**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 8 Septembre 2022

Par : Kouadio Jean-Emmanuel
Titre : Assurance Multirisque Agricole : Calibration d'un modèle technique et apport de l'*Open Data*

Confidentialité : Oui - (Durée: 2 ans)

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

***Membre présent du jury de l'Institut des Actuaires :***
Romain LAILY
Julie SURGET
Signature :

***Entreprise :***
AXA France
Signature :

***Membres présents du jury de l'EURIA :***
Françoise PENE

***Directeur de mémoire en entreprise :***
Oksana ALLAIRE
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**

Un contrat Multirisque Agricole (MRA) a pour but d'assurer l'ensemble des biens et responsabilités professionnels et privés des exploitants en activité ou non. Ce contrat est destiné aux propriétaires et locataires dont l'activité ou les biens sont à vocation exclusive agricole avec des exploitations enregistrées en nom propre ou sous forme sociétaire. Un contrat MRA est un contrat axé sur les bâtiments d'exploitations et d'habitations de l'agriculteur et non sur la récolte en terre comme en assurance récolte. Néanmoins, pour les produits d'origine végétale non récoltés, l'assureur s'engage à dédommager l'assuré en cas de sinistre d'un montant équivalent à une année de production. Par ailleurs, un contrat MRA vise aussi à couvrir le contenu agricole récolté, le mobilier, la valeur des espèces animales (en cas de décès) et le matériel agricole de l'exploitation. Cependant, les véhicules et matériels automoteurs (ex : tracteur) soumis à l'obligation d'assurance sont exclus mais couverts par le contrat Matagri d'AXA France.
L'objectif de ce mémoire est de mettre en place une modélisation de la prime pure multirisque agricole par garantie. L'étude réalisée essayera de mesurer l'apport de données *Open Data* dans les modèles de prime pure. Dans le cadre de ce mémoire, deux des garanties les plus importantes du produit multirisques agricoles seront traitées en occurrence : la garantie incendie et la garantie tempête grêle neige (TGN).

**Mots clefs:** Multirisque Agricole, Tarification, Segmentation, Prime pure, GLM, XGBoost, Zonier, Approche Bayésienne

**Abstract**

A Multirisk Agricultural Policy (MRA) is designed to insure all the professional and private property and liabilities of farmers, whether or not they are active. This contract is intended for owners and tenants whose activity or property is exclusively agricultural, with farms registered in their own name or as a company. A MRA policy focuses on the farm buildings and dweillings, instead of focusing on the crop on land as in crop insurance. Nevertheless, for unharvested plant products, the insurer undertakes to compensate the insured in the event of a claim for an amount equivalent to one year's production. In addition, a MRA policy also aims to cover harvested agricultural contents, furniture, the value of animal species (in case of death) and the farm's agricultural equipment. However, vehicles and self-propelled equipment (e.g. tractors) subject to compulsory insurance are excluded, but covered by AXA France's Matagri contract.
The objective of this thesis is to set up a modelling of the pure agricultural multi-risk premium by guarantee. The study will try to measure the contribution of the Open Data in the pure premium models. Within the framework of this thesis, two of the most important coverages of the agricultural multi-risk product will be treated in this case : the fire coverage and the storm-hail-snow coverage (SHN).

