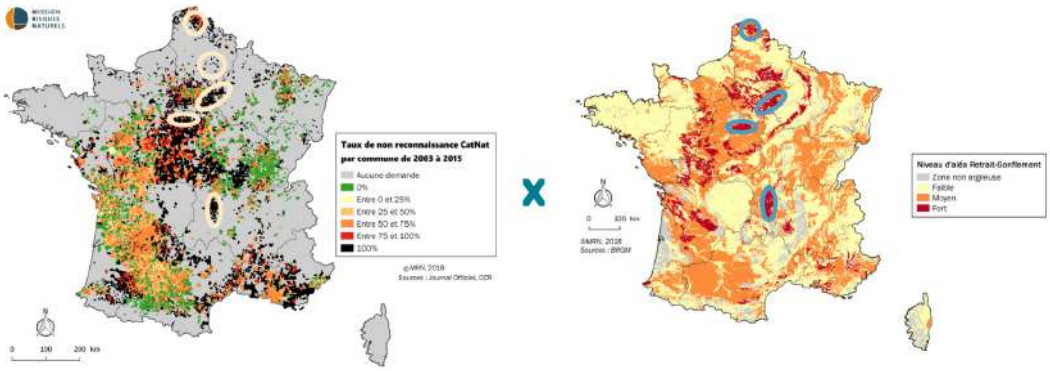


FIGURE 3.2 – Comparaison entre reconnaissance et susceptibilité - *Source- MRN*

Les travaux réalisés par l'association Missions Risques Naturels (MRN) mettent en évidence les incohérences de la cartographie de la susceptibilité qui n'est pas représentative de la sinistralité observée puisque certaines régions en zone forte subissent des taux

de non reconnaissance élevés et inversement pour d'autres localisations moins susceptibles au RGA.

La carte d'exposition se distingue de la précédente car elle intègre la sinistralité observée au cours du temps d'où la notion d'exposition. Le degré de hiérarchisation reste le même et résulte du produit entre la susceptibilité et la sinistralité observée :

- Exposition nulle = susceptibilité nulle (rien ne change)
- Exposition faible = susceptibilité faible · sinistralité faible
- Exposition moyenne =
  - susceptibilité faible · sinistralité moyenne ou forte
  - susceptibilité moyenne · sinistralité faible ou moyenne
- Exposition forte =
  - susceptibilité moyenne · sinistralité forte
  - susceptibilité forte · sinistralité moyenne ou forte

La hiérarchisation des zones sinistrogènes repose sur la densité de sinistres observés par formation argileuse dans les zones urbanisées :

- Sinistralité faible = densité de sinistre au $km^2 < 2$
- Sinistralité moyenne = densité de sinistre au $km^2 \in [2 : 10]$
- Sinistralité forte = densité de sinistre au $km^2 > 10$

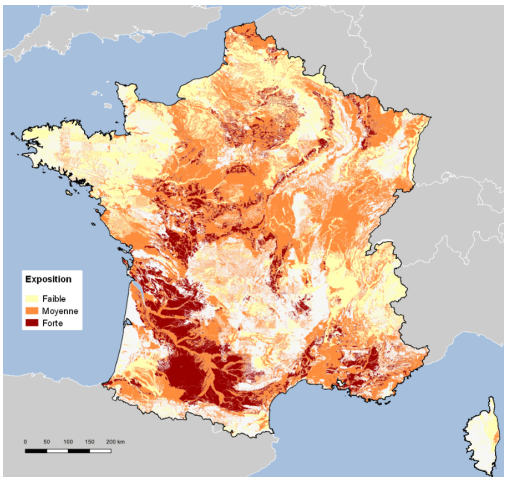
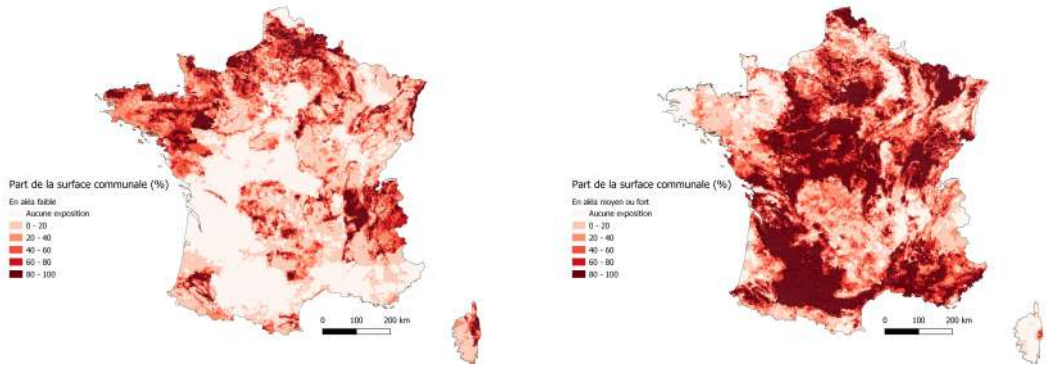


FIGURE 3.3 – Carte d'exposition au RGA - *Source : Géorisques*

La base des sinistres utilisée correspond à la base « Sinistres Indemnisés Liés aux Évènements Climatiques » (SILEC) et représente 70% du marché de l'assurance avec près de 180000 sinistres indemnisés pour le péril sécheresse entre 1989 et 2017. En plus d'avoir modifié la hiérarchisation de certaines zones, un *buffer*, c'est à dire une zone tampon de 100 mètres, a été appliquée autour des polygones spatiaux de sorte qu'une grande partie des 18000 sinistres localisés hors zone d'aléa (10% de la base) soit désormais rattachés à l'une d'entre elles. L'application de la zone tampon aurait permis de rattacher 5% des sinistres survenus hors zone d'aléa soit la moitié d'entre eux.

In fine, la zone d'exposition moyenne ou forte représente désormais 48.5% du territoire contre 23.6% pour la zone faible et 27.9% pour la zone non argileuse. En ce qui concerne le critère géotechnique, celui-ci n'est impacté que part la zone tampon qui peut augmenter légèrement la part de la surface communale en argile avéré.

Les cartes suivantes montrent l'exposition des communes en zone faible puis moyenne ou forte avec une coloration quantile.



(a) Aléa faible (b) Aléa moyen-fort

FIGURE 3.4 – Part de la surface communale selon le degré d'exposition

Nous avons extrait à partir de la carte d'exposition, la part de la surface communale au sein des 4 zones d'aléa ainsi que la superficie que cela représente.

3.1.3 La carte de l'ESDAC

L'*European Soil Data Centre* (ESDAC) est un centre de recherche européen fournissant des données sur la nature des sols pour le continent européen. Il fournit également une cartographie, cette fois-ci non hiérarchisé, de la présence d'argile. La figure 3.5 représente la concentration d'argile sur la couche superficielle du sol (0-20cm) avec une résolution de 500 mètres. Ces données sont issues d'un processus de modélisation de type MARS (*Multivariate Additive Regression Splines*) qui a permis d'extrapoler les 28000 points d'observations de la base LUCAS à l'ensemble du territoire.

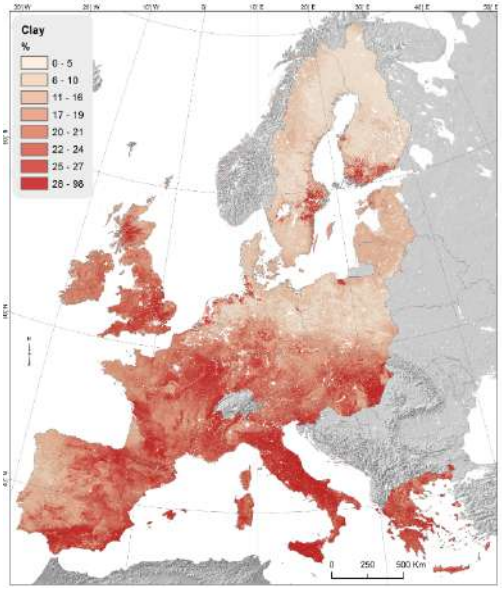


FIGURE 3.5 – Concentration des sols en argile *Source- ESDAC*

Nous avons extrait à partir de cette carte la concentration moyenne d'argile pour l'ensemble des communes du territoire en appliquant une moyenne pondé-

rée par la surface. D'autres choix sont également possibles comme par exemple retenir la concentration maximale par commune.

3.1.4 Les indicateurs d'exposition des maisons individuelles

Comme le modèle de détection ne repose sur aucunes données propre à l'assureur, nous devons ajouter aux communes des indicateurs d'exposition du bâti au phénomène.

En 2021, le « Service des Données et Études Statistiques » (SDES) a mis à jour les indicateurs d'exposition des maisons individuelles au retrait-gonflement des argiles par commune initiés en 2017 à la demande de la « Direction Générale de Prévention des risques » (DGPR). Les données résultent d'un croisement entre la cartographie d'exposition du BRGM, des fichiers cadastre/parcelle vectorisé ainsi que des fichiers démographiques d'origine fiscale sur les logements et les personnes (Fideli, INSEE). Les données parcellaires fournissent ainsi une information précise sur la localisation des biens tandis que la base Fideli permet d'ajouter les années de construction suivante : avant 1921, entre 1921 et 1945, entre 1946 et 1975 et après 1975. Ces périodes correspondent à des changements de modes constructifs délimités par les deux après-guerres et l'après-choc pétrolier.

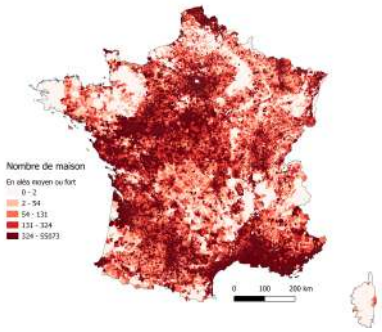


FIGURE 3.6 – Nombre de logements en zone moyenne ou forte

Le graphique 3.6 montre les communes avec le plus de maisons exposées à l'aléa moyen ou fort sans tenir compte de la période de construction. Ces indicateurs sont essentiels car pour que le risque se produise, il faut non seulement de l'aléa mais également de la vulnérabilité. Même si la part de la surface communale en aléa moyen ou fort est important, si aucun bâtiment n'est localisé sur ces couches alors il y a peu de chance pour qu'un sinistre lié au RGA survienne.

3.1.5 Conclusion

Pour la suite de l'étude, les données du portefeuille ont été croisées avec la carte d'exposition du BRGM en fonction de leur position XY ainsi qu'avec le zonier stochastique de Generali. De plus, nous avons souhaité ajouter aux communes à la fois la surface et la part de la surface communale pour chacune des couches de la cartographie d'exposition ainsi que la concentration d'argile moyenne (pondérée par la surface) issue de la carte de l'ESDAC.

Les indicateurs du SDES seront ajoutés au modèle de détection.

3.2 La base de données météorologiques

Outre les prédispositions géotechniques, il est nécessaire de quantifier l'intensité du phénomène afin de s'approcher à la fois des critères d'éligibilité à la reconnaissance mais également pour discriminer la fréquence.

Les sous-sections qui suivent introduisent la méthodologie de calcul des indices de sévérité. L'ensemble des données météorologiques nécessaires à la reproduction des indices sont issues de la base de données *ERA5-Land* produit par le centre de prévisions météorologiques européen (ECMWF pour *European Centre for Medium-Range Weather Forecasts*). Les données sont disponibles à l'open data depuis les années 1950 et jusqu'à 2 voir 3 mois avant la date actuelle. La résolution spatiale des données sont de $0,1° \times 0,1°$ soit environ 9 km^2.

3.2.1 Calcul des indices de précipitations standardisés

Standardized Precipitation Index (SPI)

Introduction [1]

Au cours du mois de décembre 2009, 54 experts représentant 22 pays autour du monde se sont réunis autour d'un atelier parrainé par des institutions météorologiques, gouvernementales et universitaires afin de débattre sur les différents indices de sécheresses météorologiques, agricoles et hydrologiques existant. La Déclaration de Lincoln a ainsi recommandé aux institutions météorologiques internationales d'utiliser l'indice de précipitation standardisé comme le critère de sécheresse à adopter en météorologie. Créé en 1993 par les scientifiques américains McKee, Doesken et Kleist, l'indice de précipitation normalisé est à la fois simple, interprétable, nécessitant peu de données, statistiquement robuste et possédant des propriétés de cohérence spatiale sur des régions aux climats différents.

Définitions

Le SPI, comme la plupart des autres indices de sécheresse, peut être calculé sur différents horizons temporels (1, 3, 6, 12 et 24 mois) afin de mesurer des sécheresses de plus ou moins long terme. Étant donnée que les sols superficiels réagissent plus vite à la sécheresse que les nappes phréatiques et cours d'eau, les indices de sécheresse seront calculés sur une période de trois mois afin de caractériser l'intensité sur une saison.

1. La référence suivante contient une description de l'indice de précipitations standardisée par l'Organisation Météorologique Mondiale : [OMM, 2012]

Le SPI calculé sur une période de 3 mois compare les précipitations moyennes sur les trois mois écoulés avec les précipitations moyennes pour cette même période de trois mois sur l'ensemble des données disponibles avec à minima 30 ans de données.

La valeur de cet indice peut être biaisé dans le cas ou le climat reste anormalement sec sur la période de 3 mois examinée. En effet si les précipitations s'écartent peu de leur moyenne long terme alors des valeurs extrêmes seront associées à un niveau de précipitation pourtant normal.

Les valeurs négatives du SPI définissent un déficit de précipitation, la période de retour de l'intensité du phénomène à un instant est donné par le tableau suivant.

Valeurs du SPI	Intensité	Période de retour
≥ 2	Extrêmement humide	1 fois tous les 50 ans
[1.5 ; 1,99]	Très humide	1 fois tous les 20 ans
[1 ; 1,49] et	Modérément humide	1 fois tous les 10 ans
[-0.99 ; 0.99]	Proche de la normale	1 fois tous les 3 ans
[-1 ; -1.49]	Modérément sec	1 fois tous les 10 ans
[-1.5 ; -1.99]	Très sec	1 fois tous les 20 ans
≤ -2	Extrêmement sec	1 fois tous les 50 ans

TABLE 3.1 – Valeurs du SPI

Pour une période de retour T, le quantile associé vaut $F_X^{-1}\left(\frac{1}{T}\right)$ avec :
- $X \sim \mathcal{N}(0,1)$
- $F_X^{-1}(\alpha) = inf\{y|F_X(y) \geq \alpha\}$ la fonction α-quantile.

Méthodologie

Notons $(X_{t,c})_{t\in\mathbf{N}}$ la série chronologique des précipitations mensuelles pour une cellule c quelconque. Pour calculer le SPI sur un horizon de trois mois, nous commençons par appliquer une moyenne glissante sur 3 mois à la série originale. Ceci consiste à remplacer $X_{t,c}$ par $\frac{X_{t,c}+X_{t-1,c}+X_{t-2,c}}{3}, t \in [1,\ldots,T]$ où T désigne la longueur de notre série temporelle. De fait les deux premières valeurs de la série ne sont pas renseignées. En introduisant l'opérateur de retard B tel que $(B^1X)_t = X_{t-1}$ et d'une manière plus générale $(B^hX)_t = X_{t-h}$ alors nous pouvons noter $(Y_{t,c})_t = \frac{(B^1+B^2+B^3)X_{t,c}}{3}$ la série de la moyenne glissante.

L'étape suivante consiste à séparer la série chronologique $(Y_{t,c})_t$ de longueur T en douze séries temporelles correspondant à chaque mois m de l'année.

$$Y_{t',m,c} = Y_{m+12*t',c},\ t' \in [0 : \lfloor T/12 \rfloor],\ m \in [1:12]$$

Chacune des 12 séries temporelles est ajustée à une loi Gamma dont la fonction de densité est donné par $f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\lambda x}, x \in \mathcal{R}^+$. Puis les séries mensuelles sont

standardisées en utilisant une projection quantile-quantile avec la loi normale centrée réduite.

$$Y^{std}_{t',m,c} = F^{-1}_{N_{0,1}}(\hat{F}_{Y\,t',m,c}(y))$$

Enfin, les douze séries mensuelles sont rassemblées afin d'obtenir la série standardisée pour la cellule c. Le processus est réitéré sur l'ensemble des cellules spatiales de sorte à obtenir une collection de cartes mensuelles d'indices standardisés depuis mars 1950.

Pour la construction de notre base trimestrielle, nous définirons les indices extrêmes sur une saison comme étant la valeur minimale de cet indice sur la saison concernée pour les indices de précipitations et la valeur maximale pour les températures. Par exemple, la valeur extrême des précipitations standardisées est définie comme :

$$ESPI_{a,s,c} = \underset{m \in s}{min}(SPI_{a,m,c})$$

où :

- $a \in \{1950, \ldots, 2021\}$
- $s \in \{$hiver, printemps, été, automne$\}$
- $m \in \{1, \ldots, 12\}$ le mois de l'année.
- c la cellule.

Les cartes de la figure 3.7 donne un aperçu de l'indice extrême $ESPI_3$ sur les 4 saisons de l'année 2018. Ces données peuvent être comparées avec la carte d'éligibilité au critère météorologique.

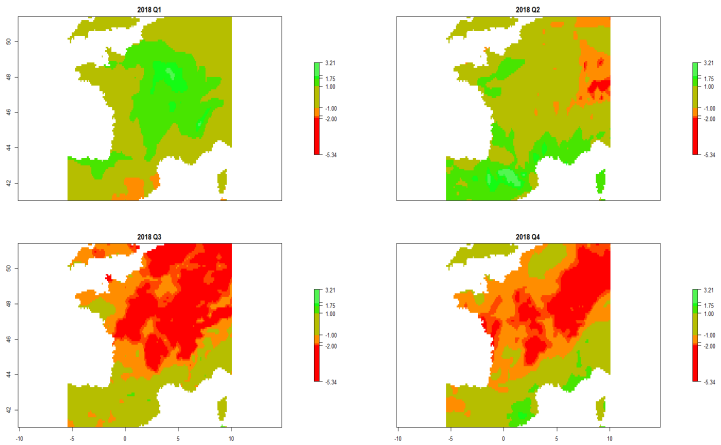


FIGURE 3.7 – Indice $ESPI_3$ sur l'année 2018

Le SPEI

Le SPEI a été introduit dans la continuité du calcul de l'indice SPI afin d'incorporer les données d'évapotranspiration potentielle. Désormais la série que l'on étudie correspond

aux précipitations nettes définies comme la différence entre les précipitations mensuelles totales (en mm) et l'évapotranspiration potentielle (en mm).

Les données d'évapotranspiration potentielle sont disponibles sur le *Climate Data Store* du centre ECMWF mais cette variable peut également être approchée par la formule de Thorntwaite (1948) à partir des températures mensuelles. D'autres méthodes plus complexes existent et nécessitent davantage de données comme celle de Penman-Monteith ou Priestley-Taylor.

$$PET_t = 16 * K * (10 * \frac{T_t}{I})^a$$

avec :

- $I = \sum_{i=1}^{12} \frac{T_t}{.} 5^{1.514}$, l'indice thermique mensuelle calculé à partir des températures moyennes mensuelles.
- T_t la température moyenne en degré celcius au temps t.
- K est un coefficient d'ajustement qui est fonction de la latitude et du mois donné
- $a = 6,75.10^{-7} * I^3 + 7,71.10^{-5} I^2 + 1,79.10^{-2} I + 0,49239 = \frac{1,6}{100} I + 0,5$

La méthodologie de calcul et l'interprétation reste la même que pour le SPI cependant la loi utilisée pour ajuster les précipitations nettes mensuelles est log-logistique et non plus gamma.

Les cartes de la figure 3.8 montre la reproduction de l'indice $ESPEI_3$ sur l'année 2018.

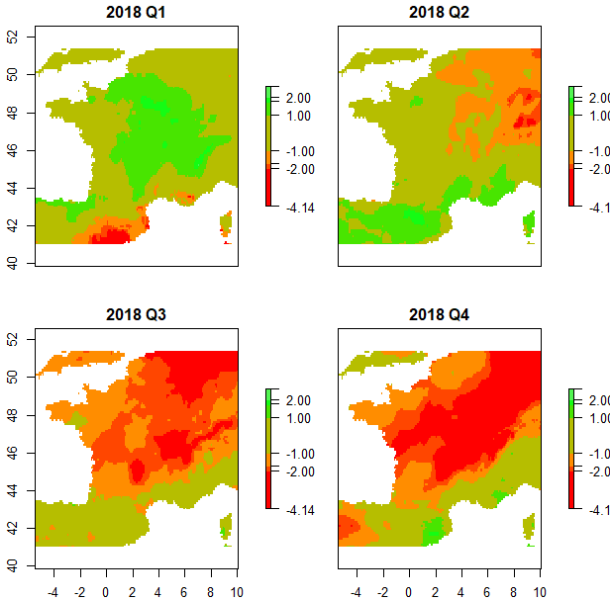


FIGURE 3.8 – Indice $ESPEI_3$ sur l'année 2018

3.2.2 L'indice d'humidité des sols superficiels

Définition du « Soil Wetness Index »

Le *Soil Wetness Index* (SWI) est un indice mesurant sur une profondeur d'environ 2 mètres l'état de la réserve en eau des sols par rapport à sa réserve utile. Il est défini comme :

$$SWI = \frac{w - w_{wilt}}{w_{wilt} - w_{field}} \tag{3.1}$$

où :

- w est le contenu intégré en eau du sol.
- w_{wilt} représente le point de flétrissement (*wilting point* en anglais), en deçà de cette quantité d'eau, les plantes ne peuvent survivre.
- w_{field} est la capacité de rétention maximale d'eau dans le sol, au delà de ce seuil le sol est saturé.

Le SWI est utilisé à la fois pour mesurer la sécheresse agricole en France au travers du SSWI (*Standardized Soil Wetness Index*) mais il joue également le rôle de critère météorologique dans une version uniforme, spécialement conçu pour le régime Cat Nat.

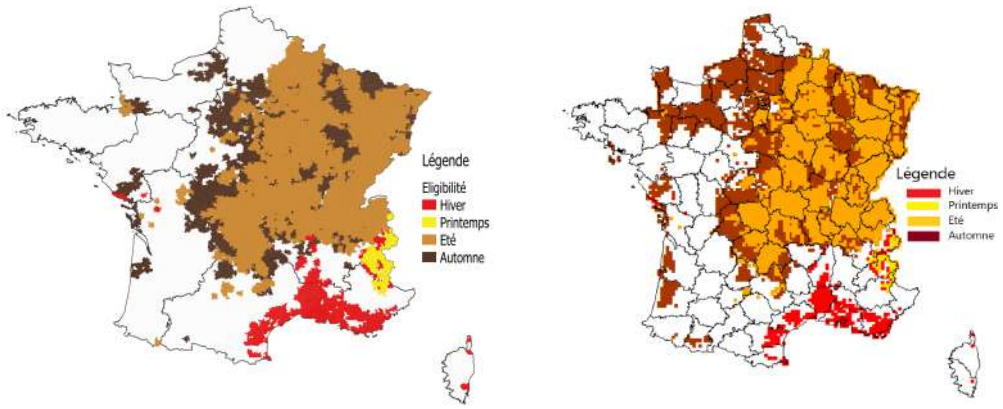
Le SWI dans sa version uniforme

Le SWI uniforme utilisé par la commission interministérielle est un output du modèle hydrométéorologique SIM pour Safran-Isba-Modcou construit par le Centre National de Recherche Météorologiques (CNRM) et utilisé par Météo France. Il s'agit d'une configuration du modèle spécifique au dispositif Cat Nat reposant sur une hypothèse de texture et de végétation uniforme sur l'ensemble du territoire :

- Choix d'une végétation de type gazon.
- Choix d'une texture très argileuse (58% d'argile, 12% de sable).

Depuis l'année 2021, Météo France met à disposition les données uniformes moyennées sur une période de trois mois de 1969 à 2020. Les valeurs moyennées sont réparties sur le maillage fixe SAFRAN avec une résolution spatiale de 64 km^2. Elles sont définies sur le système de coordonnées géographiques Lambert 2 étendu (EPSG :27572) et nous permettent de reproduire les critères de reconnaissance météorologique pour les années 2018, 2019 et 2020.

La figure 3.9 compare la carte d'éligibilité publiée par Météo France au cours de l'année 2018 avec une reproduction des critères à partir des données uniformes communiquées par Météo France. Nous pouvons apercevoir des disparités entre les cartes qui ne devraient pas exister.



(a) Reproduction des critères (b) Éligibilité Météo France

FIGURE 3.9 – Comparaison des cartes d'éligibilité

De leur propre côté, la Mission Risques Naturels (MRN) a reproduit notre travail et observe les mêmes conclusions. Certaines communes ont été déclarées en l'état de catastrophe naturelle sécheresse alors qu'elles n'étaient pas éligibles.

Le tableau 3.2 donne un aperçu du niveau de rappel entre l'éligibilité et la variable binaire « Arrêté ». En théorie, nous devrions avoir un niveau de rappel égal à 100%, tous les arrêtés Cat Nat doivent respecter les critères météorologiques. Cependant, le taux de rappel observé est en-dessous de 100%, il y donc des erreurs contenues dans les données transmises par Météo-France.