**Keywords:** Agricultural Multirisk, Pricing, Segmentation, Pure Premium, GLM, XGBoost, Zoning, Bayesian Approach

# Note de synthèse

Dans une politique d'amélioration de l'offre agricole d'AXA France, une revue de la mécanique de tarification du produit multirisque agricole (MRA) a été suggérée. L'une des particularités du contrat multirisque agricole est sa capacité à couvrir deux risques différents : l'habitation et l'exploitation de l'agriculteur. Ce contrat propose des garanties à la fois communes et spécifiques à l'habitation et l'exploitation. Ce mémoire a pour objectif de mettre en place un modèle de prime pure en vue d'améliorer la tarification du produit MRA.
La prime pure, ou prime technique, correspond au montant attendu des sinistres d'un assuré sur une période. Actuellement, il n'existe pas de modèle de prime pure pour le produit MRA, sa tarification est basée sur l'avis d'experts, en fonction de la sinistralité et des tarifs proposés par les assureurs concurrents. La création d'un modèle de prime pure s'inscrit dans la politique d'ajustement tarifaire du produit MRA. En effet, depuis 2017 des mesures tarifaires fortes ont été mises en place afin d'améliorer le tarif proposé. Ces mesures visaient une réduction du tarif affaire nouvelle et portefeuille, jugé trop cher par certains assurés. Cependant, l'usage de toutes ces mesures n'a pas permis d'obtenir un apport net positif. En outre, le zonier actuel ne réussit plus à segmenter le risque correctement. Cette difficulté de segmentation serait liée à sa structure et son ancienneté (début des années 2000).

Par conséquent, cette étude vise à répondre aux enjeux suivants :
- **Comprendre au mieux le produit multirisque agricole d'AXA France** ;
- **Mettre en place un modèle de prime pure pour le produit MRA d'AXA France** ;
- **Mettre en place un nouveau zonier**.

L'une des premières étapes de l'étude consiste à comprendre la mécanique de tarification actuelle. Cette étape est fondamentale pour créer une nouvelle structure de tarification, surtout lorsqu'il s'agit de modélisation. Un état des lieux de la méthode de tarification actuelle sert de proxy pour identifier les variables qui seraient plus ou moins pertinentes dans la modélisation. Par ailleurs, une amélioration de la qualité de segmentation d'un modèle construit avec uniquement des variables internes est envisagé au regard de la masse de données disponibles en *Open Data* et des garanties couvertes.

Le contexte et les enjeux du mémoire étant présentés, deux des garanties les plus importantes du produit multirisque agricole sont traitées : la garantie incendie et la garantie tempête grêle neige. Une base de modélisation par images de contrats de 2010 à 2021 est construite et enrichie à partir de données en *Open Data*. Les variables *Open Data* créées sont issues de bases construites dans les travaux de ce mémoire ou de bases de données disponibles directement sur internet. Ces dernières sont relatives à la distance aux centres incendie, la météo, la criminalité, l'année de construction de biens immobiliers et le prix au mètre carré.
Les statistiques descriptives univariées, l'analyse des corrélations et des distributions des variables *Open Data* révèlent les liens entre les différentes variables et la sinistralité observée. Au préalable, un traitement des données par interpolation et imputation (par la moyenne ou le mode) est réalisé avant toute analyse ou modélisation.

Pour éviter d'avoir une modélisation du coût des sinistres perturbée par des coûts extrêmes de sinistres , ceux-ci ont été écrêtés et mutualisés. L'écrêtement et la mutualisation du coût des sinistres extrêmes s'opère en utilisant un seuil déterminé par la théorie des valeurs extrêmes. En parallèle, la méthode de Chain-Ladder a servi à déterminer la charge finale des sinistres attritionnels et extrêmes non clôturés. Une étude réalisée sur la charge des sinistres a donné du sens à l'utilisation d'un unique seuil pour l'habitation et l'exploitation. De plus, une étude analogue a permis d'observer la différence entre les facteurs influant sur la fréquence de sinistres sur les risques habitation et exploitation. En revanche, l'analyse réalisée sur la modélisation du coût moyen révélait qu'une unique modélisation du coût moyen n'aurait pas d'effet néfaste sur la puissance globale du modèle et ne risquerait pas de conduire à une erreur tarifaire ou une antisélection majeure. Les variables retenues pour chacune des modélisations sont sélectionnées par analyse des corrélations et par pénalisation des coefficients.

La structure de modélisation adoptée en tenant compte des contraintes opérationnelles et du volume des données est la suivante :

$$\textit{Prime pure} = (\mathbb{1}_{Exploitation} * \textit{fréquence}_{Exploitation} + \mathbb{1}_{Habitation} * \textit{fréquence}_{Habitation}) * \textit{coût moyen global}$$

Les principaux résultats de la modélisation sont les suivants :

1. ***Modèles de fréquence***

   — *Habitation*
   La fréquence empirique par tranches de nombre de pièces et de contenu de l'habitation vient palier à l'absence de modèle stable sur la fréquence de sinistres habitation incendie. Le modèle sur la garantie TGN, quant à lui, est segmentant et stable sur la base d'apprentissage et la base de validation.

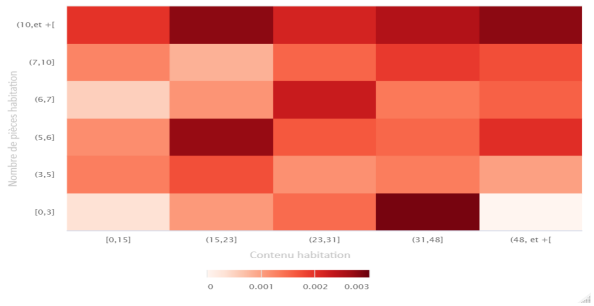



FIGURE 1 – Fréquence empirique des sinistres incendie habitation

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson TGN | 0,089 | 21,71 % | 0,089 | 21,23 % |

TABLE 1 – Résultats de la modélisation fréquence de sinistres TGN habitation

— *Exploitation*
Les modèles GLM calibrés sur le risque exploitation présentent des performances très satisfaisantes en termes de segmentation et de prédiction. La confrontation de ces modèles avec les modèles XGBoost démontrent que l'utilisation de modèles complexes ne permettraient pas une augmentation significative de la segmentation des modèles de fréquence de sinistres.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| XGBoost Poisson Incendie | 0,099 | 45,41% | 0.098 | 44,23% |
| GLM Poisson Incendie | 0,118 | 49,08% | 0,118 | 48,44% |
| XGBoost Poisson TGN | 0,145 | 37,23% | 0,145 | 36,92% |
| GLM Poisson TGN | 0,193 | 36,59 % | 0,193 | 36,22 % |

2. ***Modèles de coût moyen***
Les modèles de coût moyen semblent moins stables entre la base d'apprentissage et la base de validation selon le critère du Gini à cause du faible volume de données. Néanmoins, ces modèles décrivent assez correctement la tendance de la sinistralité observée avec une stabilité temporelle acceptable.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM gamma Incendie | 64150 | 13.22% | 64490 | 7.74% |
| GLM gamma TGN | 7185 | 9,27% | 7219 | 7,62% |

3. ***Intégration des données Open Data***
Dans les modèles de fréquence, l'intégration des données *Open Data* dans la modélisation permet une amélioration de la segmentation avec l'apparition de tendances discriminantes des risques. L'amélioration de la segmentation globale du modèle

paraît plus importante sur la garantie TGN. Par ailleurs, l'absence de tendance discriminante apparente lors de l'intégration des variables *Open Data* dans le modèle de coût moyen a conduit à l'abandon d'une modélisation du coût moyen avec ces variables.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* Incendie exploitation | 0,118 | 49,08% | 0,118 | 48,44% |
| GLM Poisson avec *Open Data* Incendie exploitation | 0,118 | 49,98% | 0,118 | 49,08% |
| GLM Poisson sans *Open Data* TGN habitation | 0,089 | 21,71 % | 0,089 | 21,23 % |
| GLM Poisson avec *Open Data* TGN habitation | 0,089 | 36,71 % | 0,089 | 33,64 % |
| GLM Poisson sans *Open Data* TGN exploitation | 0,193 | 36,59 % | 0,193 | 36,22 % |
| GLM Poisson avec *Open Data* TGN exploitation | 0,193 | 38,7% | 0,193 | 38,14% |

4. ***Ajout des zoniers***

L'intérêt de la construction d'un zonier est qu'il pourrait améliorer la capacité de prédiction et de segmentation du modèle.