Rappel	2018	2019	2020
MRN	95%	98%	99%
Reproduction	95.41%	98.72%	99.07%

TABLE 3.2 – Rappel entre éligibilité et arrêté

Malheureusement, la mise à jour du SWI uniforme n'est pas régulière ce qui nous empêche de l'utiliser dans nos modèles. Toutefois, les données présentes dans la base ERA5-Land nous permettent de calculer le niveau d'un SWI non uniforme sur plusieurs couches du sol. La section qui suit compare les deux versions de l'indice d'humidité des sols et l'application des critères en vigueur.

Comparaison du SWI uniforme et non uniforme

Standardisation du SWI non uniforme

Les données ERA5-Land fournissent la contenance volumétrique des sols en eau sur quatres couches du sol (0-7cm, 7-28cm, 28cm-1m et 1m-2,89m) ce qui laisse la liberté de construire le SWI non uniforme sur la ou les couches souhaitées. Dans cette étude, nous

avons conservé les couches intermédiaires et trois autres profondeurs (0-28cm, 0-1m et 0-2m 89) tel que :

$$SWI_{1m} = \frac{7 \cdot SWI_{0-7cm} + 21 \cdot SWI_{7cm-28cm} + 72 \cdot SWI_{28cm-1m}}{7 + 21 + 72}$$

Comme les cartes des points de flétrissement w_{wilt} et de capacité maximale de rétention des sols w_{field} sont invariants au cours du temps, les variations du SWI ne dépendent que des variations de la contenance volumétriques des sols en eau. Ainsi, la standardisation du SWI revient à la standardisation de l'eau intégrée dans le sol w.

La technique de standardisation du SWI repose sur la même méthodologie que le SPI. Toutefois en l'absence de documentation, nous avons choisi de simplement centrer-réduire les observations mensuelles plutôt que de calibrer une loi et effectuer la projection quantile-quantile.

$$SSWI_{1m} = \frac{SWI_{1m} - \mu}{\sigma}$$

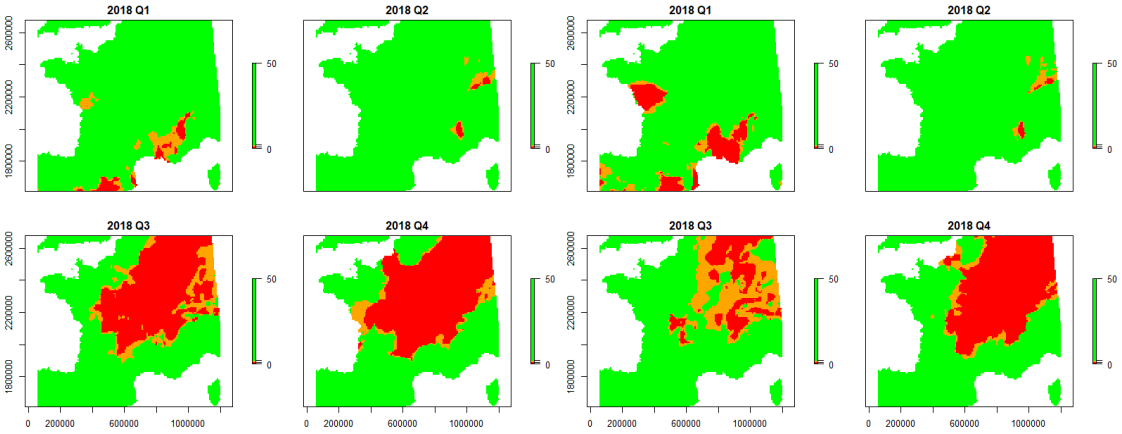
avec :

- μ la moyenne de la série mensuelle (moyennée sur une période de 3 mois)
- σ l'écart-type la série mensuelle (moyennée sur une période de 3 mois)

Reproduction des critères de reconnaissance

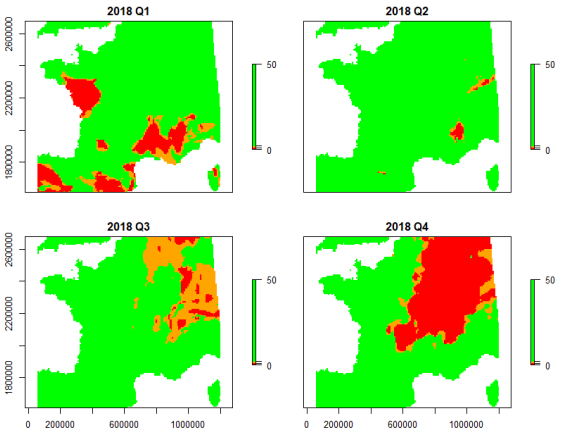
Pour construire de manière homogène les critères de reconnaissance sur les données non uniformes, nous avons commencé par projeter les données d'ERA5-Land initialement géoréférencées sur système géodésique WGS 84 vers le Lambert II étendu. Afin de se reposer sur exactement le même maillage utilisé par la commission interministérielle (maillage SAFRAN), le SWI non uniforme a été distribué sur ce dernier par interpolation bilinéaire.

Les cartes de la figure 3.10 comparent l'éligibilité reproduite avec les données uniformes et les données non uniformes pour l'année 2018. Nous pouvons observer des similarités géographiques comme des disparités entre les données uniforme et non uniforme. Même si le SWI non uniforme est plus homogène que son homologue, nous observons une forte corrélation entre les deux.

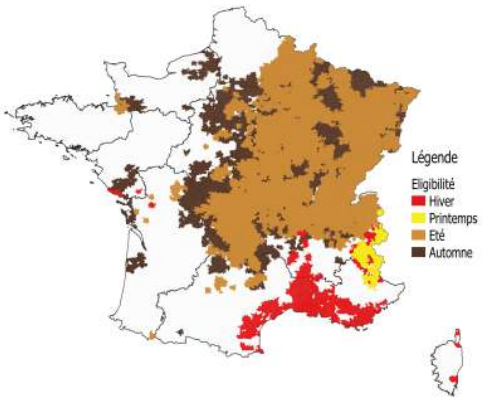


(a) Critère météorologique (0-28cm)

(b) Critère météorologique (0-1m)



(c) Critère météorologique (0-2m89)



(d) Critère météorologique uniforme

FIGURE 3.10 – Comparaison des critères météorologiques en 2018

Pour la suite, nous garderons les rangs du SWI non uniforme comme variable explicative plutôt que la variable binaire représentant l'éligibilité.

3.2.3 Conclusion

In fine, nous avons récolté une collection d'indices météorologiques extrêmes sur les saisons. La base de données météorologiques contient ainsi les variables suivantes :

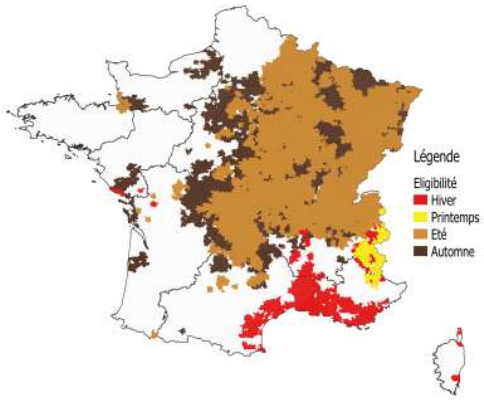
- Les valeurs $ESSWI$ sur plusieurs couches du sol.
- La reproduction des critères d'éligibilité sur plusieurs couches du sol.
- Les valeurs des indices $ESPI_3$ et $ESPEI_3$.
- Les valeurs des indices de températures standardisés $ESTI_3$.

Cette collection d'indices servira à la fois à discriminer la fréquence ainsi qu'à classifier les communes pour lesquelles nous estimons qu'un arrêté va être publié.

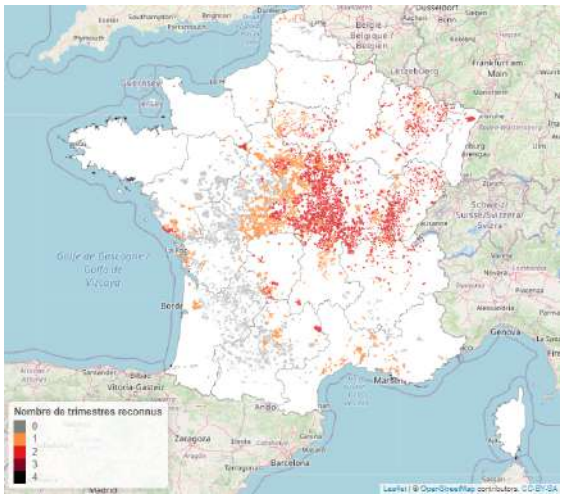
3.3 L'historique des demandes de reconnaissance

Comme nous le montre le graphique 3.11ci-dessous, nous constatons que toutes les communes éligibles au critère météorologique n'effectuent pas forcément la démarche de demande de reconnaissance Cat Nat sécheresse. Plusieurs raisons plausibles peuvent expliquer ceci :

- Aucun habitant n'a été sinistré et cela malgré l'intensité anormale de la sécheresse.
- Il n'y a pas de bâtiment exposé.
- La commune méconnaît le risque auquel elle est exposée.



(a) Éligibilité



(b) Arrêté observé

FIGURE 3.11 – Comparaison entre les arrêtés observés et l'éligibilité

Afin de capter l'habitude des communes dans leurs démarches administratives, nous avons ajouté au modèle de détection une variable recensant, au cours du temps et pour chaque commune, le nombre de demandes de reconnaissances passées. De manière à ne pas anticiper sur le futur, le nombre de demandes passées pour la commune α au temps t doit être connu à cette date. Autrement dit, les demandes comptabilisées doivent être publiées au journal officiel avant la date t.

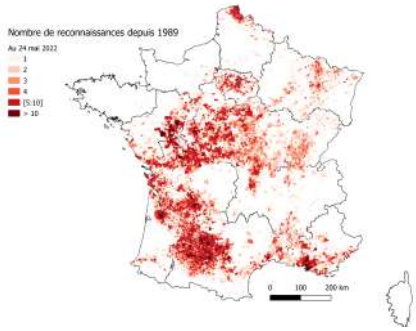


FIGURE 3.12 – Nombre de reconnaissances antérieures

Cette variable est construite à partir de la base nationale de Gestion ASsistée des

Procédures Administratives relatives aux Risques (GASPAR) contenant la liste des arrêtés Cat Nat par typologie de péril depuis les années 1989. La carte 3.12 hiérarchise les communes en fonction du nombre de reconnaissances antérieures depuis l'intégration du RGA au régime Cat Nat.

3.4 Les données risques

Afin de construire nos modèles de fréquence et de sévérité, nous avons besoin des données risques du portefeuille MRH de Generali. Cette base contient toutes les informations relatives aux biens assurés par Generali.

Afin d'adapter nos données à notre problématique, nous avons appliqué certaines transformations et hypothèses. Dans les sous-sections qui suivent, nous détaillerons les filtres et transformations appliquées à la base contrats ainsi qu'à la base sinistre.

Nous commencerons par analyser les données relatives aux sinistres car ce sont ces informations qui ont motivés nos choix. Enfin, nous présenterons les transformations appliquées à la base contrat jointe avec les sinistres.

3.4.1 Les données sinistres

Les données sinistres contiennent les informations de 2605 sinistres sécheresse ouvert ou clos sur la période 2010-2021. Les variables qui nous intéressent sont les suivantes :

- Le numéro de police
- La date de survenance
- La date de déclaration
- L'état du dossier
- Le montant du sinistre
- Les données risques associées

Choix des filtres utilisés

Nous pouvons remarquer sur la figure 3.13 qu'il y a trop peu de sinistres clos dans notre base de données et il sera très difficile de réaliser un modèle dans ces conditions, c'est pourquoi nous aimerions inclure les sinistres ouverts dans la variable cible des modèles fréquences.

De plus, nous remarquons que le taux de clotûre des sinistres est relativement long. Il y a donc beaucoup de sinistres ouverts pour les années récentes. La plupart des sinistres ouverts et survenus avant 2019 seront clôturés par la suite car les arrêtés ont été publiés. Seule une faible proportion sera classée sans-suite après décision des experts. En revanche, les sinistres ouverts et survenus après 2019 ne peuvent être inclus à la modélisation car tout les arrêtés ne sont pas publiés et une partie de ces sinistres seront classés sans-suite en l'absence d'arrêtés.

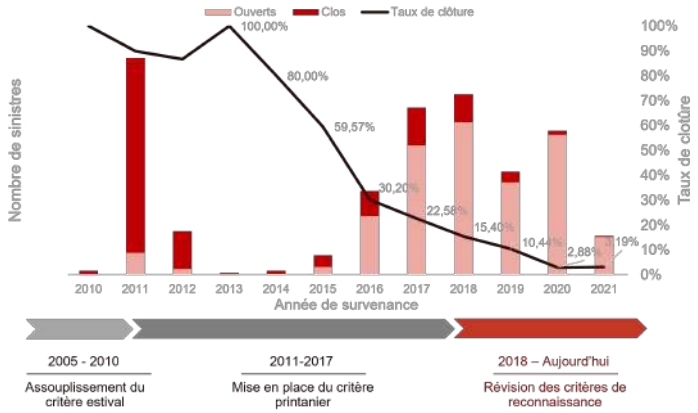


FIGURE 3.13 – Nombre de sinistres par année de survenance et état de dossier

Parmi les **2605 sinistres** entre 2010 et 2021, 2501 sinistres concernent les propriétaires de maison individuelle (occupant et non occupant) soit 96% du total. Ce second constat nous pousse à écarter les autres modalités d'habitation et d'habitant de notre base contrat étant donnée la faible volumétrie de sinistres sur ces profils.

Par la suite, la base contrat sera filtrée sur les propriétaires de maisons et les sinistres ouverts seront inclus dans la modélisation pour les années avant 2019.

Raccrochage des sinistres aux arrêtés Cat Nat

Afin d'expliquer au mieux la fréquence d'apparition d'un sinistre sécheresse en fonction de l'intensité du phénomène et sachant que la commune a été reconnue, l'ensemble des sinistres déclarés doivent être survenus au cours d'une saison reconnue en l'état de catastrophe naturelle. Malheureusement, une partie de nos sinistres déclarés semble être survenus au cours d'une saison ou d'une année non reconnue en l'état de catastrophe naturelle.

Laisser ces sinistres en-dehors d'une reconnaissance aura plusieurs conséquences :
— Nous risquons de manquer une proportion non négligeable de nos sinistres par la suite en filtrant nos données sur les arrêtés observés.
— La date de survenance des sinistres mal raccrochés pourrait avoir lieu au cours d'une période humide et induirait un biais lors de la modélisation de la fréquence.

Pour les années antérieures, certains sinistres ont été classés clos à tord par les gestionnaires et correspondent en réalité à des sinistres sans-suite. Nous pouvons les identifier en écartant les sinistres clos pour lesquels le montant du sinistre correspond au montant des honoraires de l'expert mandaté. Les autres sinistres clos mal rattachés correspondent à des erreurs ou des oublis de la part des gestionnaires de sinistres. Il est indispensable

de les raccrocher à leur véritable date de survenance correspondant à une saison reconnue.

Pour les sinistres ouverts le cas est un peu plus complexe, les sinistres en-cours les plus anciens correspondent pour la majorité d'entre eux à des litiges et le délai de publication est largement dépassé. Par conséquent, nous devons les rattacher, s'il le faut, à l'arrêté le plus proche.

Même si le taux de clôture reste faible pour les années 2018 et 2019, à cause du processus d'indemnisation qui reste long, le délai de 2 ans pour effectuer une demande de reconnaissance est dépassé et normalement aucun arrêté ne peut être publié sur ces périodes. Nous avons donc fait le choix de raccrocher les sinistres clos et ouverts survenus avant l'année 2019 (incluse) à la date d'arrêté antérieur la plus proche et cela dans la limite de deux ans. Ce délai correspond au délai maximal d'instruction pour la procédure ordinaire.

Les sinistres déjà raccrochés à une saison reconnue seront joints à la base contrat avec le numéro de police et en comparant les dates de survenance des sinistres avec les dates de début et de fin de mensuelle de nos contrats.

En revanche pour les sinistres que nous rattachons, il est impossible de déterminer le véritable profil de risque s'il en existe plusieurs pour le même trimestre. Ce cas de figure peut survenir en cas de modification du profil de risque de l'assuré sur le trimestre à raccrocher. Dans ce cas particulier, il y a autant de lignes que de profils de risques associés à la police. Nous devons effectuer un choix arbitraire pour ne pas comptabilisé en double le sinistre lors de la jointure. Le choix appliqué a été de raccrocher le sinistre au profil de risque le plus exposé sur le trimestre de raccrochage.

Le graphique 3.14 résume l'impact de l'ensemble des filtres sur le nombre de sinistres retenus pour l'apprentissage.

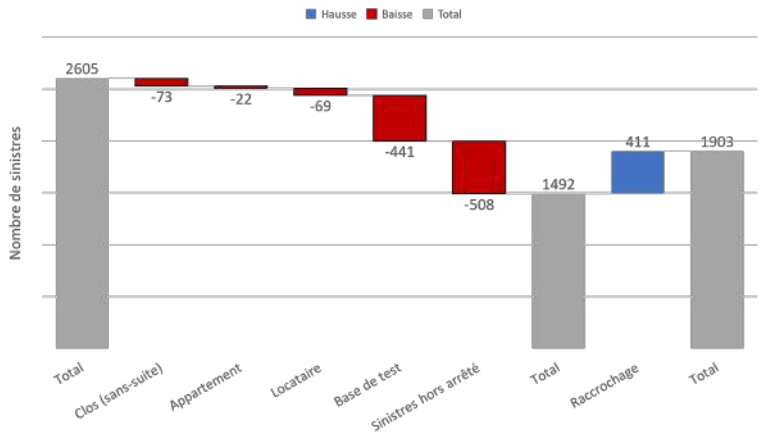


FIGURE 3.14 – Impact des filtres sur le nombre de sinistres retenus

In fine, nous ajouterons les informations suivantes à nos contrats :
- nbsin_Sécheresse : le nombre de sinistres clos ou ouvert sur le péril sécheresse (variable cible pour le module fréquence)
- nbsin_clos_Sécheresse : le nombre de sinistres clos sur le péril sécheresse (pour calculer le coût moyen sur le module sévérité)
- mtsin_Sécheresse : la charge brut des sinistre clos ou ouvert sur le péril sécheresse (module fréquence)
- mtsin_clos_Sécheresse : la charge brut des sinistres clos sur le péril sécheresse (module sévérité)

3.4.2 La base des contrats

Présentation de la base

La base de données contrats MRH contient l'ensemble des données de risques de notre portefeuille par mensuelle pour les années 2010 à 2021 inclus. Il s'agit en quelque sorte d'une photographie de notre portefeuille à un instant t. Chaque ligne de la base correspond à une police et ses données de risques au cours de la mensuelle et de l'année donnée.

En cas de modification des données de risques comme le niveau des capitaux mobiliers ou encore un déménagement qui change complètement la vision du risque alors deux lignes avec le même numéro de police sont crées sur une même mensuelle :
- L'une correspond à l'ancienne vision du risque, la date de fin de mensuelle correspond à la date de modification des risques. Cette ancienne version du contrat est enregistrée jusqu'à la fin de l'année mais avec une exposition nulle.
- L'autre ligne correspond à la vision du risque la plus récente, elle débute à compter de la date de modification des risques soit la fin de l'ancienne version et perdure jusqu'à la fin du contrat (résiliation, annulation ou modification du profil de risques.

Pour notre étude, nous nous intéresserons uniquement à une partie des variables contenues dans la base contrat. Par mesure de confidentialité, nous ne présenterons qu'une partie d'entre elles :
- k_hinc : la tranche de capitaux mobiliers hors incendie
- annee_construction : l'année de construction
- zonier exposition du BRGM
- nbpieces : le nombre de pièces

Comme notre méthode repose sur l'estimation d'un nombre de sinistres à la maille Insee, année, trimestre, nous avons dû calculer l'exposition trimestrielle pour chaque individu de notre base. Après avoir joint nos sinistres aux bons profils de risque, nous avons agrégé l'exposition, le nombre ainsi que le montant des sinistres par code Insee, année, trimestre et profils de risques.

Chapitre 4

Présentation des résultats

Les sections qui suivent détaillent les résultats de nos différentes approches pour l'estimation de la charge ultime. Pour rappel, nous avons souhaité comparer les prédictions de deux méthodes :

- Une approche fréquence/sévérité qui, à partir des données risques, des informations météorologiques, géotechniques et des habitudes communales, prédit respectivement la fréquence et le coût des sinistres liés au RGA.
- La combinaison d'un modèle de classification, détectant les communes reconnues en l'état de catastrophe naturelle sécheresse, avec une approche fréquence/sévérité qui prédit la fréquence et le coût des sinistres sachant que la commune a été reconnue.

Chacune des étapes nécessaires à la modélisation dans les différentes méthodes seront détaillées. Par mesure de confidentialité, l'ensemble des informations relatives aux données de risques intervenant dans le tarif MRH seront omises ou masquées.

Nous commencerons par analyser la première méthodologie avant de parcourir le modèle de détection puis le modèle de fréquence post-détection. Enfin, après avoir justifié l'utilisation d'un coût moyen pour la partie sévérité, nous comparerons les deux méthodes précédentes afin de sélectionner la plus pertinente d'entre elles.

4.1 1re approche : Modèle fréquence

4.1.1 Descriptif de la base de données

Ce premier modèle a pour objectif d'estimer la fréquence de sinistres à partir d'un modèle linéaire généralisé de type Poisson. La base de données utilisée résulte du croisement entre :

- La base contrats MRH du portefeuille de Generali, agrégée à la maille Insee, année et trimestre. Cette base contient les données risques de nos contrats sur la période 2010-2021 inclus.

- Les sinistres sécheresse ouverts et clos dont certaines dates de survenance ont été modifiées afin de raccrocher les sinistres à une période reconnue Cat Nat.
- Les données météorologiques contenant plusieurs indices extrêmes standardisés sur la saison. Cette base contient également les rangs, sur une période glissante de 50 ans, du niveau d'humidité des sols sur plusieurs profondeurs.
- Les données géologiques contenant la part de la surface communale sur chacune des couches de la carte d'exposition ainsi que la concentration moyenne d'argile de la commune.
- Le nombre de demandes de reconnaissance passées (favorables ou défavorables).

Le jeu de données ainsi créé contient 18 283 177 individus que nous avons séparé par la suite en trois bases :

- Une base d'apprentissage correspondant à 80% des données tirées aléatoirement et sans remises entre 2010 et 2019 inclus.
- Une base de validation dite « holdout » correspondant aux 20% des données restantes entre 2010 et 2019.
- Une base de test contenant les années 2020 et 2021.

Le tirage aléatoire des individus a été réalisé en amont sur la base complète de sorte à ce que le même échantillonage soit utilisé pour la construction des modèles de fréquence au sein des deux approches. De la même manière, nous avons défini l'année 2019 comme année de référence au cours des deux approches. Le choix de l'année de référence impacte le coefficient $\hat{\beta}_0$ et donc l'interprétation des coefficients $\hat{\beta}_j$ de nos modèles.

4.1.2 Sélection de variables

Pour sélectionner les variables optimales, nous avons procédé de la manière suivante :

- Itérations successives de la méthode *forward* pas à pas jusqu'à saturation du modèle. A chaque étape *forward*, les variables trop corrélées sont exclues du modèle. Si nécessaire, nous regrouperons les modalités (simplifications) de sorte à ce que les coefficients soient significativement non nuls.
- Lorsque le modèle est à priori stable, nous essayerons d'ajouter les variables restantes en simplifiant les modalités.
- Enfin, nous vérifierons la stabilité du modèle avec une itération *backward*.

Sélection *forward* et simplification

Après avoir ajouté et simplifié les variables faisant diminuer l'AIC avec la méthode *forward*, nous avons parcouru et retravaillé les variables qui n'ont pas été sélectionnées. Ainsi les capitaux mobiliers qui possèdent de nombreuses modalités ont pu être ajoutés au modèle ainsi que la variable « Risque 2 ». Même si l'impact de cette variable est quasiment nul sur le BIC, elle améliore un peu l'AIC et la déviance (-15 points sur l'AIC). Le tableau 4.8 en annexe résume l'impact de chaque variable dans le modèle avant et

après regroupement de modalités.