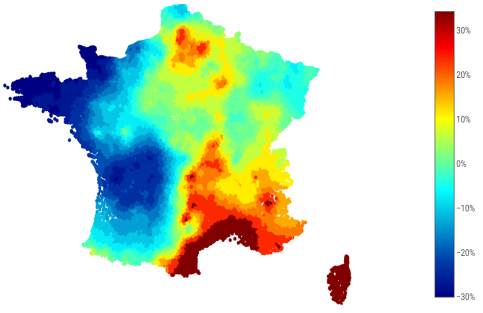
$$g(Y_i) - \hat{\beta} X_i = R_i$$
$$= \textit{facteur géographique} + \textit{aléa}$$

La construction des zoniers a été faite uniquement pour l'exploitation, en raison de la faiblesse du taux de couverture des communes pour le risque habitation.

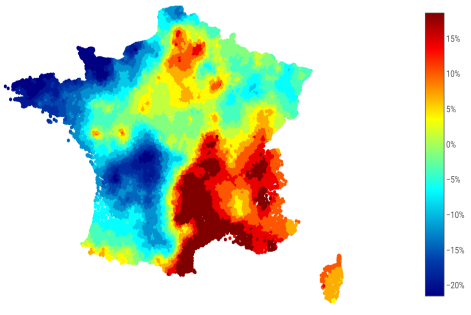
L'approche utilisée pour la construction des zoniers de fréquence de sinistres exploitation est une approche bayésienne avec une loi à priori gaussienne des facteurs géographiques. Les zoniers sont construits à partir des résidus de chaque modèle. Ces zoniers ainsi construits sont réinjectés dans leurs modèles respectifs comme de nouvelles variables. Dans ce mémoire, les zoniers construits sont des zoniers de fréquence de sinistres.

Du fait de la différence de la nature des risques modélisés au sein de chaque garantie, il est nécessaire de construire un zonier pour chaque garantie.

— *Zoniers incendie*

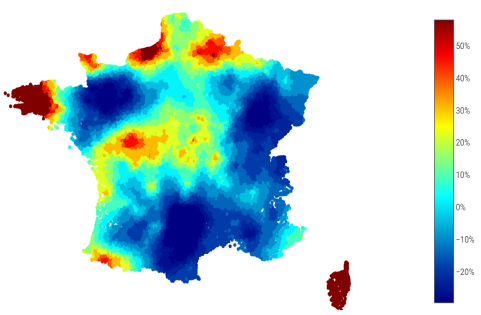



(a) Zonier issu du modèle sans *Open Data*

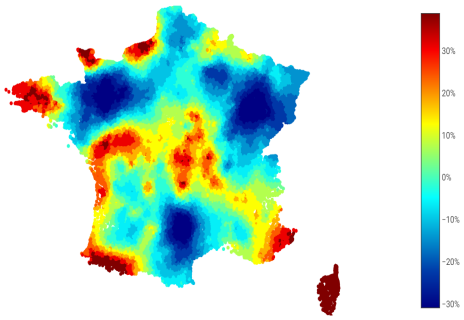



(b) Zonier issu du modèle avec *Open Data*

— *Zoniers TGN*




(c) Zonier issu du modèle sans *Open Data*




(d) Zonier issu du modèle avec *Open Data*

L'ajout des zoniers dans les modèles améliore la segmentation globale des différents modèles. De plus, l'interprétation de l'information géographique est facilitée par les tendances de sinistralité observées sur les données *Open Data*. Enfin, ces zoniers apportent une segmentation plus fine du risque en fonction de la garantie et corrigent le biais présent sur l'ancien zonier dû au fait qu'il soit calibré sur toutes les garanties.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* + zonier Incendie | 0,117 | 51,63% | 0,117 | 49,84% |
| GLM Poisson avec *Open Data* + zonier Incendie | 0,117 | 51,85% | 0,117 | 49,85% |
| GLM Poisson sans *Open Data* + zonier TGN | 0,193 | 42,24% | 0,193 | 40,49% |
| GLM Poisson avec *Open Data* + zonier TGN | 0,193 | 43,16% | 0,193 | 41,25% |