Le graphique 4.1 reprend ces impacts post-simplification sur le BIC :

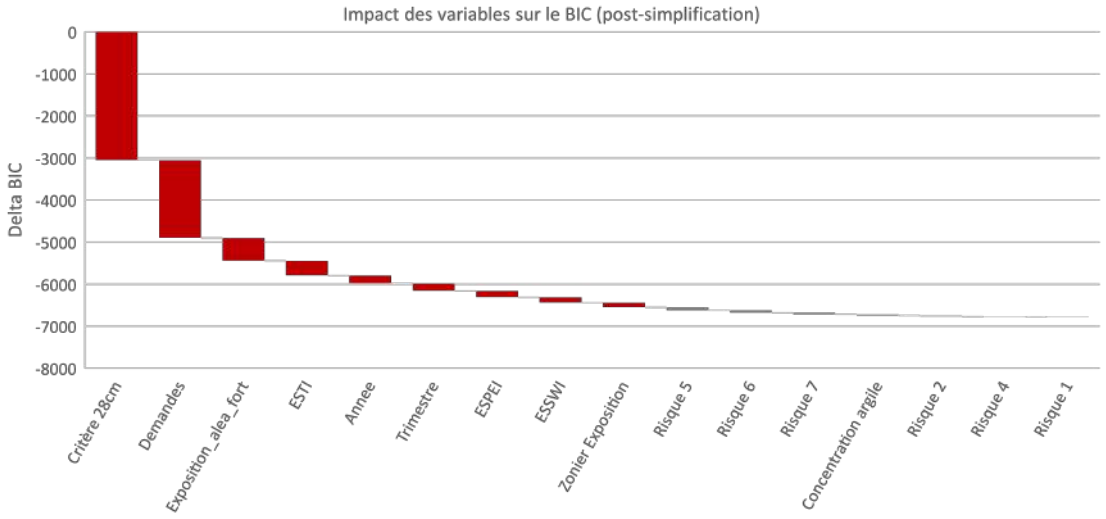
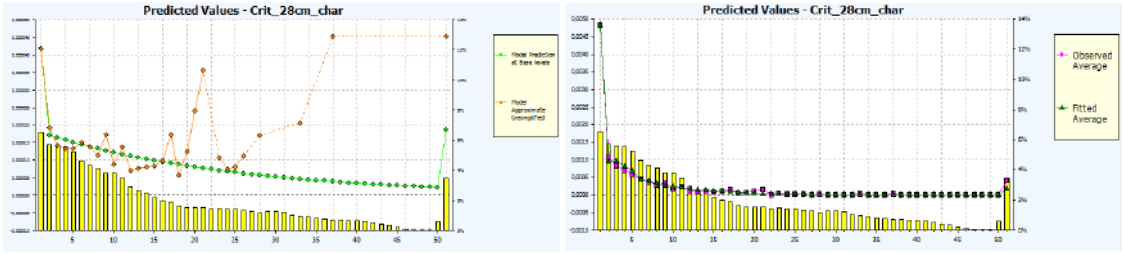


FIGURE 4.1 – Impacts des variables sur le BIC

Nous pouvons remarquer que les variables les plus influentes de notre modèle concernent des données météorologiques et géologiques. En particulier, la reproduction des critères météorologiques sur la couche superficielle du sol (0-28cm) à partir du SWI non uniforme ainsi que le nombre de demandes de reconnaissance antérieures concentrent 72% de la perte totale de BIC avec 4901 points.

Reproduction des critères sur la couche superficielle du sol

Les graphiques de la figure 4.2 montrent les simplifications utilisées ainsi que la fréquence observée et prédite sur la variable la plus importante.

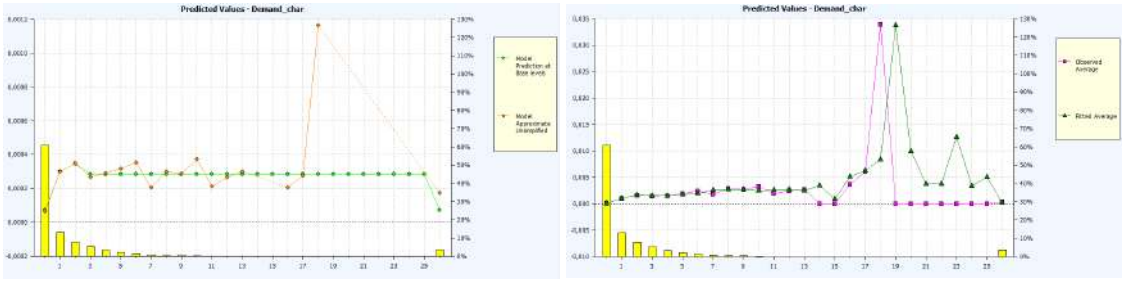


(a) Simplification (b) Comparaison fréquence observée/prédite

FIGURE 4.2 – Critère$_{28cm}$: simplification et comparaison de la fréquence prédite et observée

Nous pouvons remarquer que la fréquence observée augmente fortement lorsque le rang de la variable $\text{Critre}_{28\text{cm}}$ vaut 1, cette valeur indique une période de retour excédant 50 ans. En plus de quantifier l'intensité du phénomène, cette variable qualifie l'éligibilité de la commune vis-à-vis des critères de reconnaissance. Nous avons pu le constater précédemment avec la figure 1.3 à la page 23.

Nombre de demandes antérieures



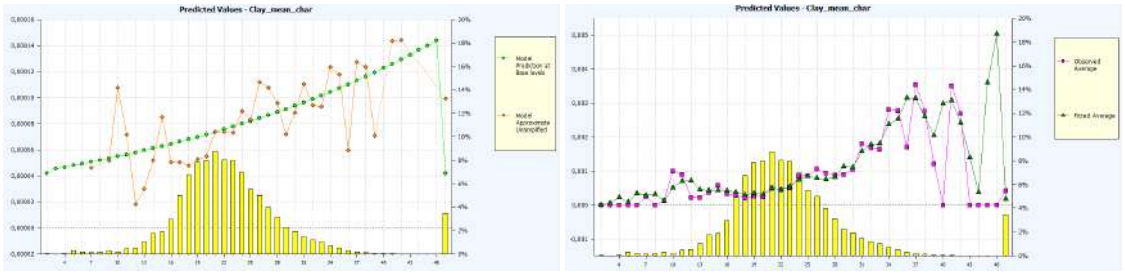
(a) Simplification (b) Comparaison fréquence observée/prédite

FIGURE 4.3 – Nombre de demandes de reconnaissance : simplification et comparaison de la fréquence prédite et observée

Afin de respecter la significativité des coefficients, nous avons dû regrouper la variable désignant le nombre de demandes de reconnaissance antérieures en seulement deux modalités. Ainsi, les communes ayant réalisées au moins une demande par le passé ont une fréquence qui est plus élevée que les autres. Elles sont donc plus susceptibles d'être sinistrées à l'avenir. Ces constats sont résumés par les graphiques 4.3a et 4.3b.

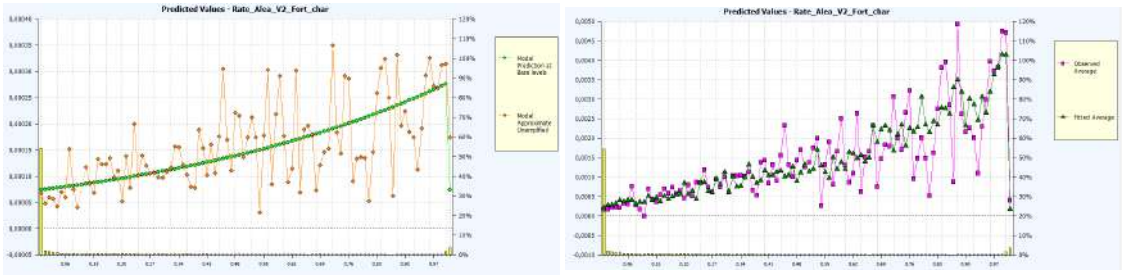
Données géotechniques

En plus des différents zoniers présents dans la base de tarification, deux variables référant à la présence d'argile ont été sélectionnées. Il se trouve que tous nos contrats ne sont pas géocodés à la maille XY. Ainsi, l'ajout de ces variables vient compléter une partie de l'information manquante sur l'exposition à l'aléa avec un indicateur communal.



(a) Simplification (b) Comparaison fréquence observée/prédite

FIGURE 4.4 – Concentration moyenne d'argile : simplification et comparaison de la fréquence prédite et observée



(a) Simplification (b) Comparaison fréquence observée/prédite

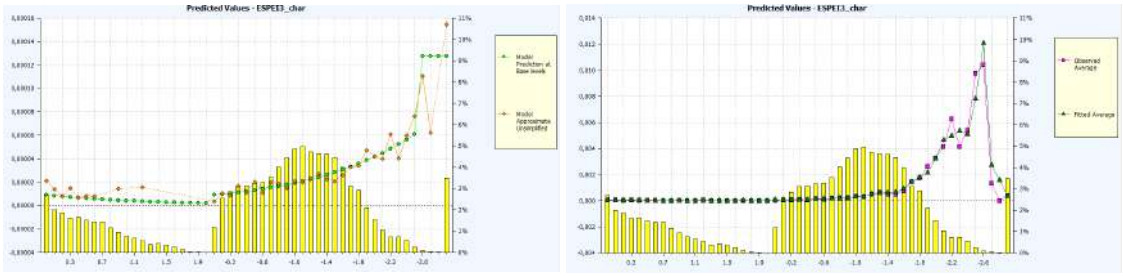
FIGURE 4.5 – Part de la surface communale en aléa fort : simplification et comparaison de la fréquence prédite et observée

Nous pouvons constater une augmentation de la fréquence observée avec la concentration d'argile dans le sol ainsi qu'avec la part de la surface communale en aléa fort.

Indices de sécheresse

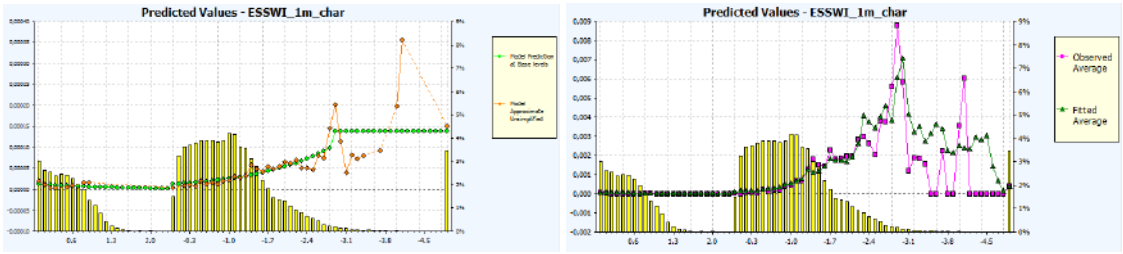
Trois indices de sécheresse ont également été retenus, il s'agit des indices relatifs à la température (ESTI), aux précipitations nettes (ESPEI) ainsi qu'au bilan hydrique des sols sur une profondeur d'un mètre (ESSWI_$1m$).

Pour ces trois indices, nous pouvons apercevoir une augmentation de la fréquence sur la queue de distribution. Autrement dit, le risque qu'un sinistre survienne augmente avec des conditions météorologiques extrêmement défavorables. Ces constats sont résumés au travers des graphiques suivants.



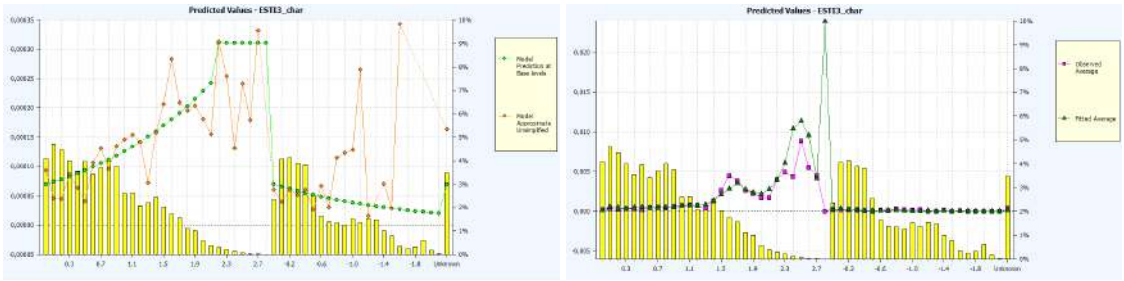
(a) Simplification- ESPEI (b) Fréquence observée/prédite - ESPEI

FIGURE 4.7 – Précipitations nettes standardisées : simplification et comparaison de la fréquence prédite et observée



(a) Simplification- ESSWI (b) Fréquence observée/prédite - ESSWI

FIGURE 4.8 – Humidité des sols standardisées : simplification et comparaison de la fréquence prédite et observée

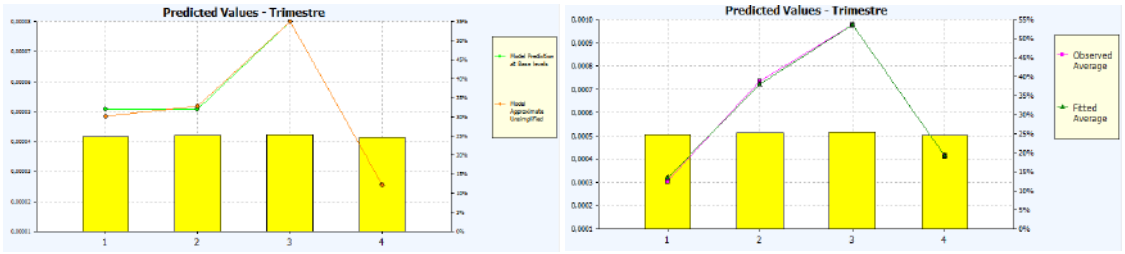


(a) Simplification - ESTI (b) Fréquence observée - ESTI

FIGURE 4.6 – Températures standardisées : simplification et comparaison de la fréquence prédite et observée

Impact de la saison

Enfin, la saison figure également dans notre modèle. Au cours de l'été, les épisodes de fortes températures accélèrent le processus d'évapotranspiration et favorisent le retrait des sols argileux d'où la fréquence plus élevée sur cette période. Les épisodes de sécheresse automnales sont plus rares dans nos données d'apprentissage d'où la fréquence plus faible.



(a) Simplification - Trimestre (b) Fréquence observée - Trimestre

Sélection backward

Lors de la sélection *backward*, la variable « Risque 5 » a été écartée du modèle car son retrait implique une amélioration du BIC de 11 points. En revanche, nous avons choisi de garder les capitaux mobiliers dans le modèle car la perte de BIC est très faible (-1 point) et son retrait engendre une augmentation d'environ 13 points sur l'AIC et la déviance.

Le résultat de la sélection *backward* est résumé avec le tableau 4.1 qui suit :

Variables	ΔAIC	ΔBIC	Δdéviance
Risque 5	17,68	-11,00	17,36
Risque 1	13,34	-1,00	13,27
Risque 2	16,07	1,73	16,03
Risque 4	17,49	3,15	17,30
ESPEI3	36,91	22,57	36,50
Concentration d'argile	37,30	22,96	36,56
Risque 6	72,97	58,63	72,73
Risque 7	92,36	63,68	91,95
Critère 0-28cm	94,87	66,19	94,39
Zonier exposition	130,48	101,80	130,05
Exposition aléa fort	163,38	149,04	163,12
ESTI3	205,07	190,73	205,44
Année calendaire	252,31	194,95	252,20
Trimestre	234,59	205,91	235,05
ESSWI_1m	251,62	237,28	252,89
Demandes	293,99	279,65	293,32

TABLE 4.1 – Sélection backward

4.1.3 Validation des hypothèses et du modèle

Comme les GLM sont des modèles paramétriques, certaines hypothèses doivent être vérifiées afin de s'assurer que les coefficients soient bel et bien interprétables. Pour cela,

nous vérifierons :

- les corrélations entre les variables.
- la stabilité temporelle des coefficients.
- la significativité des coefficients.

Corrélation des variables

Compte-tenu de la volumétrie de notre base, nous présenterons uniquement les corrélations entre les variables sélectionnées. Le contrôle des corrélations se fait en réalité en amont lors de la sélection de variable.

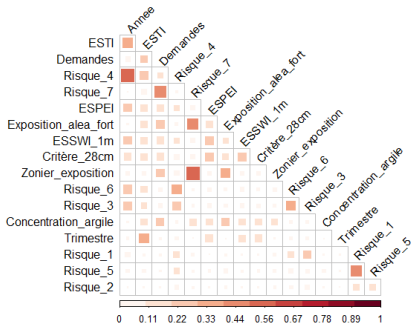


FIGURE 4.10 – V de Cramer sur les variables sélectionnées

Pour l'étude des corrélations, nous nous sommes fixés un seuil de 70% à ne pas dépasser. Ce seuil correspond au standard proposé par le logiciel de tarification Emblem.

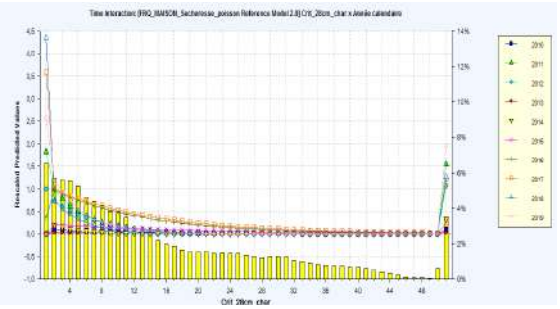
Comme nous pouvons le voir sur la figure 4.10, les variables qualitatives ou discrétisées ne sont pas significativement corrélées les unes aux autres donc les coefficients du modèle sont bien interprétables. Les variables « année » et « Risque 4 » ont une valeur de 0,575 sur le V de Cramer contre 0,58 entre le zonier sécheresse et le zonier stochastique.

Stabilité temporelle

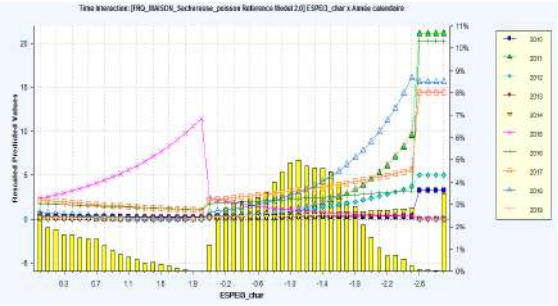
La stabilité temporelle est une étape très importante car elle permet de s'assurer que les coefficients et les simplifications réalisées restent stables au cours du temps. Elle consiste à ajuster un modèle pour chacune des années présentes dans notre base d'apprentissage et de vérifier que nos simplifications ont les mêmes tendances.

Si l'une de nos variables présente des tendances différentes selon les années alors elle est retirée du modèle. Compte-tenu de la faible volumétrie de sinistres pour les années 2010, 2013, 2014 et 2015, ces dernières peuvent montrer des incohérences liées à un effet volume. Nous ne tiendrons pas compte de ces années pour nos décisions sur la stabilité.

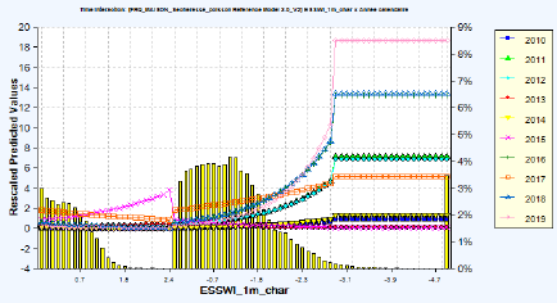
Les graphiques suivants montrent les résultats de la stabilité temporelle sur les variables présentées. Les variables de risques ne sont pas présentées mais sont stables au cours du temps.



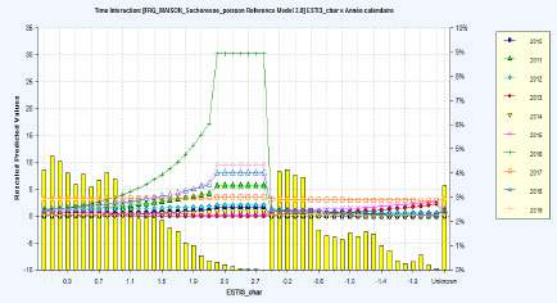
(a) Critère météorologique (0-28cm)



(b) ESPEI

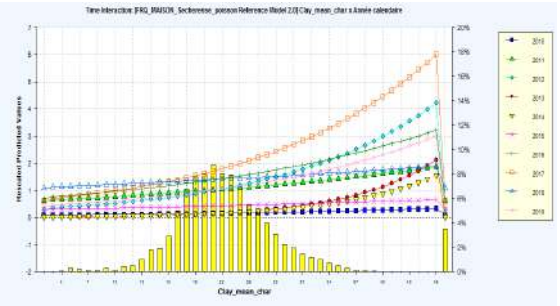


(c) $ESSWI_{1m}$

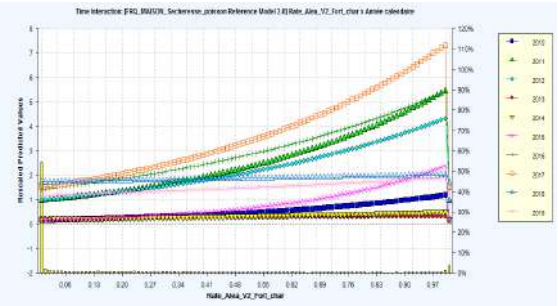


(d) ESTI

FIGURE 4.11 – Stabilité temporelle des variables météorologiques

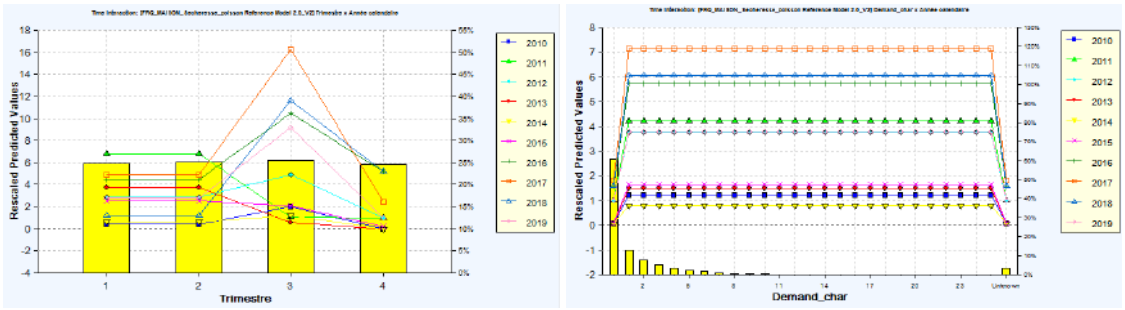


(a) Concentration moyenne d'argile



(b) Exposition à l'aléa fort

FIGURE 4.12 – Stabilité temporelle des variables géologiques



(a) Saison (b) Nombre de demandes

FIGURE 4.13 – Stabilité temporelle des autres variables

Nous pouvons remarquer sur la figure 4.13a que la sécheresse printanière de 2011 ne suit pas la tendance estivale. Toutefois, nous observons les même tendances sur les autres années.

Pour les autres variables, les regroupements effectués sont similaires au cours des années. Nous pouvons remarquer des tendances différentes pour l'année 2015 sur la variable ESPEI ou ESSWI, ceci s'explique par la faible volumétrie de sinistre cette année.

Comme l'ensemble de nos variables et regroupements présentent les même tendances, aucune variable se sera écartée du modèle.

Significativité des coefficients

Le tableau 4.2 montre les estimations des paramètres de notre modèle et leur significativité. Par défaut, le logiciel de tarification Emblem utilise le rapport $\frac{\hat{\sigma}(\hat{\beta})}{\hat{\beta}}$ comme un indicateur de la significativité des coefficients. Plus le pourcentage d'erreur-type est élevé, moins il y a de différences statistiques entre ce paramètre et le niveau de base. Pour un pourcentage d'erreur-type supérieur 50%, l'estimation du paramètre se situe à moins de deux erreurs standards du niveau de base.

Par conséquent, nos coefficients sont jugés significativement différents de la référence lorsque le pourcentage d'erreur-type est inférieur à 50%. Nous pourrons également apprécier la p_{value} du test de Wald afin de tester la significativité d'une variable ou de ses simplifications en annexe.

TABLE 4.2 – Estimation des coefficients et significativité

Coefficient $\hat{\beta}$	$\hat{\beta}$	$\frac{\hat{\sigma}(\hat{\beta})}{\hat{\beta}}$	$\exp(\hat{\beta})$
$\hat{\beta}_0$	-10.55	1.6	$2,62 \cdot 10^{-5}$
$\hat{\beta}_1$	0,6850	11,2	1,9838
$\hat{\beta}_2$	1,1363	6,8	3,1151

$\hat{\beta}_3$	-0,6514	11,6	0,5213
$\hat{\beta}_4$	-0,4306	31,1	0,6501
$\hat{\beta}_5$	1,4008	6,2	4,0585
$\hat{\beta}_6$	0,3417	7,6	1,4074
$\hat{\beta}_7$	-1,2062	6,2	0,2993
$\hat{\beta}_8$	-0,5110	16,5	0,5999
$\hat{\beta}_9$	0,3186	15,5	1,3752
$\hat{\beta}_{10}$	0,6276	7,2	1,8731

avec :
- $\hat{\beta}_0$, l'intercept.
- $\hat{\beta}_1$, le coefficient pour la saison hivernale et printanière.
- $\hat{\beta}_2$, la valeur du coefficient pour la saison estivale.
- $\hat{\beta}_3$, le coefficient associé à l'indicatrice : $\mathbf{1}_{\text{Critère\_}28cm>1}$.
- $\hat{\beta}_4$, la pente de la fonction affine pour la variable Critère_$28cm$.
- $\hat{\beta}_5$, le coefficient pour un nombre de demandes antérieures supérieur à 1.
- $\hat{\beta}_6$, la pente de la droite affine pour la variable correspondant à l'exposition en aléa fort.
- $\hat{\beta}_7$, la pente de la droite affine pour la variable $ESSWI_{1m}$.
- $\hat{\beta}_8$, la pente de la droite affine pour la variable *ESPEI*.
- $\hat{\beta}_9$, la pente de la droite affine pour la concentration moyenne en argile.
- $\hat{\beta}_{10}$, la pente de la droite affine pour la température standardisée.

Le pourcentage d'erreur-type associé à chaque coefficient de notre GLM est inférieur à 40% donc ils sont bel et bien significatifs. Lors de la sélection de variable, nous avons simplifié les modalités pour faire en sorte que ces coefficients soient significatifs à chaque étape.

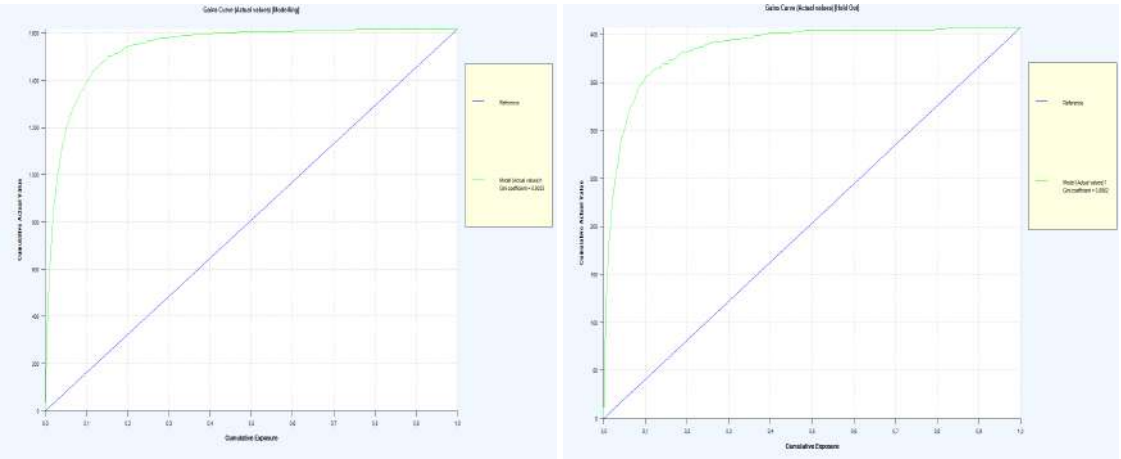
4.1.4 Performances et résidus sur la base de validation

Après avoir validé notre modèle, nous avons fixé les paramètres de ce dernier sur la base d'apprentissage. Désormais, les coefficients $\hat{\beta}_j$ ne sont plus des variables aléatoires mais des constantes. Ceci nous permet de comparer les performances du modèle sur l'échantillon de validation (holdout).

Cette étape a pour objectif de vérifier :
- l'absence de sur-apprentissage ou de sous-apprentissage en comparant l'indice de Gini entre la base d'apprentissage et de test.
- que nos résidus soient centrés en zéro.

Performances sur la base de validation

Les graphiques 4.14a et 4.14b montrent les performances du modèle figé ou « offset » sur l'échantillon d'apprentissage et de validation.



(a) Apprentissage (b) Validation

FIGURE 4.14 – Comparaison de la courbe de gain entre apprentissage et test

Sur ces graphiques, nous observons deux choses :

- La courbe de gain croît très rapidement. En sélectionnant 10% de notre exposition avec les plus grandes prédictions, nous pouvons capter environ 1400 sinistres soit plus de 85% des sinistres contenus dans la base d'apprentissage.
- L'indice de Gini sur l'échantillon hold-out (0,8922) reste proche de celui observé sur l'échantillon d'apprentissage (0,9033). Cela traduit l'absence de sur-apprentissage et de sous-apprentissage via notre modèle.

La valeur élevée de l'indice de Gini s'explique par le fait que les variables utilisées sont très discriminantes. En effet, les communes en rang 1 selon la variable Critère_28cm qui de plus ont déjà fait une demande de reconnaissance par le passé sont peu nombreuses mais concentrent la majeure partie des sinistres.

Analyse des résidus

Avec la volumétrie de notre base, il n'a pas été possible de compiler les résidus « crunched ». Mais l'histogramme des résidus de déviance standardisés représenté sur la figure 4.15 montre que ces derniers sont centrés en zéro. Cela signifie que notre modèle ne sous-prédit pas la fréquence ni ne la sur-prédit.

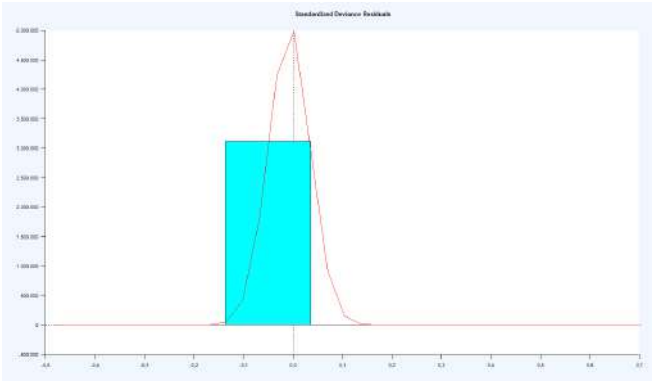


FIGURE 4.15 – Histogramme des résidus

FIGURE 4.16 – Analyse des résidus sur les données de validation

4.1.5 Résultats du modèle et conclusion

Conformément à nos attentes, les indices de sécheresse figurent parmi les variables les plus importantes de notre modèle. Nous pouvons y retrouver :

- Les précipitations nettes standardisées (ESPEI).
- L'indice d'humidité des sols standardisés sur la couche 0-1m.
- La température à 2 mètres du sol standardisée (ESTI).
- La reproduction des critères météorologiques sur la couche 0-28cm.

Finalement les deux plus importantes variables concernent la reproduction des critères météorologiques sur la couche superficielle du sol (0-28cm) ainsi que la connaissance du nombre de demandes de reconnaissance antérieures. La reproduction des critères météorologiques permet de caractériser la potentielle éligibilité au dispositif Cat Nat tandis que la seconde variable permet de déterminer, parmi les communes potentiellement éligibles, celles qui sont souvent sinistrées.

Après avoir vérifié les corrélations, la significativité des coefficients ainsi que la stabilité temporelle de ces derniers, nous avons conclu que les résidus du modèle étaient bien centrés. De plus le modèle n'est pas sujet au sur-apprentissage ce qui est positif concernant sa capacité à se généraliser.

Le graphique suivant donne un aperçu de la fréquence annuelle moyenne sur la base de validation. Si le modèle semble être proche de la fréquence observée pour les années 2010 à 2013, l'écart entre ces deux courbes augmente par la suite. Le modèle a tendance à sur-estimer la fréquence au cours des années peu sinistrées et à sous-estimer les épisodes de sécheresse intense entre 2016 et 2018.

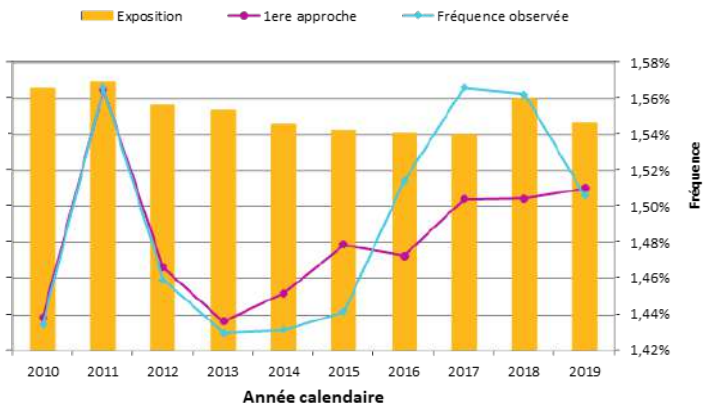


FIGURE 4.17 – Prévisions sur la base de validation

Pour les années 2020 et 2021, tous les arrêtés Cat Nat n'ont pas encore été publiés. La majorité des sinistres sur cette période concerne des sinistres ouverts et une partie d'entre eux seront classés sans-suite en l'absence d'arrêté. En plus de cela, seule une partie des sinistres ont été déclarés étant donnée la cinétique lente de la sécheresse. Ceci rend difficilement comparable les prévisions de notre modèle avec les données réelles.

Le graphique suivant compare les prévisions du modèle avec les données observées sur l'ensemble de la base. Nous remarquons que la fréquence prédite est plus faible pour l'année 2020 et coïncide sur l'année 2021. Il est difficile de tirer des conclusions sur les années de test car il est possible que certains sinistres ouverts deviennent sans-suite en l'absence d'arrêté de publié.

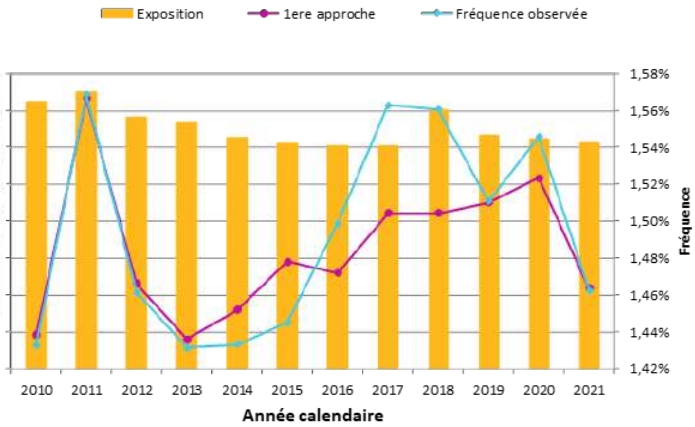


FIGURE 4.18 – Prévisions sur l'ensemble des données

4.2 2e approche : Modèle de détection des arrêtés Cat Nat

La seconde méthodologie propose une alternative au modèle à zéro-inflation en combinant à la fois un modèle de classification supervisé et un modèle de comptage de type Poisson. Le modèle de détection a comme objectif de déterminer les communes qui vont être reconnues en l'état de catastrophe sécheresse afin de retirer la partie intrinsèque d'absence de sinistralité.

Dès lors, le modèle de comptage permettra d'identifier au sein d'une commune reconnue le nombre de sinistres survenus qu'il soit nul ou non.

Les données utilisées pour l'entraînement des forêts aléatoires reposent sur le croisement entre :

- Les données météorologiques contenant à la fois plusieurs indices standardisés sur la saison mais également les rangs de l'humidité des sols sur une période glissante de 50 ans et sur plusieurs couches du sol.
- Les données géologiques contenant :
  - La part de la surface communale sur chacune des couches de la carte d'exposition ainsi que la superficie en question.
  - La concentration moyenne d'argile de la commune sur la couche superficielle du sol.
  - Le nombre de maisons individuelles exposées par zone d'exposition.
  - La proportion de maisons individuelles correspondant à une période de construction par couche de la carte d'exposition.
- Le nombre de demandes de reconnaissance passées (favorable ou défavorable) ainsi que le nombre de jours entre le début du trimestre en question et la dernière publication au journal officiel.
- Les coordonnées XY du centroïde de chaque commune afin d'ajouter une dimension spatiale au modèle.
- Les données de l'Insee contenant la densité de population au km^2 et la proportion de maison parmi les logements.

La base de données à la maille INSEE, année et trimestre contient ainsi 2 682 554 observations réparties entre 2003 et 2021 inclus.

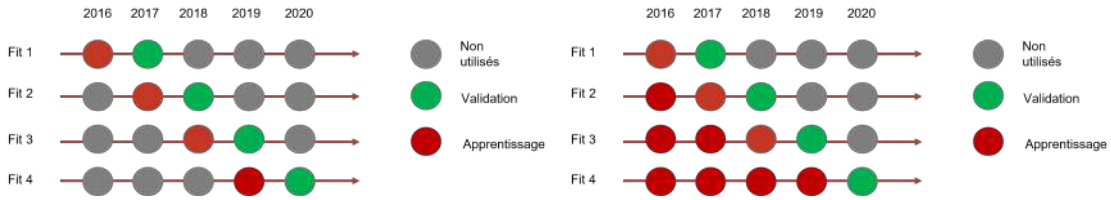
Schéma de validation

Comme les critères de reconnaissance ont évolué au cours du temps, les arrêtés observés au sein de notre base reposent sur des critères hétérogènes. En effet, certaines communes aujourd'hui non reconnues en l'état de catastrophe naturelle aurait pu l'être si d'anciens critères avait été appliqués et inversement. Afin de vérifier la stabilité du modèle au changement de critère, nous avons adopté un schéma de validation croisée spécifique. Celui-ci consiste à répéter les phases d'apprentissage et de validation sur des années distinctes et indépendantes.

Le schéma de validation proposé par la figure 4.19b propose d'ajouter à chaque étape,

l'année de validation précédente. Cette méthode permet de conserver tout l'historique pour prédire sur une année future mais elle est plus coûteuse en temps de calcul. De plus, ce schéma de validation conserve la problématique d'apprentissage sur les critères hétérogènes.

Le schéma 4.19a répond à cette problématique mais peut s'avérer difficile à mettre en place dans le cas où une année est peu marquée par la sécheresse. Il y aura alors trop peu d'arrêtés Cat Nat sécheresse pour l'apprentissage de notre modèle sur l'itération en question.



(a) Validation séquentielle (b) Validation agrégée

FIGURE 4.19 – Schéma de validation croisée séquentielle

Comme les années 2016 à 2020 ont été particulièrement marquées par la sécheresse, nous pouvons nous permettre d'apprendre sur une seule année calendaire et valider sur la suivante. Cette méthode nous permettra de voir si le modèle est capable d'apprendre et de restituer sur d'anciens critères (apprentissage en 2016, validation en 2017). Il nous permettra également de voir si le modèle est capable de prédire correctement lors d'un changement de critère (apprentissage 2017 et validation sur 2018). Enfin il nous permet d'apprécier sa capacité à restituer de bonnes prévisions sur la vision des critères actuellement en vigueur.

4.2.1 Impact du rééchantillonnage sur les métriques

Au sein de notre jeu de données, nous comptons seulement 32207 trimestres et communes reconnues entre les années 2003 et 2021 sur les 2 682 554 individus soit seulement 1,2% de notre base. En nous limitant aux années 2016 à 2020 comme cela a été proposé, cette proportion atteint 2,91% mais il subsiste toujours un fort déséquilibre de classes.

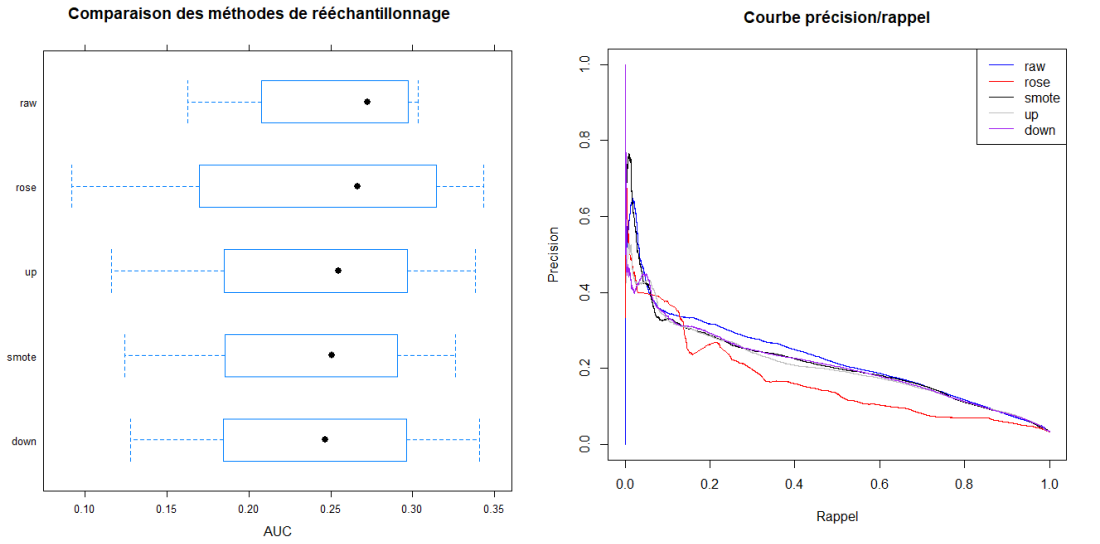
Ce constat nous impose d'utiliser des métriques de performances sensibles à ce déséquilibre et offrant une bonne mesure de comparaison entre les modèles. Nous avons pu définir à la fois le F_1 score à la page ainsi que l'aire sous la courbe précision-rappel à la page 54. En tenant compte du déséquilibre contenu dans nos données, nous pouvons préciser la valeur de ces métriques pour un modèle aléatoire.

$$\text{F}_1 \text{ score aleatoire} = \frac{2 \cdot p}{p+1} = \frac{2 \cdot 2.91\%}{1.0291} \simeq 5.66\%$$

$$\text{AUC aleatoire} = p = 2.91\%$$

En toute conscience de ce déséquilibre de classes, nous avons souhaité comparer plusieurs méthodes de rééchantillonnage sur les performances de notre modèle. Pour ce faire, nous avons commencé par centrer et réduire nos données. Puis, nous avons entraîné les forêts aléatoires sur les données d'apprentissage rééchantillonnées avec le schéma de validation 4.19a.

Ainsi, pour chacune des méthodes de rééchantillonnage, nous disposons de 4 mesures de l'AUC sous la courbe précision/rappel, du F_1 score ainsi que les probabilités associées à chaque base de validation.



(a) AUC sur les différentes méthodes

(b) Courbe précision/rappel sur les folds de validation

FIGURE 4.20 – Résultats du rééchantillonnage sur l'AUC

Nous remarquons sur les graphiques 4.20a et 4.20b que l'aire moyenne sous la courbe précision/rappel est la plus élevé pour le modèle sans rééchantillonnage avec une valeur de $0,2526$ contre $0,2421$ pour le rééchantillonnage ROSE et $0,2379$ pour la méthode SMOTE. Malgré le fait que cette métrique n'est été calculée que sur 4 bases de tests, nous remarquons une dispersion des résultats plus élevée avec le rééchantillonnage.

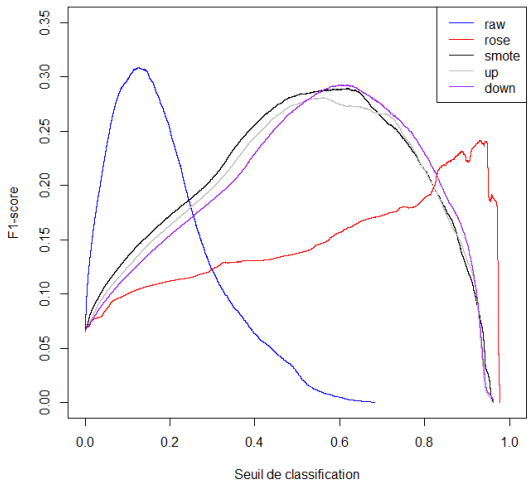


FIGURE 4.21 – Impact du rééchantillonnage sur le score F1

Enfin, nous pouvons constater que les courbes du F_1 score se sont décalées vers la droite en utilisant les méthodes de rééchantillonnage. Pour autant, le score F_1 reste plus élevé sans rééchantilloner les données.

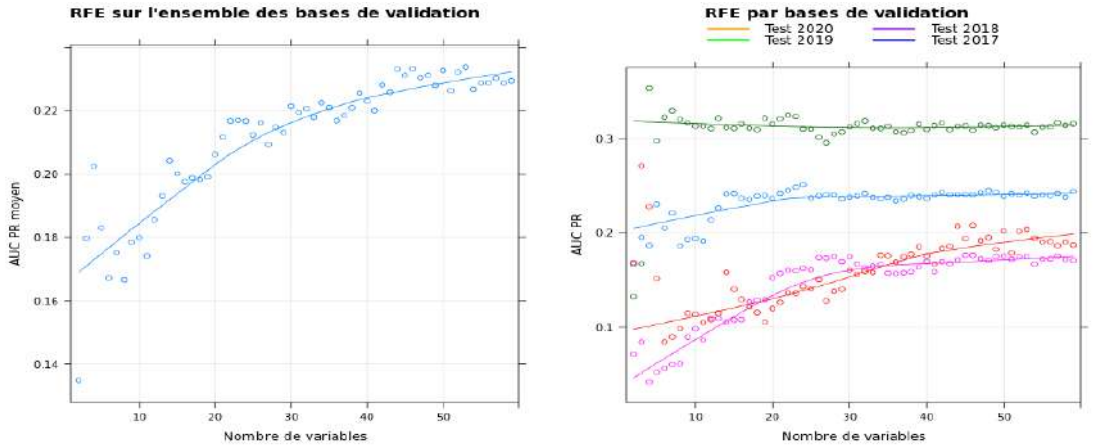
Pour la suite de notre étude, les données ne seront ni rééchantillonnées ni centrées-réduites car ce processus était nécessaire pour l'application des méthodes de rééchantillonnage mais les forêts aléatoires ne sont pas sensibles à l'échelle de nos variables.

4.2.2 Sélection de variables

La sélection de variables est une étape essentielle du processus de modélisation, elle permet de limiter notre base à un sous-ensemble de prédicteurs plus restreint mais offrant d'aussi bonne performance que le modèle complet. Ceci permet d'alléger les temps de calcul ainsi que d'identifier les variables les plus discriminantes. Nous avons fait le choix d'utiliser l'importance de notre modèle pour sélectionner ce sous-ensemble optimal. La méthodologie appliquée repose sur une élimination *backward* récursive des variables les moins importantes aussi connue sous le nom de « *Recursive Feature Elimination* » (RFE).

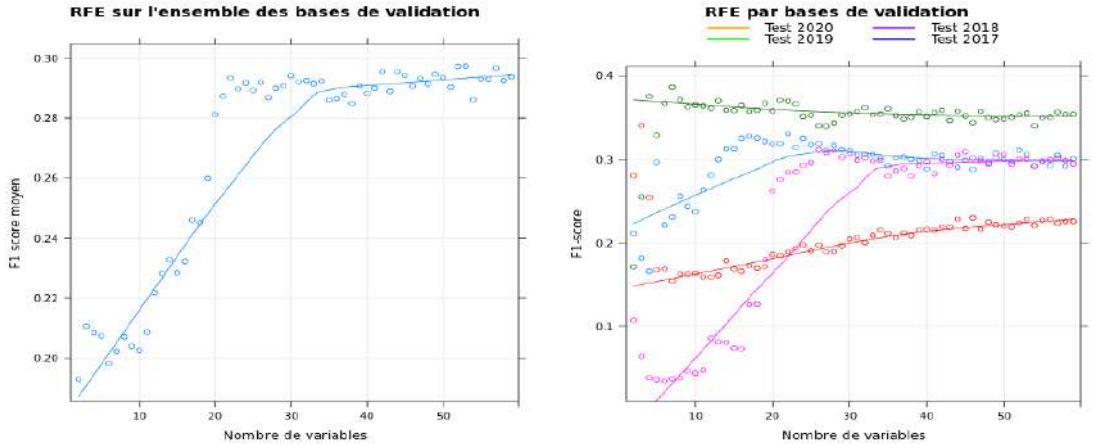
Les figures 4.22a et 4.23b détaillent les résultats de la procédure au travers de la validation croisée séquentielle. Le seuil de classification choisi pour compiler la métrique de F1-score a été fixé arbitrairement à 20%.

Si l'on regarde le niveau des performances sur les bases de tests, nous pouvons remarquer que le modèle restitue de meilleurs résultats lorsque ce dernier apprend et valide sur des bases aux critères homogènes. La baisse de performance sur l'année 2020 est expliquée par le fait que tous les arrêtés n'ont pas encore été publiés donc le modèle pré-



(a) AUC moyen (b) AUC par itérations

FIGURE 4.22 – Sélection de variable avec l'AUC



(a) F1-score moyen (b) F1-score par itérations

FIGURE 4.23 – Sélection de variables avec le F1-score

dit d'avantages d'arrêtés que ce que l'on observe pour le moment (faible précision mais rappel élevé). En revanche, les performances sur l'année 2019 sont relativement stable quelque soit le nombre de variables explicatives. Ce résultat surprenant, nous poussera à choisir un nombre de variables optimal en fonction des performances moyennes.

Sans surprise, le F_1 score moyen ainsi que l'AUC moyen sous la courbe précision/rappel sont croissants avec le nombre de variables explicatives. Nous pourrons retenir le nombre de variables correspondant au point d'inflexion de la courbe du F_1 score ou bien retenir une tolérance sur le niveau d'AUC moyen :

$$S_{tol} = \inf \left\{ S_i \mid \alpha \geq \frac{AUC^{PR}_{optimal} - AUC^{PR}_{S_i}}{AUC^{PR}_{optimal}} \right\}$$

Nous avons choisi de retenir le point d'inflexion sur la courbe du F_1 score comme étant le nombre de variables optimales pour la prédiction. La liste 4.6 en annexe décrit l'ensemble des variables sélectionnées par ordre d'importance sur les années 2016 à 2020 avec cette méthode.