5. ***Prime pure***

   Les modèles de primes pures résultant des différents modèles de fréquence et de coût moyen présentent des performances satisfaisantes en termes de segmentation et de prédiction de la prime pure observée.

| | Incendie habitation | Incendie exploitation | TGN habitation | TGN exploitation |
|---|---|---|---|---|
| Gini sans *Open Data* | 27,48% | 50,7% | 26,74% | 44,5% |
| Gini avec *Open Data* | Non concerné | 50,7% | 34,79% | 44,7% |

En somme, les primes pures incendie et TGN modélisées permettent de mesurer le niveau de sinistralité réel des assurés. De plus, une segmentation améliorée du risque géographique est obtenue avec les nouveaux zoniers. Ces modèles de primes pures constituent un nouvel atout majeur dans l'ajustement tarifaire de la prime globale multirisque agricole.

# Executive summary

As part of a policy to improve AXA France's agricultural offer, a review of the pricing mechanism of the multi-risk agricultural product (MRA) was suggested. One of the particularities of the multi-risk agricultural contract is its ability to coverage two different risks : the farmer's home and his farm. This policy offers both common and specific cover for the home and the farm. The objective of this master thesis is to develop a pure premium model to improve the pricing of the MRA product.
The pure premium, or technical premium, is the expected amount of claims over a period of time. Currently, there is no pure premium model for the MRA product, its pricing is based on expert opinion, depending on the claims historical data and the rates offered by competing insurers. The creation of a pure premium model is a part of the tariff adjustment policy for the MRA product. Indeed, since 2017, strong tariff measures have been put in place in order to improve the proposed premium. these measures aimed to reduce the new business and portfolio tariff, which was considered too expensive by some policyholders. However, the use of all these measures has not resulted in a positive net contribution. In addition, the current zoning system no longer ensures a correct segmentation of the risk. This segmentation difficulty would be linked to its structure and its age (early 2000s).

Therefore, this study aims to address the following issues :
- **Understanding AXA France's multi-risk agricultural product** ;
- **Implementing a pure premium model for AXA France's MRA product** ;
- **Implementing a new zoning system**.

One of the first steps in the study is to understand the current pricing mechanism. This step is fundamental to create a new pricing structure, especially when modelling is involved. An inventory of the current pricing methodology serves as a proxy for identifying variables that would be more or less relevant in the modelling. In addition, an improvement in the segmentation quality of a model built with only internal variables is envisageable by using Open Data.

With the context and issues of the brief presented, two of the most important coverages of the multi-risk agricultural product are discussed : fire and storm-hail-snow coverages. An image-based modelling database of contracts from 2010 to 2021 is constructed and

enriched using Open Data. The variables created are taken from the databases constructed in in this work or from databases available directly on the internet. The latter relate to distance to fire centers, weather, crime, year of construction of properties and price per square meter.
Univariate descriptive statistics, the correlation and distribution analysis of the variables reveal the links between the different variables and the observed loss experience. Prior to any analysis or modelling, the data is processed by interpolation and imputation (by mean or mode).
In order to avoid having a claims cost model disturbed by extreme claims costs, these have been capped and mutualised. The capping and pooling of extreme claims costs is done using a threshold determined by the extreme value theory. In parallel, the Chain-Ladder method was used to determine the final amount of attritional and extreme claims which were not closed. A study of the claims burden made sense of using a single threshold for both home and business. In addition, a similar study showed differences between the factors influencing the frequency of claims for house and business risks. On the other hand, the analysis carried out on average cost modelling. In contrast, the analysis of average cost modelling showed that a single average cost model would not adversely affect the overall power of the model and would not be likely to lead to major rate error or adverse selection. The variables chosen for each of the models are selected by correlation analysis and coefficient penalisation.
The modelling structure adopted, taking into account operational constraints and the volume of data, is as follows :

$$Pure\ premium = (\mathbb{1}_{farm} * frequency_{farm} + \mathbb{1}_{housing} * frequency_{housing}) * average\ overall\ cost$$

The main results of the modelling are as follows :

1. ***Frequency models***

   — *Housing*
   The empirical frequency by number of rooms and contents of the dwelling contents makes up for the absence of a stable model for the frequency of fire claims. The SHN guarantee model is segmenting and stable on the learning base and the validation base.