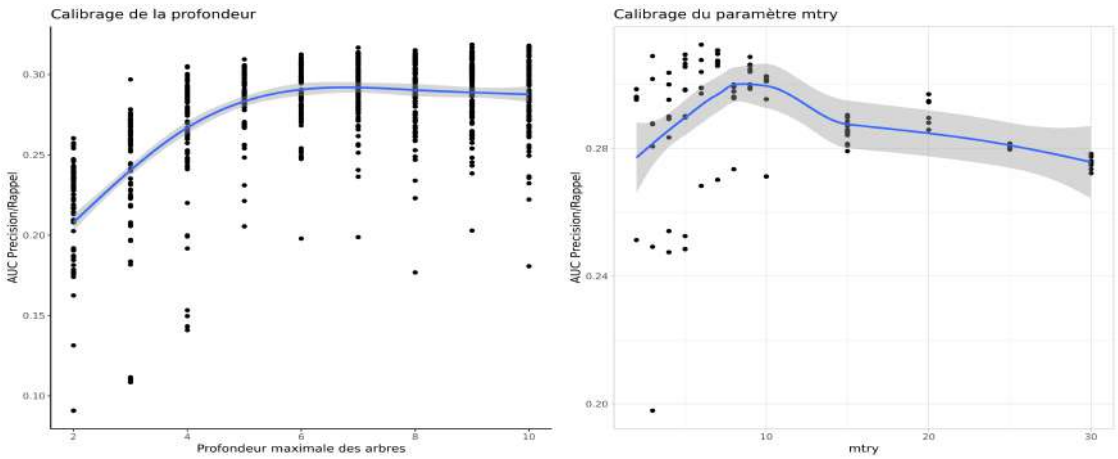
4.2.3 Calibrage des hypers-paramètres

Par la suite, les forêts aléatoires ne seront entraînées que pour les années 2018 à 2019. Nous avons choisi de retreindre la base de données aux années les plus récentes afin que le modèle puisse apprendre à classer les individus sur les critères en vigueur. In fine, l'objectif reste de prédire au mieux les années futures se basant sur ces critères.

Même si le calibrage des hypers-paramètres ne permettra pas d'accroître très significativement les performances de notre modèle, cela reste néanmoins une étape indispensable de la modélisation.

Pour cela, nous avons choisi d'optimiser l'aire sous la courbe précision/rappel, qui est une mesure sensible au déséquilibre de classe et indépendant du seuil de classification. Nous avons fait varier aléatoirement les hypers-paramètres. Cette méthode est communément appelée « *Random Grid Search* » et consiste à effectuer un tirage aléatoire et sans remises parmi l'ensemble des combinaisons définies.

Nous avons commencé par entraîner 1000 modèles sur l'année 2018 récoltant ainsi les valeurs de 1000 métriques sur l'année 2019.

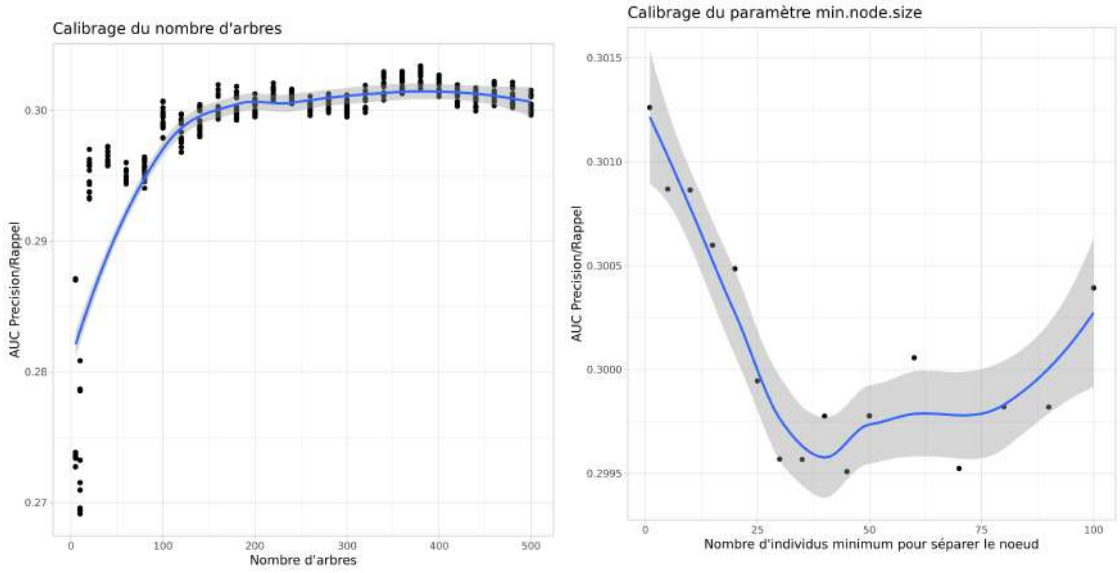


(a) profondeur maximale (b) Nombre de variables tirés aléatoirement

FIGURE 4.24 – Calibrage de la profondeur maximale et du paramètre *mtry*

Indépendamment de la valeur des autres hypers-paramètres, la profondeur maximale égale à 7 possède un bon compromis entre des arbres peu profonds et une valeur de l'AUC élevé. Après avoir filtré les premiers résultats sur cette profondeur maximale, il reste suffisamment de combinaisons explorées pour déterminer la valeur du paramètre *mtry* optimal. Avec un nombre de variables sélectionnées aléatoirement pour chaque arbre égale à 7, le modèle offre les meilleurs performances.

Afin de calibrer le nombre d'arbres que composent la forêt ainsi que le nombre d'instance minimale pour séparer un noeud (*min node size*), nous avons effectué une seconde recherche aléatoire de 500 combinaisons en fixant les valeurs de *mtry* et *max.depth* à 7.



(a) Nombre d'arbres

(b) Nombre d'instances minimale pour séparer les noeuds

FIGURE 4.25 – Calibrage du nombre d'arbres et du nombre d'instance minimale

Nous pouvons remarquer que les performances commencent à se stabiliser à partir d'un nombre d'arbres égale à 150. Cependant, nous sélectionnerons une profondeur égale à 200 pour la suite de la modélisation.

Après avoir filtré sur ce nombre d'arbre, il ne reste plus beaucoup de combinaisons explorées mais nous pouvons remarquer que les performances varient très peu en fonction de ce dernier paramètre. Le nombre d'instances minimales pour pouvoir séparer un noeud est étroitement lié à la profondeur maximale des arbres. Par conséquent, il n'a que peu d'intérêt car les arbres seront construits jusqu'à la profondeur maximale égale à 7.

Le modèle finale est donc entraîné sur les années 2018 et 2019, pour lesquels les critères de reconnaissance sont homogènes, avec les paramètres suivants :

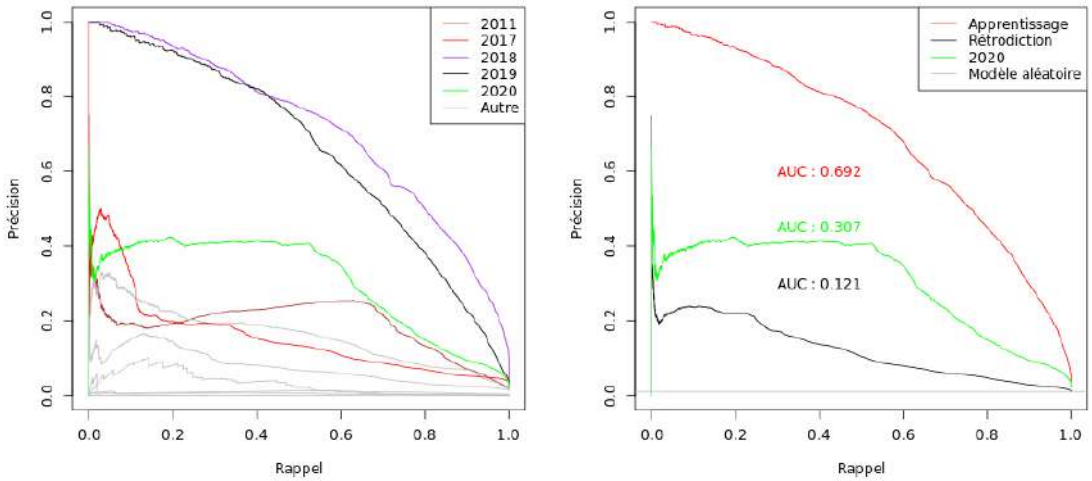
- $max.depth = 7$
- $mtry = 7$
- $ntrees = 200$
- $min.node.size = 30$

4.2.4 Performances et choix du seuil de classification

Performances du modèle

Après avoir calibré notre modèle, nous allons comparer les performances de celui-ci sur la base d'apprentissage, sur les années antérieures ainsi que sur l'année test de 2020. Le contrôle des performances sur les différentes bases permettra de constater si le modèle est propice au sur ou sous-apprentissage. Ensuite, nous devrons choisir un seuil de classification adapté afin d'attribuer les classes aux individus à prédire.

La figure 4.26a montre l'aspect des courbes précision-rappel pour chaque année. Nous pouvons remarquer que les aires sous les courbes d'apprentissage sont supérieurs aux autres. Même si ce constat est vérifié avec la figure 4.26b, nous ne pouvons pas conclure sur un éventuel sur-apprentissage de notre modèle. Si le modèle avait appris par coeur les données d'apprentissage alors la valeur de l'AUC aurait été plus importante encore. De plus, la valeur de l'AUC est bien au delà de sa valeur aléatoire. Il ne faut pas oublier que la variable cible est difficile à anticiper c'est pourquoi ces résultats restent satisfaisant.



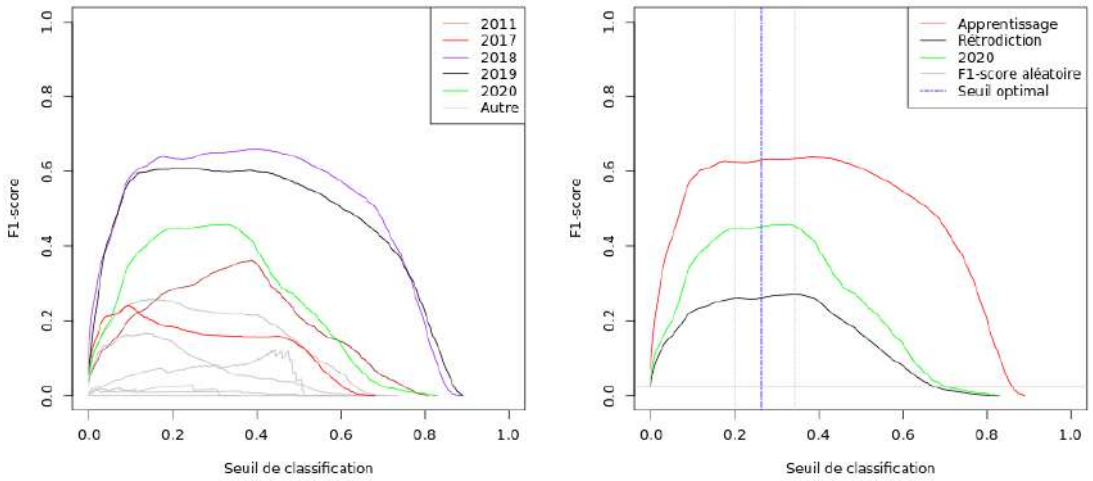
(a) (b)

FIGURE 4.26 – Courbe précision-rappel et AUC

Choix du seuil de classification

Le choix du seuil de classification est une étape très importante puisqu'il conditionne le niveau du F1 score ainsi que le nombre d'arrêtés prédits par le modèle. Si le seuil est trop faible, le modèle prédira trop d'arrêtés Cat Nat et inversement.

Les figures 4.27a et 4.27b montrent une fois de plus l'écart de performances entre les données d'apprentissage et de validation. Même si cet écart est non négligeable, le F1-score reste bien au delà du modèle non informatif et la modélisation à son intérêt. Le modèle restitue de meilleurs performances en prédiction plutôt qu'en rétrodiction.



(a) (b)

FIGURE 4.27 – Courbe de F1 score et seuil de classification

Comme nous le montre la figure 4.27a, non seulement le seuil de classification par défaut n'est pas toujours un bon choix mais le seuil optimal n'est pas identique chaque année. Plutôt que de prendre le seuil maximal sur l'année 2020, nous avons choisi de prendre le seuil maximisant le F_1 score sur l'ensemble des années n'ayant pas servi à la modélisation de sorte à ce qu'en moyenne ce seuil soit optimal. **Le seuil retenu a été fixé à 0.2639561** pour la suite de l'étude et correspond à la droite bleue sur la figure 4.27b.

4.2.5 Comparaison observés/prédits

Le tableau 4.3 récapitule l'ensemble des performances de notre modèle entre les années 2010 et 2020. Nous pouvons constater que nos métriques sont supérieures au modèle non informatif quelque soit l'année.

Si nous regardons les rétrodictions du modèle, nous remarquons qu'il n'y a pas toujours le même ordre de grandeur entre les faux positifs et les faux négatifs. Parfois le modèle a tendance à sur-prédire le nombre de saisons reconnues Cat Nat comme en 2011 tandis que d'autre fois le modèle sous-prédit comme en 2017. Ce constat est directement lié au seuil de classification qui n'est pas toujours le meilleur suivant l'année.

Année	TP	FP	FN	F1 score	F_1^{alea}	AUC	AUC^{alea}
2010	1	29	109	1.4%	0.15%	1.1%	0.07%
2011	1935	8030	766	30.55%	3.8%	19.09%	1.94%
2012	477	1950	1270	22.85%	2.48%	16.16%	1.25%
2013	0	14	23	0	0.03%	2.28%	0.01%
2014	28	607	57	7.77%	0.12%	3.72%	0.06%
2015	6	807	403	0.98%	0.59%	1.02%	0.29%
2016	137	881	1645	9.79%	2.53%	7.64%	1.28%
2017	695	2547	4444	16.59%	7.12%	15.65%	3.69%
2018	4425	3135	1776	64.31%	8.53%	71.68%	4.45%
2019	2373	1290	1794	60.61%	5.81%	65.71%	2.99%
2020	1578	2461	1366	45.19%	4.14%	30.69%	2.11%

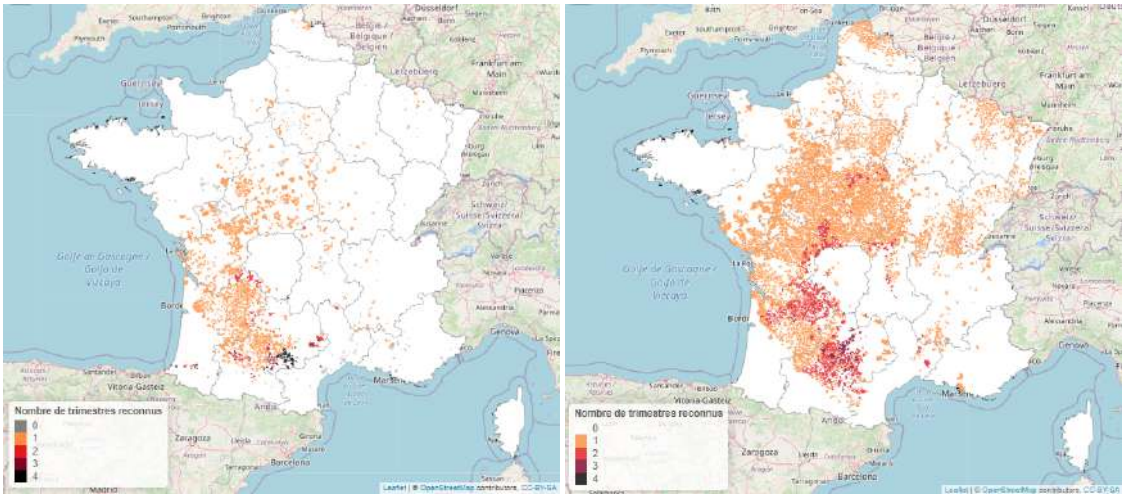
TABLE 4.3 – Résumé des prévisions du modèle

Les faux négatifs ont deux origines plausibles :

- Le modèle n'a pas détecté de conditions météorologiques anormales (éligibilité météorologique). Ceci peut survenir en cas de disparités avec les critères uniformes utilisés par la commission interministérielle.
- Le modèle n'a pas détecté la formulation de la demande. Soit la commune est primo-demanderesse, soit le sinistre survenu est un cas atypique.

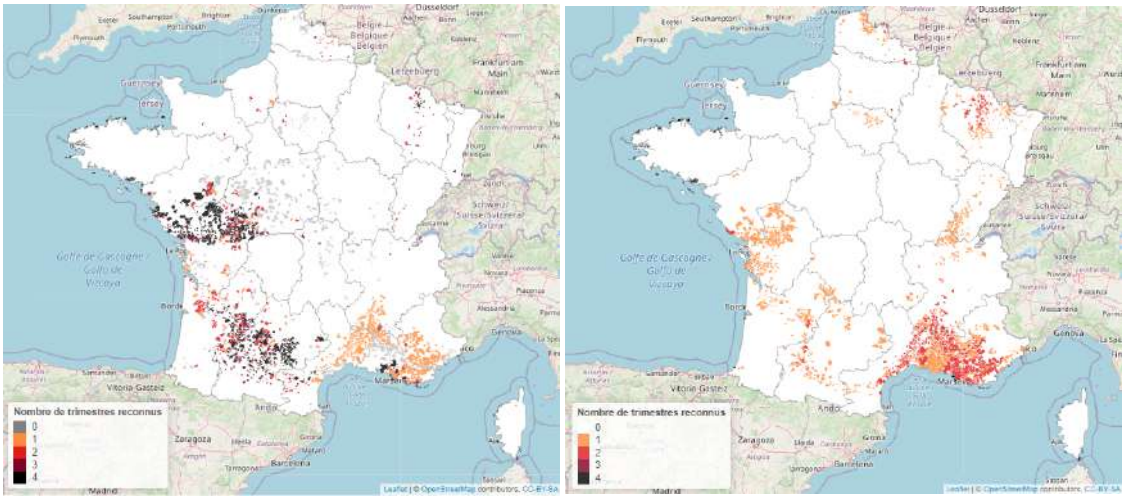
Les faux positifs correspondent à une détection à tord de l'éligibilité vis-à-vis des critères de reconnaissance pour les communes demanderesses et/ou à une détection à tord de la demande de reconnaissance.

Les graphiques suivant comparent les prévisions de notre modèle avec les arrêtés observés ainsi que les demandes défavorables. Nous invitons le lecteur à consulter l'ensemble des prévisions en annexe à la page 122.



(a) Arrêtés observés - 2011 (b) Rétrodiction - 2011

FIGURE 4.28 – Comparaison entre observés et prédits sur l'année 2011

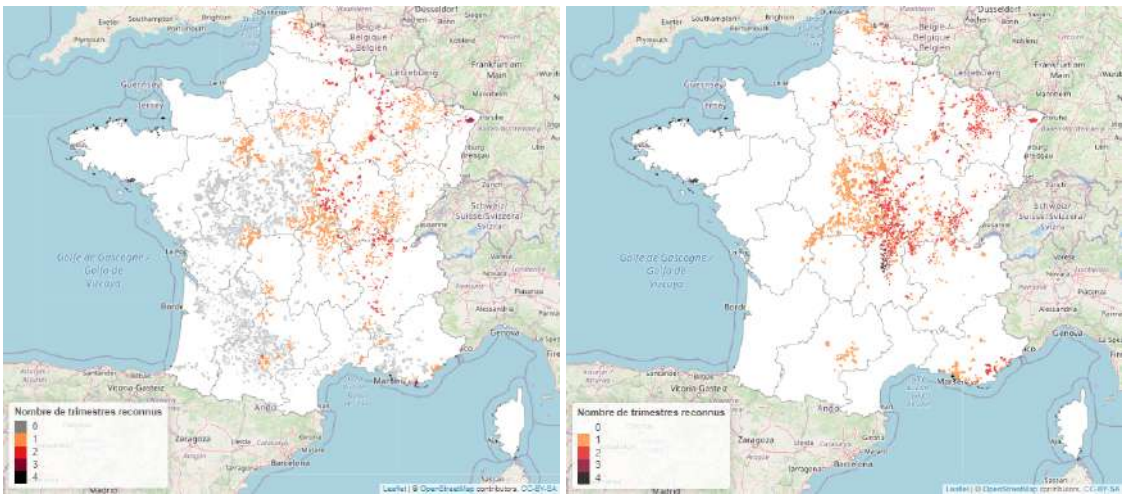


(a) Arrêtés observés - 2017 (b) Rétrodiction - 2017

FIGURE 4.29 – Comparaison entre observés et prédits sur l'année 2017

Pour l'année 2011, notre modèle détecte bien les bonnes zones géographiques mais ces dernières sont élargies par rapport à ce que l'on observe à cause du seuil de classification qui n'est pas optimal. Les erreurs du modèle sont localisées sur des zones pour lesquelles les communes n'ont pas formulé de demandes de reconnaissance. Le modèle semble avoir faussement détecté les demandes mais nous ne pouvons pas vérifier si il s'est trompé sur l'éligibilité envers les critères.

Pour l'année 2017, notre modèle semble se tromper sur l'éligibilité vis-à-vis des critères de reconnaissance. En effet, certaines communes du nord de la France ont été faussement prédites alors que leurs demandes ont reçu un avis défavorable par la commission. A l'inverse, le modèle n'a pas détecté assez de saison dans la région ouest et sud-ouest.



(a) Arrêtés observés - 2020 (b) Prévisions - 2020

FIGURE 4.30 – Comparaison entre observés et prédits sur l'année 2020

Les prévisions de notre modèle sur l'année 2020 sont très correctes puisque les communes ayant reçues un avis défavorable n'ont pas été faussement prédites. Au globale, nous observons une bonne cohérence spatio-temporelle sur cette année. Ce résultat est encourageant pour les prévisions sur les années futures utilisant les mêmes critères de reconnaissance.

4.2.6 Conclusion

Lors de la sélection de variables, la validation croisée séquentielle a mis en évidence l'instabilité temporelle de notre modèle. En effet, l'hétérogénéité des critères de reconnaissance au cours du temps ne facilite pas l'apprentissage du modèle. C'est la raison pour laquelle nous avons souhaité apprendre sur une base récente, plus restreinte afin que les prédictions futures s'effectuent avec des critères comparables.

Même si les performances du modèle peuvent sembler faible à priori, nous avons pu montrer qu'elles étaient bien supérieures à celle proposée par un modèle aléatoire. Le choix du seuil de classification a également un impact sur les prévisions. Si celui-ci semble correcte pour les prédictions futures, les rétrodictions sont plus hétérogènes avec trop d'arrêtés prédits en 2011 (seuil trop faible) et pas assez d'arrêtés prédits en 2017

(seuil trop important).

Finalement, les métriques de notre modèle ne tiennent pas compte de la proximité géographique entre nos prévisions et la réalité. Que notre modèle puisse se tromper c'est une chose mais qu'il se trompe grossièrement dans l'espace en est une autre.

4.3 2e approche : Modèle fréquence

Après avoir construit le modèle de détection, nous devons introduire un modèle de comptage afin de déterminer le nombre de sinistres parmi les communes identifiées Cat Nat. Ce second modèle de fréquence repose sur les données risques augmentées d'indices de sécheresse auquel nous avons appliqué un filtre sur les communes et périodes reconnues en l'état de catastrophe naturelle.

Ce filtre est appliqué en fonction des arrêtés observés et non pas modélisés. Ceci aura pour effet :

- d'émettre une hypothèse forte sur le modèle de détection. Effectivement, l'apprentissage du GLM reposera sur la réalité mais les prévisions pour les années futures s'effectueront sur la base des prévisions de la classification. Toute erreur de classification entraînera soit une sur-estimation de la fréquence soit une sous-estimation de celle-ci.
- de neutraliser le pouvoir prédictif des indices de sécheresse puisque l'application de ce filtre place nos données dans les conditions météorologiques les plus défavorables.

4.3.1 Sélection de variables

Comme pour le premier modèle GLM, nous avons appliqué la même méthodologie de sélection de variables.

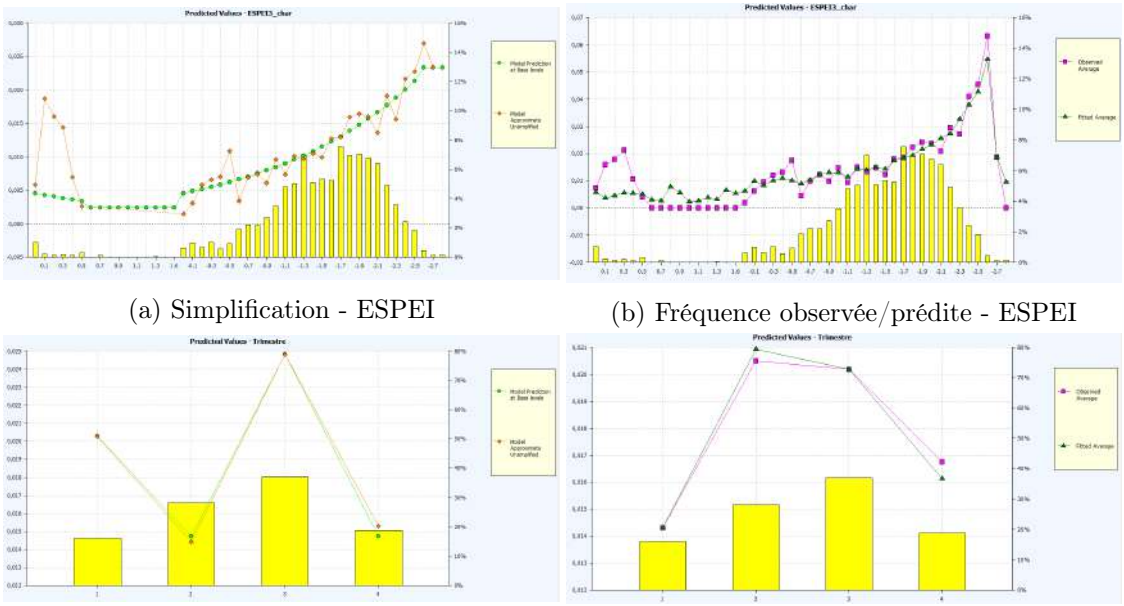
Le tableau 4.4 résume l'impact de chaque variable sur le modèle.

Étapes	Variables	Sélection forward			Post simplification		
		AIC	BIC	Déviance	AIC	BIC	Déviance
Étape 1	Zonier sécheresse	-329	- 253	-328	-257	-246	-259
Étape 2	ESPEI	-293	+182	-314	-308	-287	-312
Étape 3	Risque 5	-120	-88	-122	-122	-101	-126
Étape 4	Risque 6	-58	-15	-58	-60	-60	-64
Étape 5	Année calendaire	-59	38	-58	-99	-99	-105
Étape 6	Risque 4	-26	-5	-26	-38	-16	-42
Étape 7	Trimestre	-51	+122	-51	-118	-97	-122
Étape 8	Risque 3	-2	+203	-1	-10	-0,19	-12

Table 4.4 – Sélection forward

Parmi l'ensemble des variables ajoutées aux données de risques, seules les précipitations nettes standardisées ainsi que la saison ont été sélectionnées. Comme nous avons pu l'évoquer, le filtre a pour effet de placer les données dans les queues de distribution de nos variables météorologiques et neutralise leur effet.

Les graphiques de la figure 4.31 donnent un aperçu des simplifications utilisées et comparent la fréquence prédite avec la fréquence observée. Nous pouvons tout de même apercevoir une augmentation de la fréquence avec l'intensité de la sécheresse au travers de l'indice $ESPEI$. L'application du filtre a eu un effet également sur la simplification de la saison. En effet, par un levier exposition, la fréquence observée de la saison printanière est proche de la saison estivale. Ceci est directement lié à l'année 2011 qui a été fortement sinistrée.



(a) Simplification - ESPEI

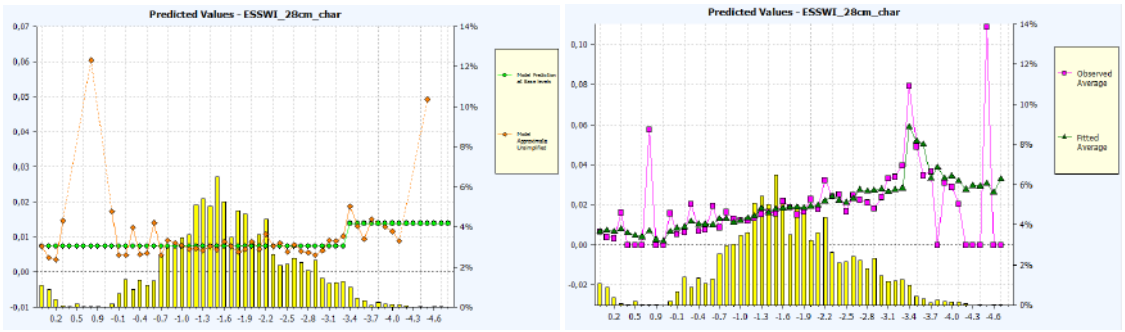
(b) Fréquence observée/prédite - ESPEI

(c) Simplification - Trimestre

(d) Fréquence observée/prédite - Trimestre

FIGURE 4.31 – Simplification des précipitations nettes standardisés et de la saison

A la suite de la sélection *forward*, nous avons ajouté les capitaux mobiliers, la variable « Risque 2 » et l'indice d'humidité des sols standardisés sur la couche 0-28 cm car les performances sur les différentes métriques se sont également améliorées après le regroupement des modalités.



(a) Simplification - ESSWI 28 cm (b) Fréquence observée/prédite - ESSWI 28 cm

FIGURE 4.32 – ESSWI 28cm

Le tableau 4.5 montre les résultats de la sélection *backward* sur notre modèle :

Variable	Sélection backward		
	AIC	BIC	Déviance
Risque 5	18	-3	18
Risque 2	16	5	16
Risque 3	17	6	17
Risque 1	22	11	22
$ESSWI_{28cm}$	27	17	27
Risque 4	30	19	30
Trimestre	55	44	55
Risque 6	65	55	65
Année calendaire	135	102	134
ESPEI	166	145	166
Zonier exposition	323	312	323

TABLE 4.5 – Sélection backward

Si l'on retire la variable « Risque 5 » alors le BIC s'améliore, toutefois cette perte est marginale au vue du gain de l'AIC donc nous pouvons la laisser dans le modèle pour le moment.

Pour conclure, la filtration sur les périodes déjà reconnues en l'état de catastrophe naturelle empêche d'inclure davantage de variables météorologiques à nos données. Seules les précipitations nettes standardisées et dans une moindre mesure l'humidité des sols superficiels sont discriminantes ce qui signifie que plus la sécheresse est intense au sens de cet indice, plus la fréquence est élevée.

4.3.2 Validation des hypothèses

Corrélation des variables

De la même manière, nous n'afficherons le V de Cramer que pour les variables figurant dans le modèle. Comme nous le montre la figure 4.33, les variables de notre modèle respectent le niveau de corrélation limite de 0,7.

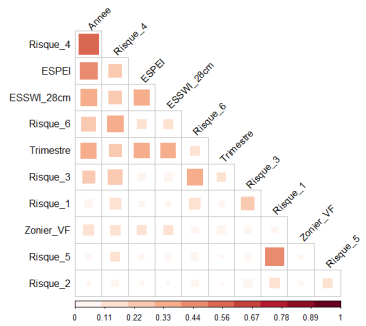
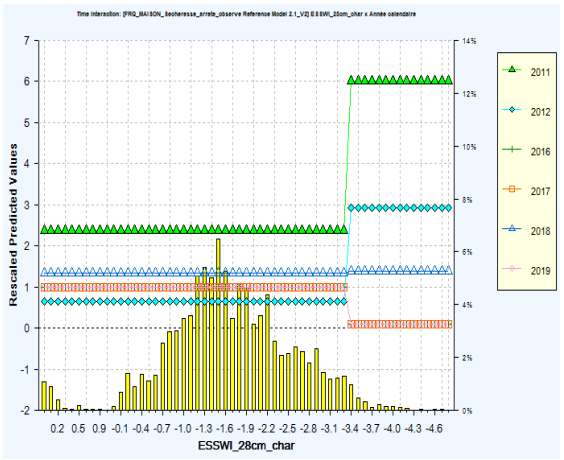


FIGURE 4.33 – V de Cramer

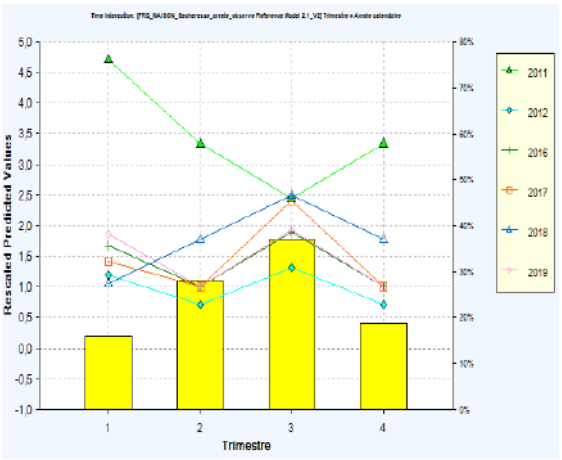
Stabilité temporelle

Les figures ci-dessous montre la stabilité temporelle des indices météorologiques ainsi que la variable trimestre. De la même manière que pour le modèle sans filtration, nous ne tenons pas compte des années 2010, 2013, 2014 et 2015 car le manque de sinistres provoque des instabilités.

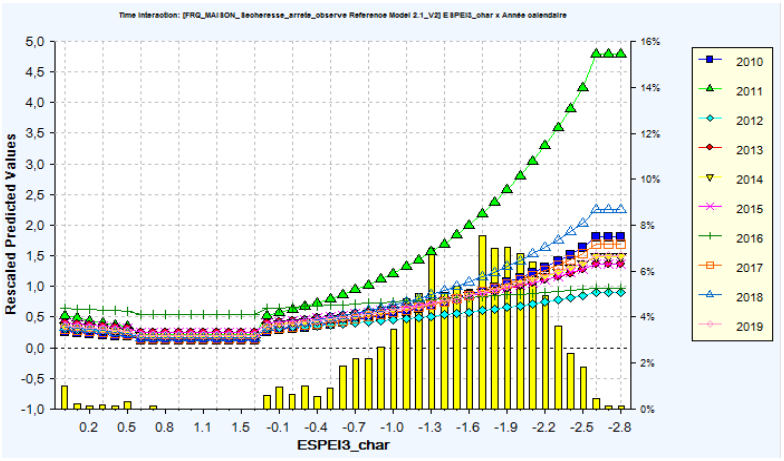
La saison ainsi que la variable $ESSWI_{28cm}$ sont exclus du modèle car les tendances ne sont pas identiques selon les années. En effet, l'année 2011 se distingue par son épisode printanier tandis que l'année 2018 ne suit pas la même tendance sur la période hivernale. En ce qui concerne l'indice d'humidité des sols sur la couche superficielle, la plupart des années n'ont pas d'augmentation de fréquence sur la queue de distribution. Finalement seule la variable $ESPEI$ sera ajoutée aux données risques.



(a) $ESSWI_{28cm}$



(b) Saison



(c) Précipitation nettes standardisés

FIGURE 4.34 – Stabilité temporelle (2e approche)

Significativité des coefficients

Par mesure de confidentialité, nous n'afficherons pas les valeurs des coefficients estimés pour les variables de risques intervenant dans notre modèle. Le tableau 4.6 donne un aperçu du coefficient associé à la variable $ESPEI$ ainsi que sa significativité.

Coefficient $\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\frac{\hat{\sigma}(\hat{\beta})}{\hat{\beta}}$	$\exp(\hat{\beta})$
$\hat{\beta}_0$	-4,5699	1.5	0,0104
$\hat{\beta}_1$	-0.5725	9.4	0.5641

TABLE 4.6 – Coefficients du modèle

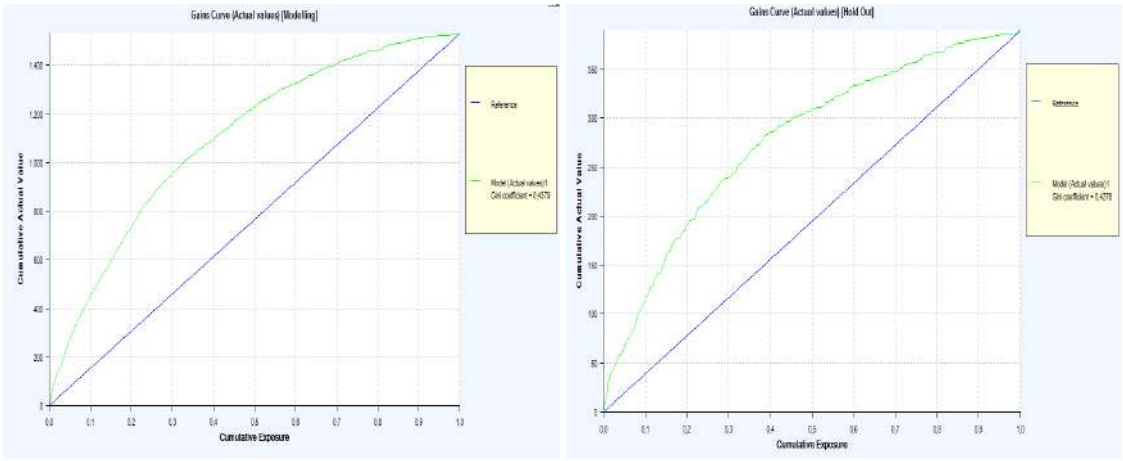
avec :
- $\hat{\beta}_0$, l'intercept.
- $\hat{\beta}_1$, le coefficient du polynôme de degré 1 attribué à la variable $ESPEI$.

Les erreurs-types des coefficients de notre GLM sont inférieures à 50% donc ceux-ci sont significatifs pour notre modèle. De plus comme nous le montre la valeur prise par ces coefficients, la fréquence augmente lorsque l'intensité de la sécheresse augmente selon la variable $ESPEI$.

4.3.3 Performances et résidus sur la base de validation

Performances sur la base de validation

Après avoir fixé la valeur de nos coefficients sur la base d'apprentissage, nous pouvons comparer les courbes de gain ainsi que l'indice de Gini entre l'échantillon d'apprentissage et de validation.



(a) Apprentissage (b) Validation

FIGURE 4.35 – Comparaison de la courbe de gain entre apprentissage et test

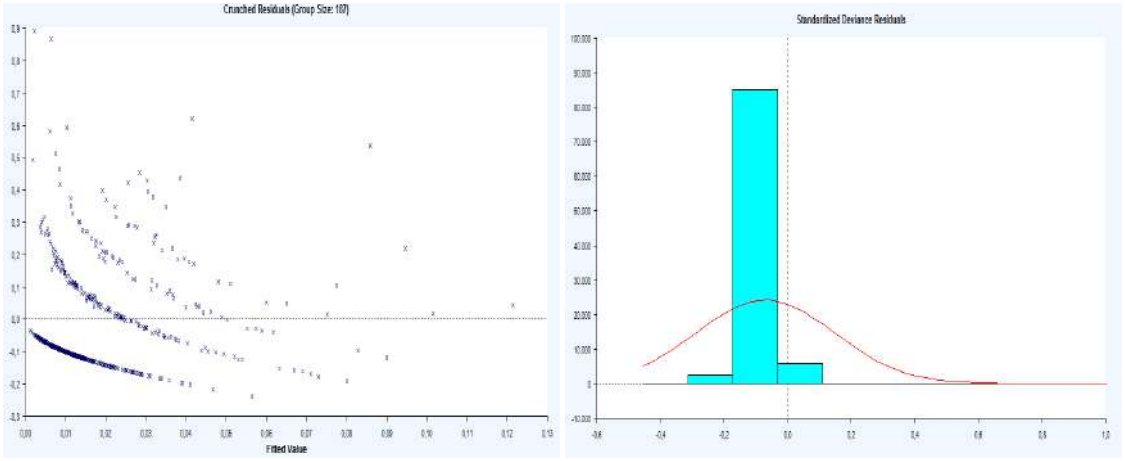
Nous remarquons que l'allure des deux courbes sont similaires, il n'y a donc pas de sur-apprentissage avec notre modèle. En revanche l'indice de Gini vaut environ 0.44 ce qui est bien inférieur à l'indice de Gini du modèle de fréquence abordé précédemment. En sélectionnant les 10% de notre exposition avec les plus grandes fréquences prédites par le modèle, nous captons environ 450 sinistres sur la base d'apprentissage.

Si le niveau de l'indice de Gini est plus faible, il n'est pas pour autant comparable au précédent modèle de fréquence. En effet, la population étudiée n'est pas identique donc l'exposition est totalement différente. La courbe de gain et l'indice de Gini n'est pas comparable avec la première méthode. Il aurait fallu appliquer le modèle restreint

aux arrêtés Cat Nat à l'ensemble de nos données puis modifier les prévisions de ce dernier de sorte à ce que la fréquence prédite soit nulle en l'absence d'arrêté Cat Nat.

Analyse des résidus

Comme pour ce modèle les données ont été filtrées sur les arrêtés observés, le logiciel Emblem nous permet d'afficher cette fois-ci les résidus « crunched » en plus de l'histogramme des résidus de déviance standardisés. Nous pouvons constater que les résidus de déviance standardisés ne sont pas tout à fait centré en zéro ce qui traduit une légère sur-estimation de la fréquence car les résidus sont en moyenne négatifs. Le graphique des résidus « crunched » avec 500 groupes aboutit à la même conclusion.



(a) Résidus « crunched » (b) Résidus de déviance standardisés

FIGURE 4.36 – Analyse des résidus

4.3.4 Résultats du modèle et conclusion

Nous avons vu que les deux modèles de fréquence n'étaient pas comparables entre eux de manière directe car les deux populations étudiées sont différentes. Nous avons donc fait le choix d'appliquer le modèle de fréquence post-détection sur l'ensemble de notre base de données puis de multiplier la fréquence prédite par la probabilité de déclarer un arrêté Cat Nat.

La probabilité de déclarer un arrêté Cat Nat est issue des prévisions de notre forêt aléatoire. Nous avons choisi de conserver les probabilités brutes ainsi que l'application du seuil de classification. La classification ré-attribue les probabilités en deçà du seuil à la valeur 0 (pas d'arrêté) et les autres à la valeur 1. Le graphique 4.37 compare la fréquence annuelle moyenne observée avec la fréquence prédite par l'utilisation du modèle de détection et du modèle linéaire généralisé.

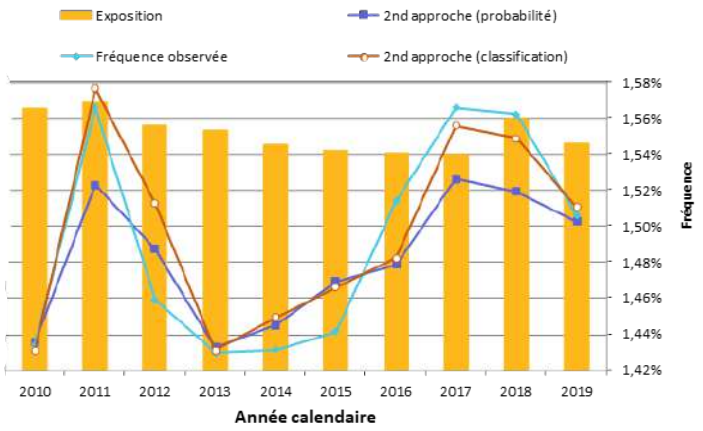


FIGURE 4.37 – Fréquence moyenne sur la base de validation

Nous pouvons remarquer que la fréquence prédite suit bien la tendance observée au fil des années. Comme le modèle de détection a appris sur les années 2018 et 2019, il faut regarder l'écart entre l'observé et le prédit sur les années antérieures. Le passage de la courbe bleue marine à la courbe orange montre l'effet de la classification sur la probabilité d'être concernée par un arrêté Cat Nat. Nous constatons que l'application du seuil de classification améliore nos prévisions en ré-haussant la fréquence.

Le graphique 4.38 compare les prévisions du modèle avec l'observé sur l'ensemble de la base. Pour les années tests, nous remarquons que la fréquence prédite par le modèle est plus faible que la fréquence observée mais assez proche. Pour l'année 2021 la fréquence prédite est relativement faible. Même si nous n'avons qu'une information partielle sur le nombre de reconnaissance, à la vue des indices de sécheresse l'année 2021 semble être moins impactée.

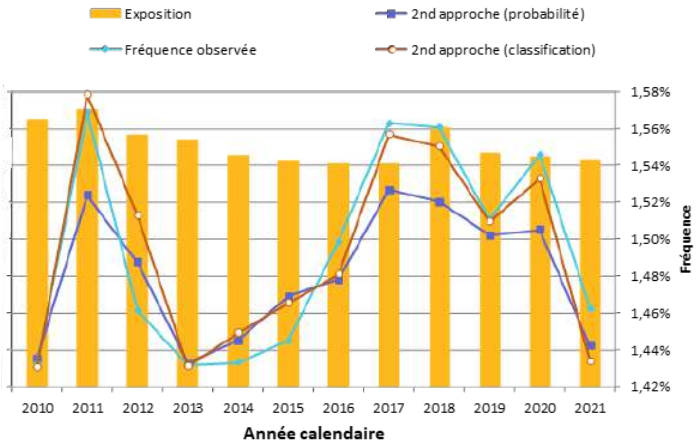


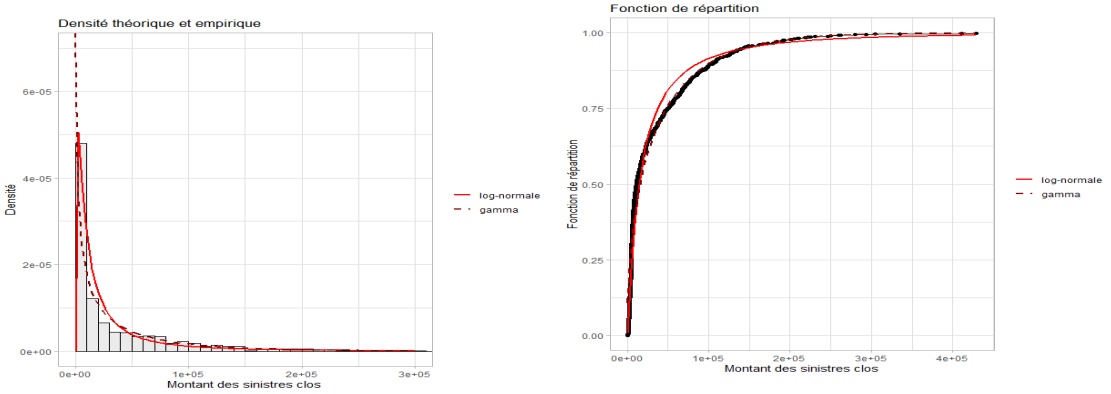
FIGURE 4.38 – Fréquence moyenne sur la base complète

4.4 Modèle de sévérité

Après avoir estimé la fréquence avec l'une des deux approches définies précédemment, nous devons à présent modéliser la sévérité c'est à dire estimer le coût moyen du sinistre en fonction des indices de sécheresses et des caractéristiques de l'individu.

En temps normal, la modélisation doit s'effectuer avec la variable cible représentant le coût moyen des sinistres clos car ces sinistres représentent la charge réelle payée par l'assureur à l'assuré et non pas une valeur estimative comme le sont les sinistres ouverts. Cependant, nous avons déjà très peu de sinistres (environ 2000 réparti sur 10 ans) et les sinistres clos ne représentent que 37,8% de ce volume. En nous limitant aux sinistres clos, la modélisation a peu de chance d'être concluante. Pour pouvoir intégrer les sinistres ouverts à la modélisation, il ne faut pas que leurs montants forfaitaires biaisent la distribution de la charge. Autrement dit, nous devons nous assurer qu'il n'y ait pas une seconde « bosse » sur la distribution empirique correspondant aux montants forfaitaires car cela se répercuterait inévitablement sur les résidus du modèle.

Les graphiques de la figure 4.39 comparent la distribution des sinistres clos avec les densités théoriques de la loi gamma et log-normale dont les coefficients ont été estimés en maximisant la vraisemblance ou via la méthode des moments le cas échéant.



(a) Histogramme à densité

(b) Fonction de répartition théoriques et empiriques

FIGURE 4.39 – Comparaison de la distribution des montants clos avec la loi gamma et log-normale

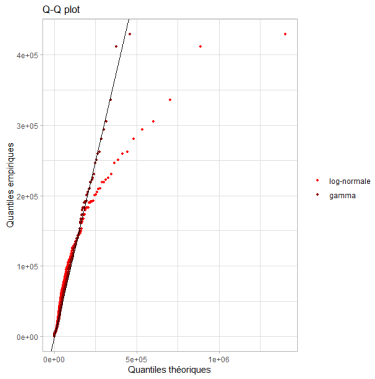
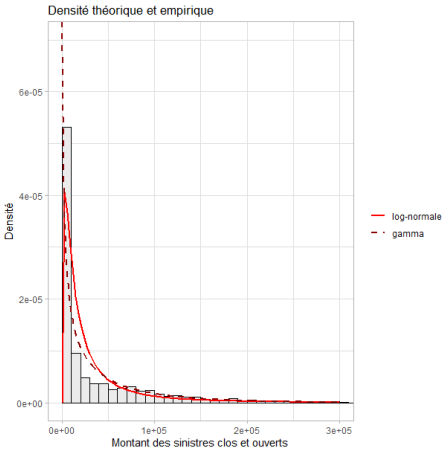
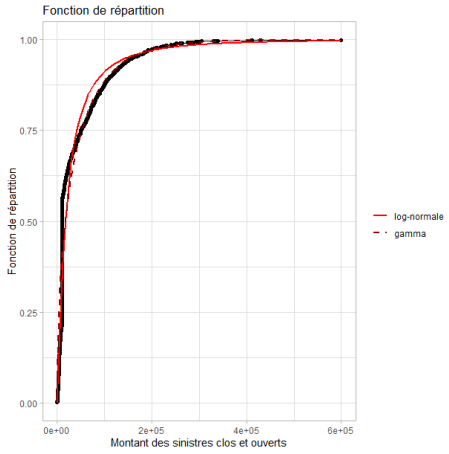


FIGURE 4.40 – Montant clos : diagramme quantiles-quantiles

Nous pouvons remarquer visuellement que la charge de nos sinistres clos se rapproche plus d'une loi gamma. Nous pouvons tester l'adéquation de notre échantillon à la loi théorique avec le test de kolmogorov-smirnov. La p_{value} du test vaut respectivement $2.2 \cdot 10^{-16}$ et $7.089 \cdot 10^{-7}$ pour l'adéquation à la loi gamma puis à la loi log-normale ce qui nous conduit à rejeter l'hypothèse nulle selon laquelle la fonction de répartition empirique est identique aux fonctions de répartition théoriques. En revanche, la statistique du test de kolmogorov-smirnov est la plus faible pour l'adéquation à la loi log-normale avec $D_{KS} = 0.092338$ contre $D_{KS} = 0.14999$ pour la loi gamma. Malgré la divergence entre la statistique du test et les résultats visuels, nous adopterons la loi gamma car les quantiles empiriques et théoriques sont plus proches.