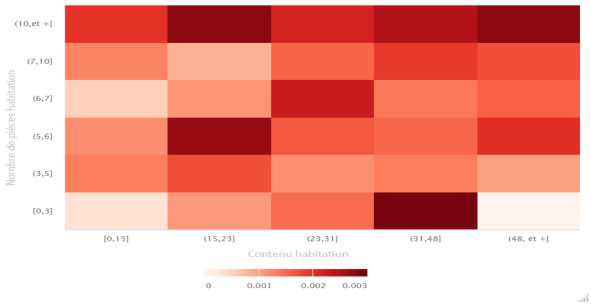



FIGURE 2 – Empirical frequency of home fire claims

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson SHN | 0,089 | 21,71 % | 0,089 | 21,23 % |

TABLE 2 – Results of the SHN housing loss frequency modelling

— *Farm*
The GLM models calibrated on the operational risk present very satisfactory performances in terms of segmentation and prediction. The comparison of these models with the XGBoost models shows that the use of complex models would not allow a significant increase in the segmentation of the loss frequency models.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| XGBoost Poisson fire | 0,099 | 45,41% | 0.098 | 44,23% |
| GLM Poisson fire | 0,118 | 49,08% | 0,118 | 48,44% |
| XGBoost Poisson SHN | 0,145 | 37,23% | 0,145 | 36,92% |
| GLM Poisson SHN | 0,193 | 36,59 % | 0,193 | 36,22 % |

2. ***Average cost models***
The average cost models seem to be less stable between the learning base and the validation base according to the Gini criterion because of the small volume of data. Nevertheless, these models describe quite correctly the observed claims trend with an acceptable temporal stability.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM gamma fire | 64150 | 13.22% | 64490 | 7.74% |
| GLM gamma SHN | 7185 | 9,27% | 7219 | 7,62% |

3. ***Integration of Open Data***
In the frequency models, the integration of Open Data into the modelling improves the segmentation with the appearance of discriminating risk trends. The improvement in the overall segmentation of the model seems to be more important for the SHN guarantee. In addition, the absence of an apparent discriminating trend when integrating the variables Open Data into the average cost model led to the abandonment of an average cost model with these variables.

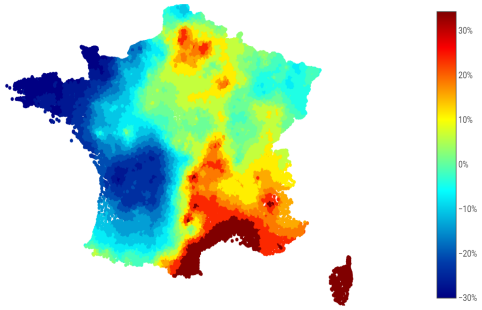
| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson without Open Data fire farm | 0,118 | 49,08% | 0,118 | 48,44% |
| GLM Poisson with Open Data fire farm | 0,118 | 49,98% | 0,118 | 49,08% |
| GLM Poisson without Open Data SHN housing | 0,089 | 21,71 % | 0,089 | 21,23 % |
| GLM Poisson with Open Data SHN housing | 0,089 | 36,71 % | 0,089 | 33,64 % |
| GLM Poisson without Open Data SHN farm | 0,193 | 36,59 % | 0,193 | 36,22 % |
| GLM Poisson with Open Data SHN farm | 0,193 | 38,7% | 0,193 | 38,14% |

4. ***Adding of zoning plans***
   The interest of building a zonier is that it could improve the prediction and segmentation capacity of the model.