(a) Histogramme à densité



(b) Fonction de répartition théoriques et empiriques

FIGURE 4.41 – Comparaison de la distribution des montants clos et ouverts avec la loi gamma et log-normale

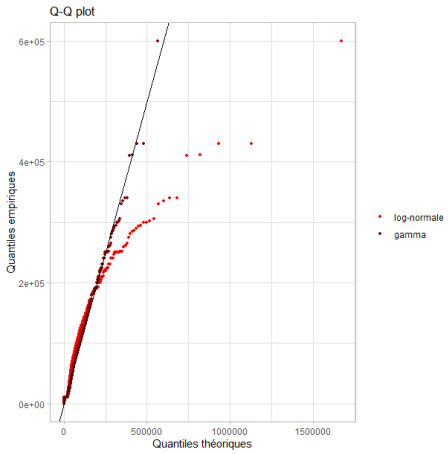


FIGURE 4.42 – Montant sinistres ouverts et clos : Diagramme quantiles-quantiles

En ajoutant les sinistres ouverts, nous nous apercevons que la distribution empirique n'est pas biaisé par les montants forfaitaires. A la vue du diagramme quantile-quantile, nous concluons sur l'adéquation du montant des sinistres clos et ouvert avec la loi gamma.

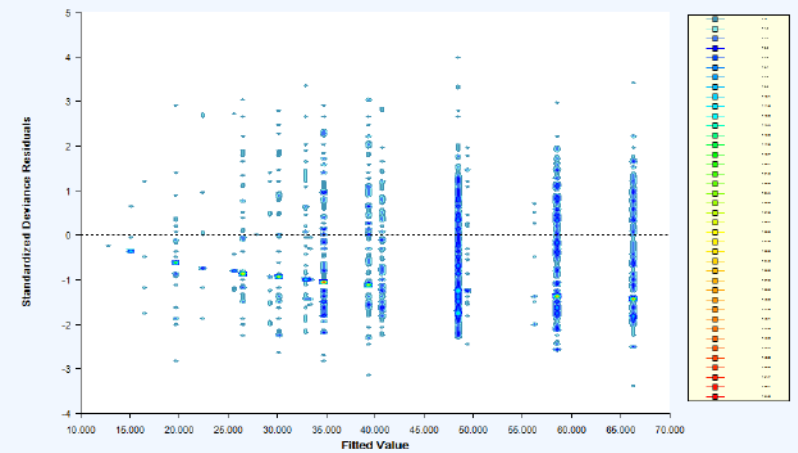


FIGURE 4.43 – Résidus de déviance standardisés du modèle de sévérité

Malgré cet effort, seulement quelques variables de risques sont sélectionnées dans notre modèle et leurs impacts sur les métriques restent limités. Les résidus de déviance standardisés ne sont pas centrés en zéro et présentent des clusters de résidus. Le manque de données et de variables discriminantes nous empêchent de construire ce modèle de sévérité. Ce constat n'est pas surprenant et coïncide le modèle utilisé pour le tarif MRH.

Par conséquent, nous nous contenterons d'utiliser le coût moyen des sinistres clos pour la partie sévérité. Nous avons choisi de retenir le montant moyen des sinistres clos sur

la période 2011-2015 car les taux de clôture chutent fortement après 2015 comme nous l'a montré la figure 3.13. En raison de la faible volumétrie, le coût moyen appliqué est uniforme sur l'ensemble du territoire et ne sera pas divulgué par mesure de confidentialité.

Dès lors, il suffira de multiplier le nombre de sinistres prédits par l'une des approches précédentes avec ce coût moyen pour obtenir la charge. Enfin, nous agrégeons ces montants par année pour obtenir l'ultime :

4.5 Estimation à l'ultime : comparaison des approches

4.5.1 Comparaison des fréquences prédites

Le graphique 4.44 compare les résultats des deux approches sur l'échantillon de validation. Sur la période 2016-2018, les prévisions sont plus proches de la réalité lorsque nous utilisons le modèle de détection et sa classification. Sur les années antérieures, ce modèle sur-prédit la fréquence, par principe de prudence nous préférerons un modèle sur-estimant la fréquence plutôt que l'inverse.

Les deux autres approches sous-estiment la fréquence sur la période 2016-2018. Nous pouvons également comparer les fréquences moyennes sur l'ensemble de la base de validation, la figure 4.61 en annexe confirme que la seconde approche utilisant la classification est meilleure en moyenne.

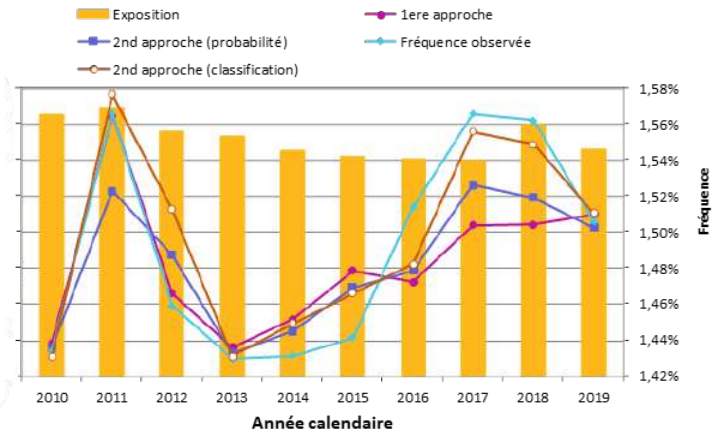


FIGURE 4.44 – Fréquence moyenne sur la base holdout

Le graphique 4.45 permet de comparer les prévisions de nos approches sur chacun des individus de notre base de validation. L'axe des abscisses représente la variation des prévisions (en pourcentage) entre la seconde approche et la première approche $\left(\frac{2^{e}\text{ approche}}{1^{re}approche} - 1\right)$. Ainsi nous pouvons déterminer, lorsque les prévisions des deux approches divergent, laquelle est la plus proche de la réalité.

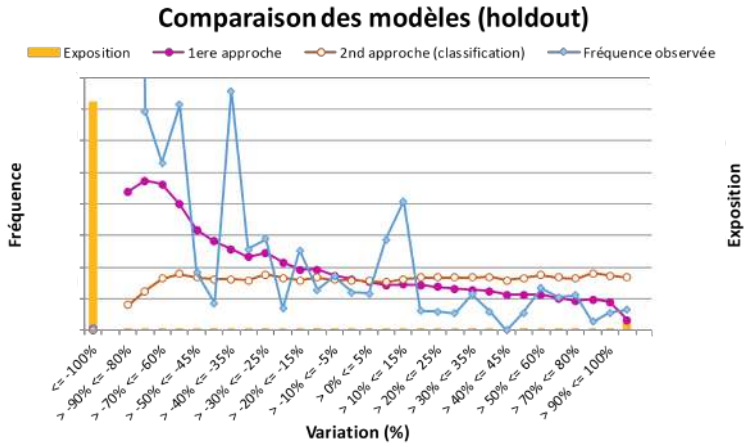


FIGURE 4.45 – Comparaison des modèles sur la base de validation

Ici, la première approche est plus proche de la réalité et nous pouvons le confirmer avec le tableau 4.7. Malgré cela la première approche sous-estime trop le nombre de sinistres pour les années les plus récentes, c'est pourquoi nous préférerons la seconde approche (avec la classification) à la première.

RMSE	$1^{re}approche$	$2^{e}approche$
Apprentissage	0.06901	0.06906
Validation	0.05989	0.05994

TABLE 4.7 – RMSE sur la base d'apprentissage et de validation

Le graphique 4.46 compare la fréquence moyenne sur l'ensemble des données avec les deux méthodes. Ceci nous permet d'apprécier les prédictions sur les années 2020 et 2021. Il est difficile de juger de la qualité des modèles sur ces années car une partie des sinistres ouverts pourront être classés sans-suite en l'absence d'arrêté, de la même façon d'autres sinistres pourront être déclarés ultérieurement pour ces années de survenance.

Pour l'année 2020, la fréquence prédite par la seconde approche utilisant la classification est plus proche de la fréquence observée donc le modèle estime un nombre de passage à l'état sans-suite plus faible que la première approche ou un nombre plus élevé de sinistres tardifs. Comparé aux récentes années, l'année 2021 a été beaucoup moins touché par la sécheresse : la première méthode coïncide avec la fréquence observée tandis que la seconde approche prédit un nombre de sinistres moins important.

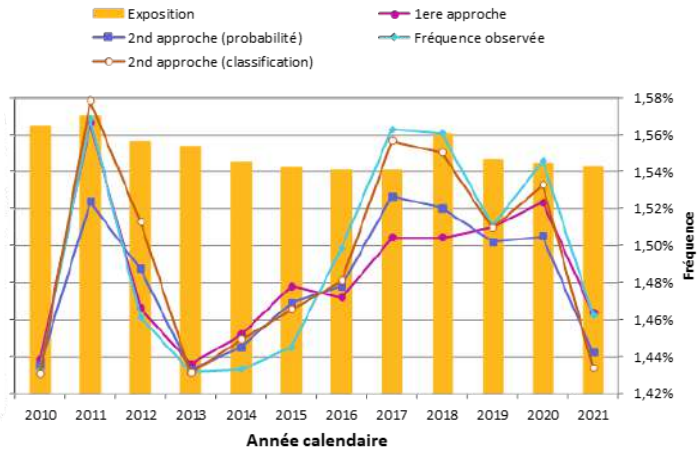


FIGURE 4.46 – Fréquence moyenne sur l'ensemble de la base

4.5.2 Comparaison de la charge totale prédite

Après avoir multiplié les prédictions de ces méthodes avec le coût moyen des sinistres clos sur la période 2011-2015, nous pouvons comparer la charge à date avec les estimations de notre modèle.

Pour les années avant 2015, la charge totale est comparable avec les estimations de nos modèles. En revanche pour les années après 2015, il existe une incertitude sur la charge finale des dossiers ouverts. La charge de ces sinistres peut être amenée à s'accroître si les dégâts s'aggravent comme décroître si les gestionnaires ont sur-estimés le montant restant dû.

Le graphique 4.47 montre les prédictions de la charge ultime par nos modèles sur la base de validation. La seconde approche utilisant la classification reste meilleure et plus prudente sur les années récentes. L'utilisation du coût moyen montre ses limites car l'écart entre les prédictions et l'observé s'est accentué par rapport à la figure 4.37 sur les années 2016 et 2017.

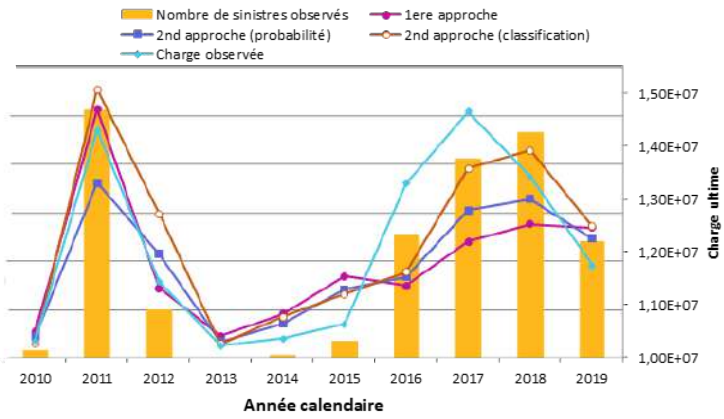


FIGURE 4.47 – Charge totale sur la base de validation

Comme nous le montre la figure 4.48, le coût moyen utilisé est plus faible que le coût moyen des sinistres ouverts et clos pour les années 2016 et 2017. A partir de l'année 2018, les sinistres clos rapidement correspondent aux dégâts mineurs tandis que la charge des sinistres ouverts est relativement faible par rapport aux années précédentes.

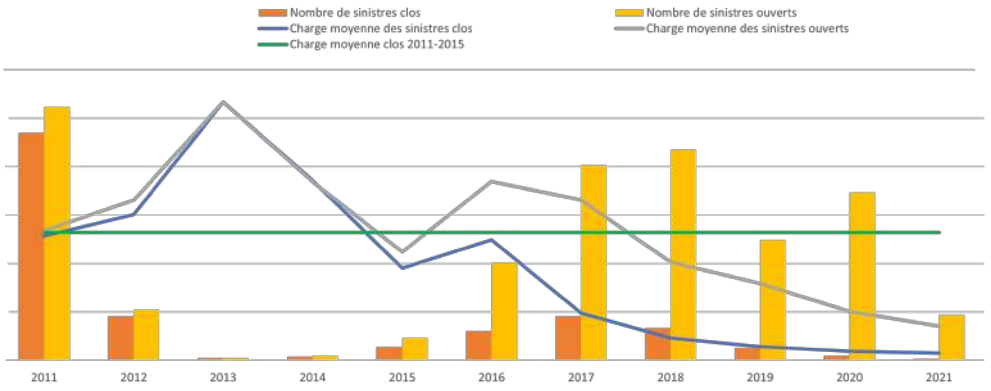


FIGURE 4.48 – Evolution de la charge moyenne par état de dossier

Par conséquent, les prévisions à l'ultime du modèle anticipent la survenance de nouveaux sinistres, le passage de certains dossiers ouverts en sans-suite ainsi que les réajustements de la charge sur les dossier en-cours.

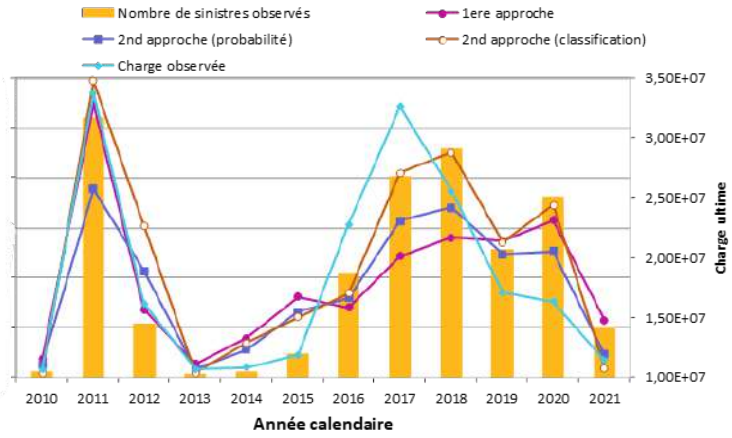


FIGURE 4.49 – Charge totale sur l'ensemble de la base

4.6 Conclusion de fin de chapitre

Au cours de la première méthode, nous avons pu remarquer que les indices de sécheresse étaient très discriminants sur la fréquence. En particulier, la reproduction des critères d'éligibilité sur la couche superficielle du sol ainsi que l'historique des demandes de reconnaissance antérieures ont un fort impact sur les métriques. Les résultats de ce modèle sont satisfaisants d'un point de vue global même si le modèle sous-prédit la fréquence sur les années 2016 à 2019.

En ce qui concerne la seconde méthodologie, nous avons pu voir que le modèle de détection était plus à même de se généraliser sur les années futures. En effet, l'année test de 2020 offre les meilleurs performances vis-à-vis de nos métriques. Le seuil de classification choisi n'est pas toujours le plus optimal ce qui a conduit à une sur-estimation du nombre d'arrêté en 2011 et une sous-estimation en 2017.

Pour le modèle de fréquence post-détection, la filtration sur les communes et périodes reconnues a eu pour effet de neutraliser le pouvoir prédictif de la plupart des indices de sécheresse. De plus, la population étudiée n'étant plus la même, les deux modèles de fréquence ne sont plus directement comparables.

La combinaison du modèle de détection avec le second GLM a un effet favorable sur les prédictions lorsque nous utilisons la classification. Malgré la sur-estimation du modèle de détection pour l'année 2011 et la sous-estimation du modèle en 2017, la fréquence prédite est proche de la fréquence observée.

Lorsque nous comparons simultanément les prévisions de nos modèles, la première approche est meilleure sur les années avant 2015 inclus puis la 2^{e} approche incluant la classification se démarque puis qu'elle ajuste mieux la tendance sur l'année 2016 et 2017.

Pour les années suivantes les modèles ne sont pas comparables car le modèle de détection a appris sur les années 2018 et 2019, pour l'année 2020 et 2021 c'est la fréquence et la charge observée qui seront amenés à évoluer.

Finalement, l'utilisation d'un coût moyen a montré ses limites puisque le passage de la fréquence à l'ultime a augmenté l'écart entre les courbes observés et prédites en 2016 et 2017.

Conclusion

L'objectif de ce mémoire consistait à définir une méthodologie robuste pour l'estimation de la charge ultime liée au RGA.

Le premier chapitre de ce mémoire a abordé la sécheresse dans sa globalité. Nous avons pu y définir ses origines et conséquences ce qui nous a permis de retenir des indices de sécheresses adaptés pour quantifier l'intensité du phénomène ainsi que des indices d'exposition ou de vulnérabilité à l'échelle communale. Enfin, nous avons pu définir le cadre réglementaire ainsi que l'évolution des critères de reconnaissance, ceci nous a permis de définir la méthodologie de calcul des critères actuellement en vigueur en vue de les reproduire.

Comme l'indice d'humidité des sols utilisé par la commission interministérielle (SWI uniforme) n'est mis à jour que tardivement, nous avons fait le choix de reproduire les critères de reconnaissance à partir des données provenant d'un centre météorologique européen. La reproduction des critères météorologiques s'est avérée très proche de l'éligibilité originale.

Lors de la construction de nos bases de modélisation, nous avons remarqué qu'une partie non négligeable de nos sinistres était à priori survenu lors de périodes non reconnues en l'état de catastrophe naturelle sécheresse. Pour ne pas associer un indice de sécheresse humide à ces sinistres, nous avons choisi de raccrocher arbitrairement nos sinistres à la période reconnue la plus proche dans la limité du délai de déclaration de 2 ans. Cette hypothèse nous a permis d'éviter un biais lors de la modélisation et de conserver le plus de sinistres possible.

Le premier modèle de fréquence que nous avons construit repose sur un modèle linéaire généralisé de type Poisson. Cette première approche ne tient pas compte de la sur-proportion de sinistres nuls liés à l'absence d'apparition de l'évènement et nous a servi de référence pour quantifier l'apport de la seconde méthodologie.

Au sein de ce modèle, nous avons pu voir que la fréquence augmente très fortement lorsque la période de retour de l'humidité des sols superficiels excède 50 ans. La connaissance de ce critère associée à l'historique des demandes de reconnaissance passées améliore très significativement notre modèle. Si les prévisions de ce modèle sont quasiment parfaites en moyenne pour les premières années, l'écart avec les valeurs observées se creuse après 2016 et nous sous-estimons la fréquence. Malheureusement, nous n'avons pas eu le temps de regarder la cohérence spatiale entre les sinistres prédits et observés mais cette étude sera réalisée à l'avenir.

La seconde méthodologie propose de modéliser la fréquence en deux sous-parties distinctes. Le premier modèle a comme objectif de détecter les communes qui vont être déclarées Cat Nat tandis que le second prédit le nombre de sinistres pour les communes

identifiées.

Comme il existe peu de saisons et communes reconnues Cat Nat, la variable cible du modèle de détection est fortement déséquilibrée. Nous avons alors définis plusieurs métriques sensibles à ce déséquilibre et nous avons étudié l'impact de plusieurs techniques de rééchantillonage sur nos performances. Malgré cet effort, les résultats de notre modèle ne se sont pas améliorés. Nous avons donc poursuivi l'apprentissage sur les années 2018 et 2019 afin que notre modèle se généralise au mieux sur les années futures pour lesquels les critères de reconnaissance sont identiques. Les résultats de la détection sont satisfaisant pour les prédictions futures sur nos métriques mais aussi d'un point de vue spatial. Cependant, le seuil de classification n'est pas toujours le plus adapté lors de la rétrodiction. De ce fait certaines années sont sur-estimées ou sous-estimées.

Par la suite, nous avons construit un second modèle généralisé de type Poisson mais cette fois-ci en filtrant nos données sur les communes et périodes reconnues. Ce filtre a eu pour effet de placer nos données dans des conditions météorologiques déjà défavorables. Par conséquent, un unique indice de sévérité a été sélectionné : les précipitations nettes standardisées.

Comme les deux approches ne sont pas comparables entre elles avec les métriques présentées, nous avons comparé les fréquences annuelles moyennes ainsi que la fréquence moyenne sur l'échantillon de validation avec l'observé. Nous nous apercevons que la seconde méthode est plus proche de la fréquence observée. En particulier, la classification a un réel impact sur les prédictions.

Enfin, le manque de données et de variables discriminantes nous empêche de construire un modèle sévérité. Afin d'obtenir la charge ultime sur le péril, nous avons dû nous contenter d'appliquer le coût moyen des sinistres clos sur la période 2011-2015. L'utilisation de ce coût moyen accentue les écarts sur l'estimation finale. Au global, les prévisions sont cohérentes et satisfaisantes.

A l'avenir, le processus sera automatisé en interne de sorte à avoir un suivi trimestriel du risque pour GENERALI. Il va permettre de conforter la décision sur les montants à provisionner pour les IBNyR et s'avère porteur d'autres projets. En effet, les indices de sécheresse utilisés ainsi que la reproduction des critères météorologiques pourront faire l'objet d'études pour la mise à jour du générateur et du zonier stochastique.

Afin d'améliorer les prévisions de notre modèle, nous pourrions intégrer plusieurs données issues de la télématique comme le niveau de pentification sur la parcelle, le nombre et la distance des arbres par rapport aux bâtiments ou encore la forme géométrique des maisons.

Il serait également intéressant de comparer les résultats de cette approche avec les modèles à zéros-inflation (ZIP ou ZINB). Ces derniers intègrent simultanément un modèle logistique et un modèle de comptage mais ne sont pas implémentables sur le logiciel de

tarification ce qui rend la modélisation plus chronophage.

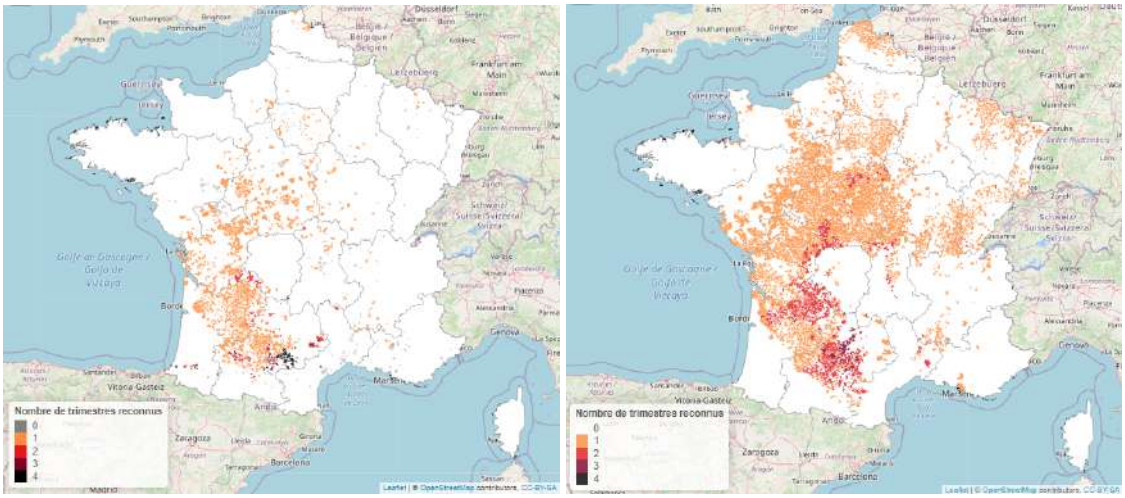
Annexe

A - Sélection de variables de la 1re méthode

Étape	Variables	Pré simplification			Post simplification		
		AIC	BIC	Déviance	AIC	BIC	Déviance
Étape 1	Critère 28cm	-3 107	- 2390	-3147	-3099	-3056	-3103
Étape 2	Nombre de demandes	-1882	-1509	-1903	-1888	-1845	-1893
Étape 3	Part de la surface en aléa fort	-552	+895	+6	-557	-542	-559
Étape 4	ESTI	-436	+295	-445	- 364	-350	-366
Étape 5	Année calendaire	-343	+832	-425	-247	-190	-256
Étape 6	$ESSWI\_1m$	-198	+848	-251	-160	-132	-164
Étape 7	Concentration moyenne d'argile	-220	-1271	-286	-42	-27	-44
Étape 8	Zonier exposition sécheresse	-166	-65	-165	-140	-111	-144
Étape 9	Trimestre	-158	-115	-157	-218	-175	-223
Étape 10	ESPEI	-198	+518	-222	-131	-153	-137
Étape 11	Risque 5	-198	-100	-57	-102	-73	-106
Étape 12	Risque 6	-82	-25	-82	-77	-49	-82
Étape 13	Risque 7	-198	+518	-222	-74	-45	-77
Étape 14	Risque 2	-2	+413	-11	-38	-24	-40

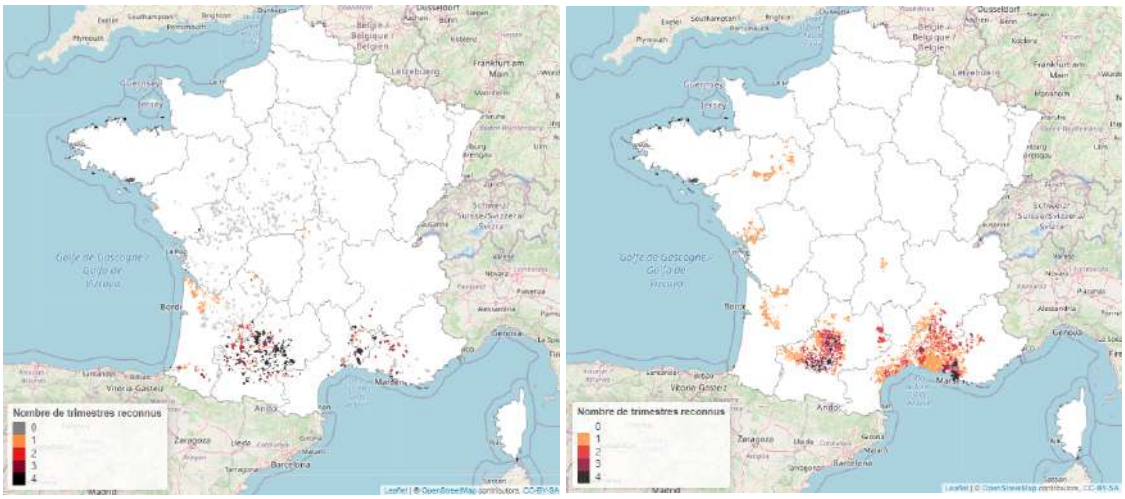
Table 4.8 – Sélection forward - méthode n°1

B - Comparaison entre les arrêtés Cat Nat observés et les prévisions du modèle de détection



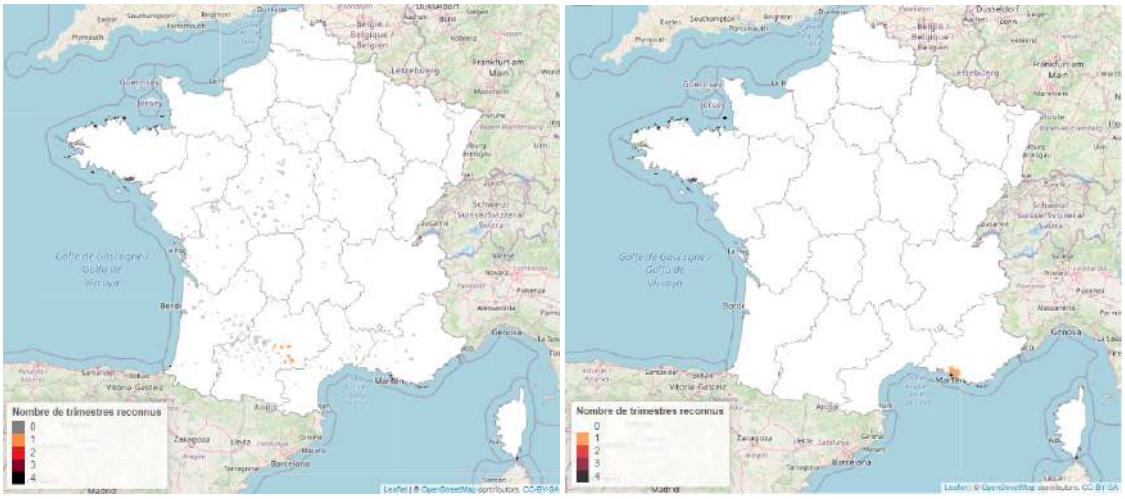
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.50 – Comparaison entre observés et prédits sur 2011



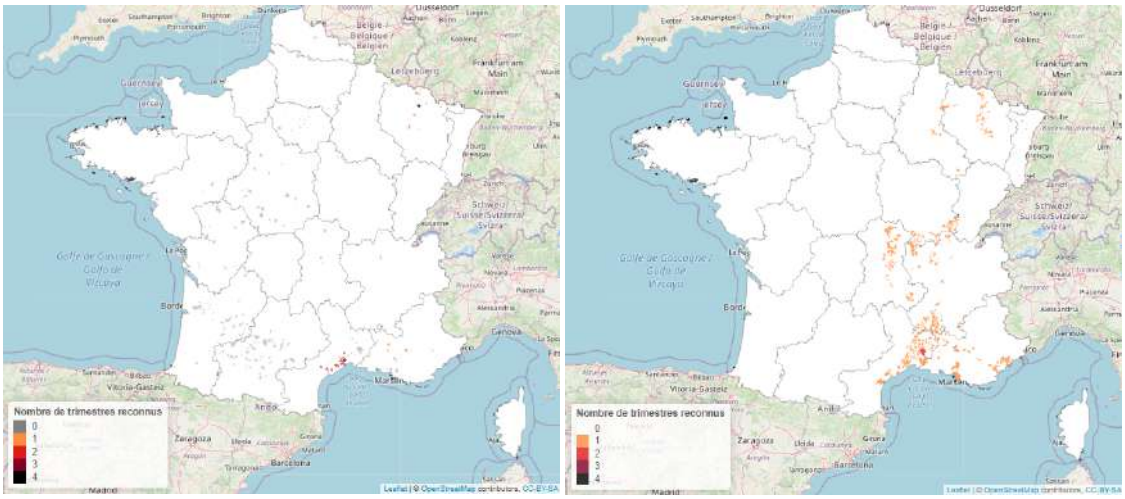
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.51 – Comparaison entre observés et prédits sur 2012



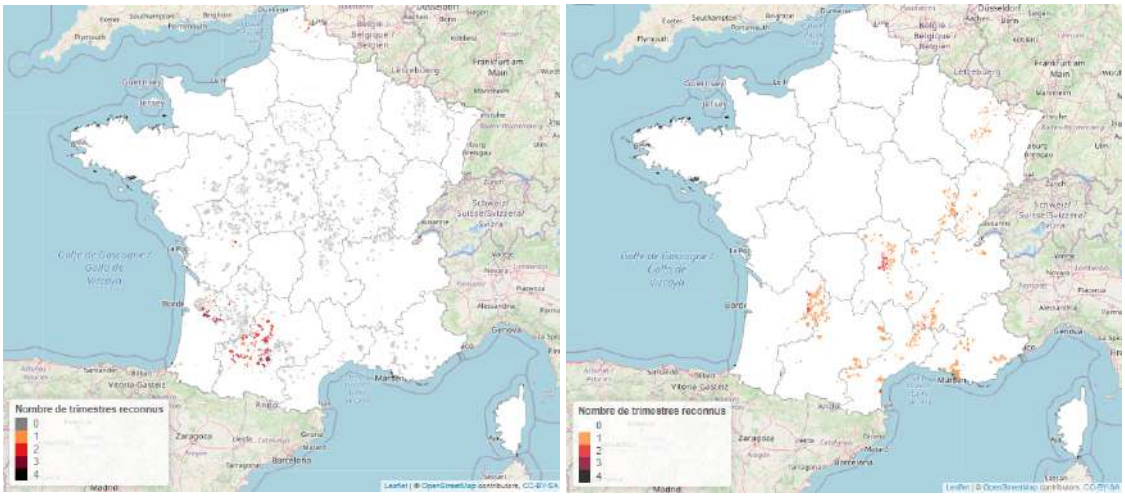
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.52 – Comparaison entre observés et prédits sur 2013



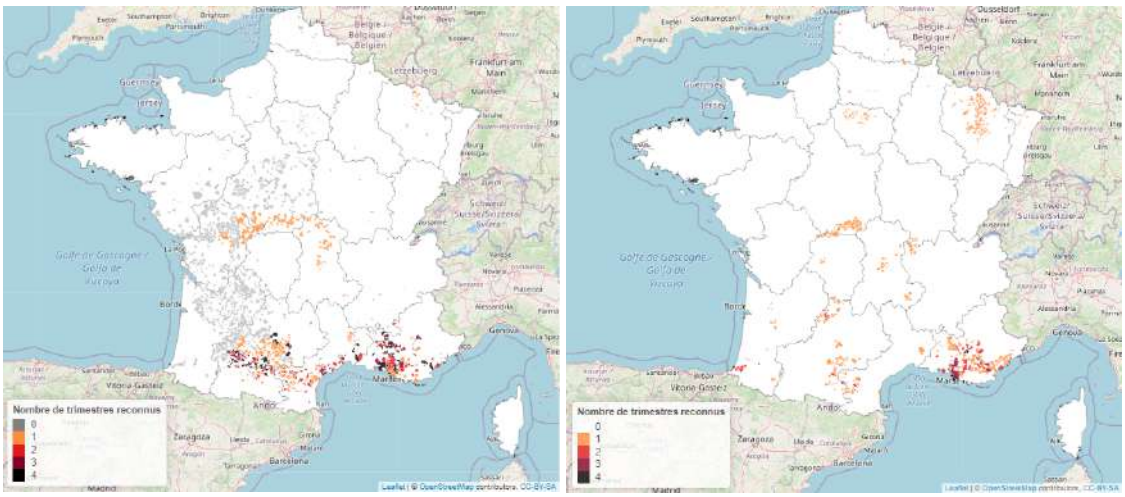
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.53 – Comparaison entre observés et prédits sur 2014



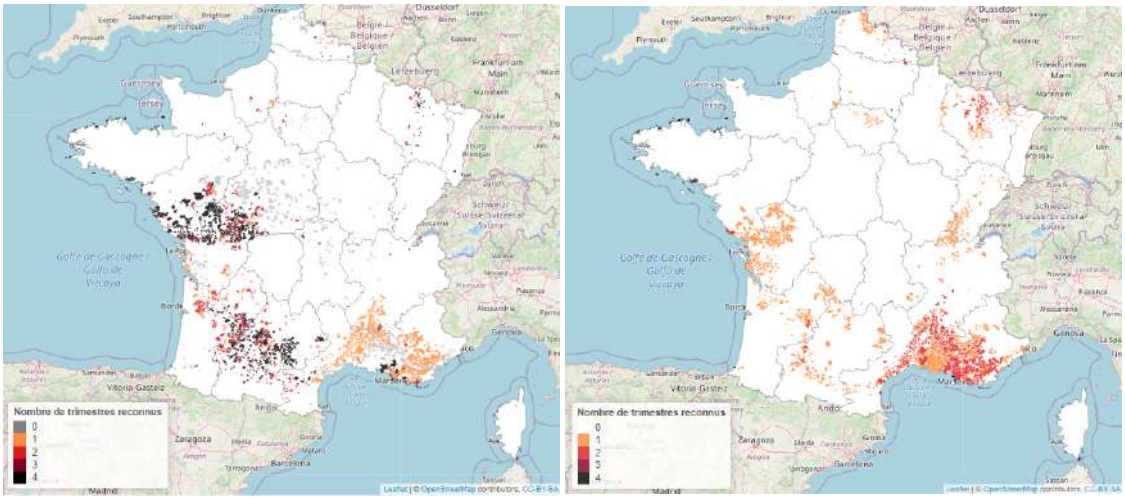
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.54 – Comparaison entre observés et prédits sur 2015



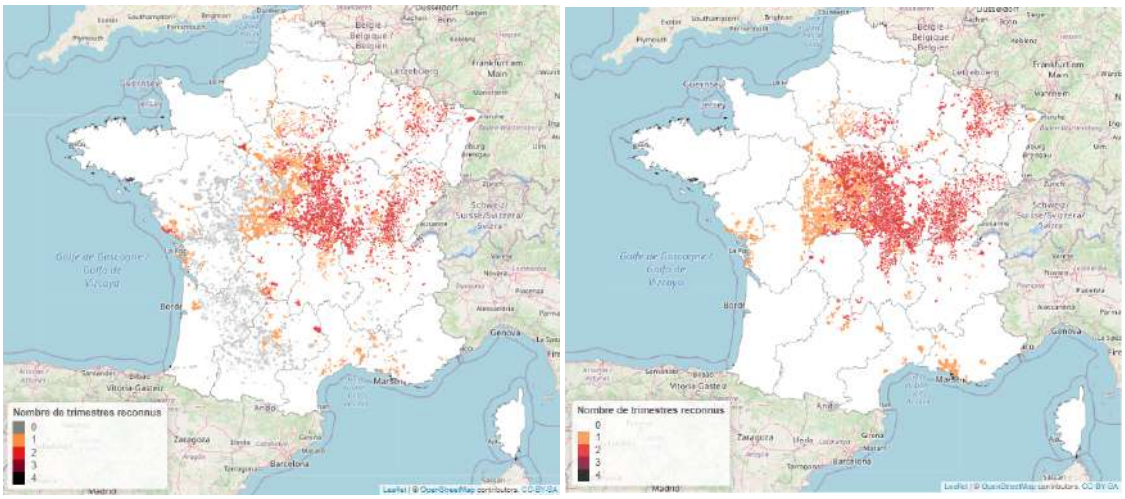
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.55 – Comparaison entre observés et prédits sur 2016



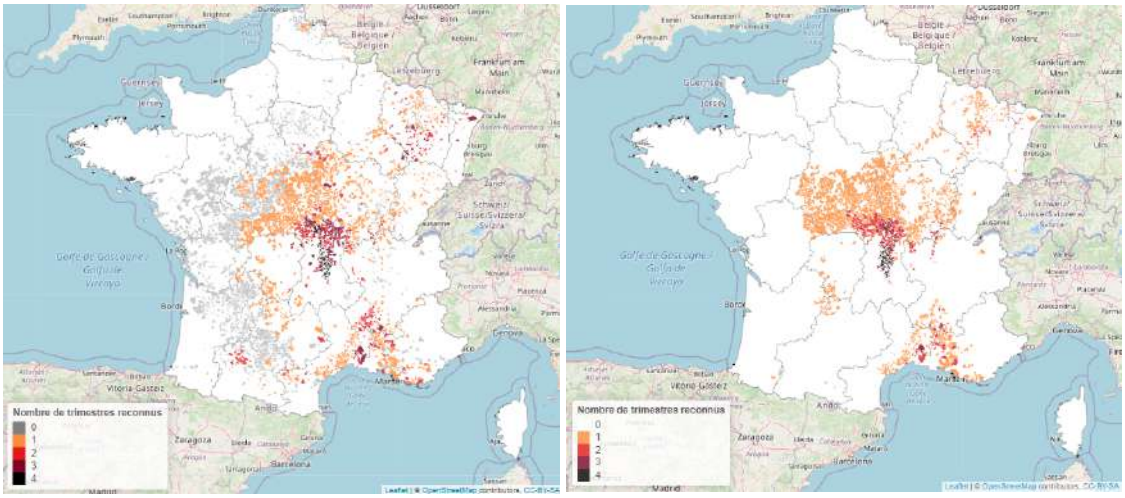
(a) Arrêtés observés (b) Rétrodiction

FIGURE 4.56 – Comparaison entre observés et prédits sur 2017



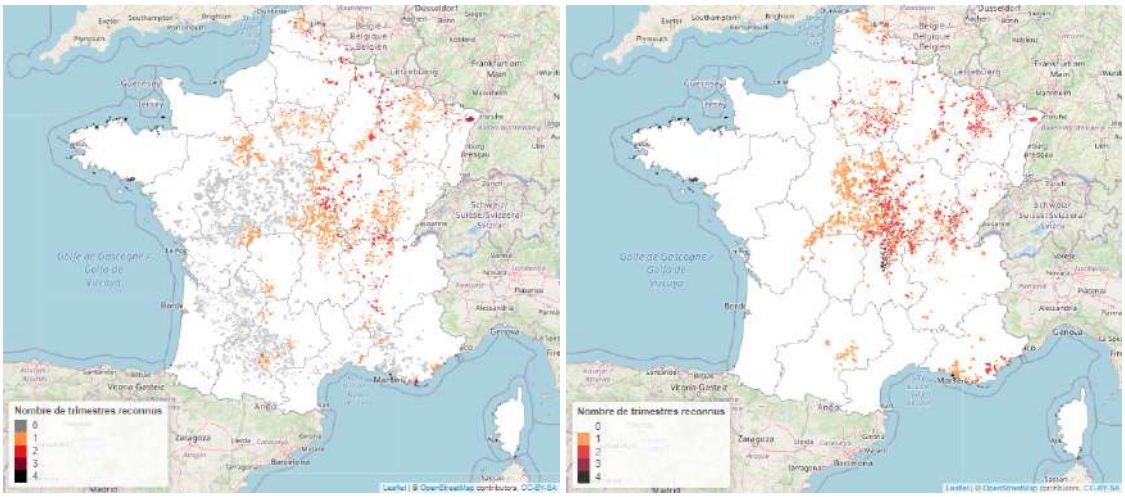
(a) Arrêtés observés (b) Apprentissage

FIGURE 4.57 – Comparaison entre observés et prédits sur 2018 (apprentissage)



(a) Arrêtés observés (b) Prévisions (apprentissage)

FIGURE 4.58 – Comparaison entre observés et prédits sur 2019



(a) Arrêtés observés (b) Prévisions

FIGURE 4.59 – Comparaison entre observés et prédits sur 2020

FIGURE 4.60 – Prévisions - 2021

C - Sélection de variables pour le modèle de détection

1. Nombre de jours depuis la dernière publication au journal officiel
2. Latitude
3. Nombre de demandes de reconnaissance antérieures
4. Longitude
5. Eswvl1 (SWI non uniforme standardisé sur la couche 0-28cm)
6. ESPI3
7. ESTI3
8. ESPEI3
9. $ESSWI_{28cm}$
10. Part de la surface communale en aléa moyen ou fort
11. Eswvl2 (SWI non uniforme standardisé sur la couche 8-24cm)
12. Part de la surface communale en aléa fort
13. Rang du SWI non uniforme sur la couche 7-28cm)
14. ESRO3 (Ruissellement standardisés)
15. Concentration moyenne d'argile
16. Nombre de maisons exposées à l'aléa fort
17. $ESSWI_{2m}$
18. Surface communale en aléa moyen ou fort
19. Rang du SWI non uniforme sur la couche (0-28cm)
20. Eswvl4 (SWI non uniforme standardisé sur la couche 1m-2m89)
21. Nombre de maisons exposées à l'aléa moyen
22. Nombre de maisons construites avant 1920 exposées à l'aléa moyen ou fort
23. $ESSWI_{1m}$
24. Nombre de maisons construites après 1975 exposées à l'aléa moyen ou fort
25. Part de la surface communale en aléa moyen
26. Eswvl3 (SWI non uniforme standardisé sur la couche 28cm-1m)
27. EscPDSI (indice de Palmer)
28. EscPHDI (indice de Palmer)
29. EscWPLM (indice de Palmer)

D - Comparaison de la fréquence et de la charge totale

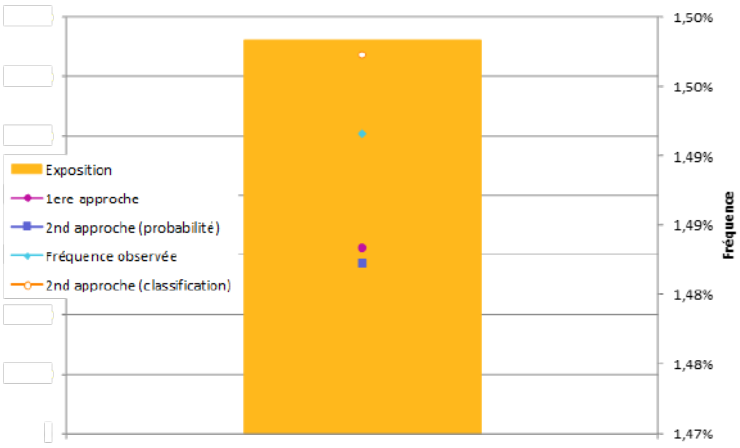


FIGURE 4.61 – Fréquence moyenne sur la base holdout

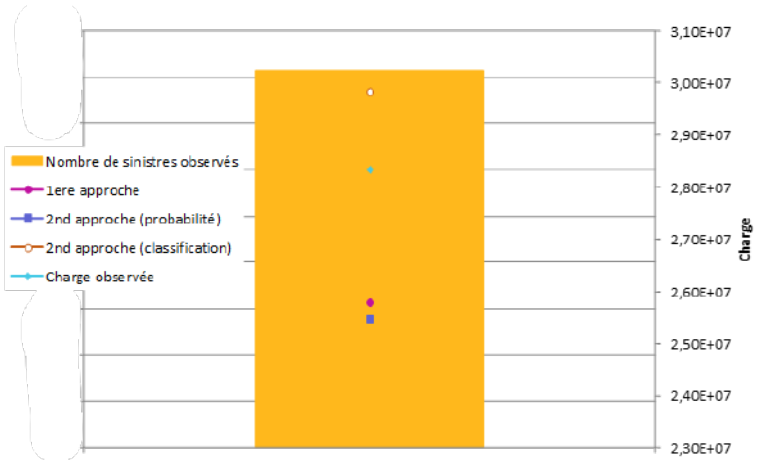


FIGURE 4.62 – Charge totale sur la base holdout

Table des figures

Liste des tableaux

Bibliographie

[CC, 2022] (2022). *Des dommages en forte progression, un régime de prévention et d'indemnisation inadapté*. Cour des comptes.

[ARNAUD, 2016] ARNAUD, E. (2016). Modélisation du risque sécheresse en france. Mémoire de D.E.A., Dauphine.

[CCR, 2021] CCR (2021). Rapport scientifique. https://www.ccr.fr/-/ccr-rapport-scientifique-2020.

[CCR, 2022] CCR (2022). L'indemnisation des catastrophes naturelles en france. https://www.ccr.fr/documents/35794/35836/indemnisation+cat-nat.pdf/ff905a8f-ccb3-44e2-a0d0-b92c6d2e352e?t=1452598764000.

[CHARPENTIER, 2013] CHARPENTIER, A. (2013). Modèle linéaires généralisés. http://freakonometrics.free.fr/slides-2040-4.pdf.

[CHARPENTIER *et al.*, 2021] CHARPENTIER, A., JAMES, M. et ALI, H. (2021). Predicting drought and subsidence risks in france. https://nhess.copernicus.org/articles/22/2401/2022/.

[CLIMSEC, 2011] CLIMSEC (2011). Apports opérationnels pour le monitoring des sécheresses. https://www.cnrm.meteo.fr/IMG/pdf/20110630_climsec_p2_mblanchard.pdf.

[DUTANG, 2017] DUTANG, C. (2017). Some explanations about the iwls algorithm to fit generalized linear models. https://hal.archives-ouvertes.fr/hal-01577698/document.

[ECOTO *et al.*, 2021] ECOTO, G., BIBAUT, A. et CHAMBAZ, A. (2021). One-step ahead sequential super learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. https://arxiv.org/abs/2107.13291.

[FRÉCON et KELLER, 2009] FRÉCON, J.-C. et KELLER, F. (2009). Sécheresse de 2003 : un passé qui ne passe pas. https://www.senat.fr/rap/r09-039/r09-039.html.

[JOETZJER, 2014] JOETZJER, E. (2014). Variabilité inter-annuelle des sécheresses et leur réponse au changement climatique : Quels indicateurs ? Mémoire de D.E.A., ENSAIA.

[LUNARDON *et al.*, 2014] LUNARDON, N., MENARDI, G. et TORELLI, N. (2014). Rose : A package for binary imbalanced learning. https://journal.r-project.org/archive/2014/RJ-2014-008/RJ-2014-008.pdf.

[Légifrance, 2019] Légifrance (2019). Révision des critères de reconnaissance : circulaire n° inte1911312c. https://www.legifrance.gouv.fr/circulaire/id/44648.

[Légifrance, 2021] Légifrance (2021). Loi baudu (n° 2021-1837) du 28 décembre 2021 relative à l'indemnisation des catastrophes naturelles. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044589864.

[MRN, 2018] MRN (2018). Sécheresse géotechnique : De la connaissance de l'aléa à l'analyse de l'endommagement du bâti. https://www.mrn.asso.fr/rapport-mrn-secheresse-geotechnique/.

[OMM, 2012] OMM (2012). Organisation météorologique mondiale - guide d'utilisation de l'indice de précipitations normalisé. https://www.droughtmanagement.info/literature/WMO_standardized_precipitation_index_user_guide_fr_2012.pdf.

[Préfecture, 2018] Préfecture (2018). Procédure générale de demande de reconnaissance de l'état de catastrophe naturelle.

[RAKOTOMALALA, 2012] RAKOTOMALALA, R. (2012). Zero inflated poisson regression (zip). https://eric.univ-lyon2.fr/ricco/cours/slides/zip_regression.pdf.

[SCHULTE, 2016] SCHULTE, J.-F. (2016). Modélisation du risque subsidence en france métropolitaine. Mémoire de D.E.A., ISUP.

[SDES, 2021] SDES (2021). Cartographie de l'exposition des maisons individuelles au retrait-gonflement des argiles. https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2021-06/note_methode_croisement_retrait_gonflement_argiles_juin2021v3.pdf.

[SOUBEYROUX *et al.*, 2012] SOUBEYROUX, J.-M., KITOVA, N., BLANCHARD, M., VIDAL, J.-P., MARTIN, E. et DANDIN, P. (2012). Caractérisation des sécheresses des sols en france et changement climatique : Résultats et applications du projet climsec. https://hal.archives-ouvertes.fr/hal-00757327.

[WELLS *et al.*, 2004] WELLS, N., GODDARD, S. et J.Hayes, M. (2004). A self-calibrating palmer drought severity index. https://journals.ametsoc.org/view/journals/clim/17/12/1520-0442_2004_017_2335_aspdsi_2.0.co_2.xml.