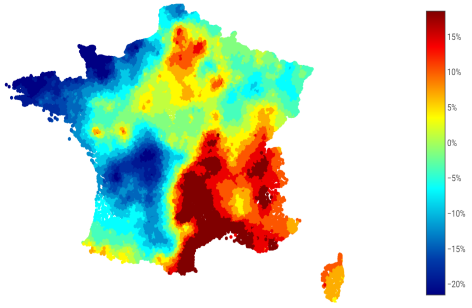
$$g(Y_i) - \hat{\beta} X_i = R_i$$
$$= \textit{geographical factor} + \textit{hazard}$$

   The construction of zoning plans was done only for farm, because of the low coverage rate of the communes for the housing risk. The approach used for the construction of zoning plans of frequency of exploitation claims is a Bayesian approach with a gaussian distribution of the geographical factors. Zoning plans are constructed from the residuals of each model. These zoning plans are then fed back into their respective models as new variables. In this paper, zoning plans constructed are claims frequency zoning. Due to the different nature of the risks modelled within each cover, it is necessary to construct a zoning plan for each cover.

   — *Fire zoning plan*

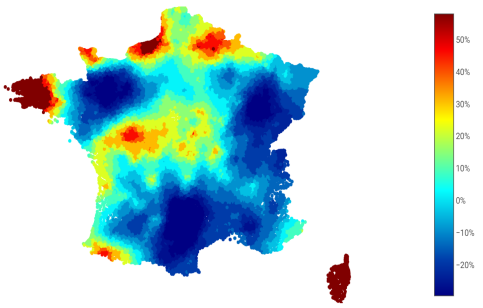



(a) Zoning from the model without Open Data

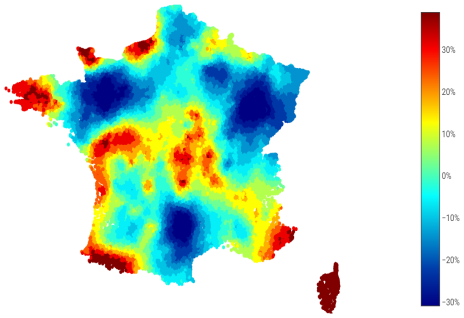



(b) Zoning from the model with Open Data

   — *SHN zoning*




(c) Zoning from the model without Open Data




(d) Zoning from the model with Open Data

   The addition of zoning plans in the models improves the overall segmentation of the different models. In addition, the interpretation of geographical information

is facilitated by the loss trends observed in the data. Finally, these zoning plans provide a finer segmentation of risk according to the cover. As well, these zoning plans correct the bias present in the old a zoning plan due to the fact that it was calibrated on all covers.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson without *Open Data* + Fire zoning | 0,117 | 51,63% | 0,117 | 49,84% |
| GLM Poisson with *Open Data* + Fire zone | 0,117 | 51,85% | 0,117 | 49,85% |
| GLM Poisson without *Open Data* + SHN zoning | 0,193 | 42,24% | 0,193 | 40,49% |
| GLM Poisson with *Open Data* + TGN zoning | 0,193 | 43,16% | 0,193 | 41,25% |

5. ***Pure premium***
   The pure premium models resulting from the different frequency and average cost models perform well in terms of segmentation and prediction of the observed pure premium.

| | fire housing | fire farm | SHN housing | TGN exploitation |
|---|---|---|---|---|
| Gini without Open Data | 27,48% | 50,7% | 26,74% | 44,5% |
| Gini with Open Data | Not relevant | 50,7% | 34,79% | 44,7% |

In conclusion, the modelled pure fire and TGN premiums allow for the measurement of the actual loss experience of the insured. In addition, an improved segmentation of the geographical risk is obtained with the new zoning plans. These pure premium models constitute a new major asset in the tariff adjustment of the global agricultural multi-risk premium.

# Remerciements

Je tiens à adresser mes remerciements les plus sincères à Mme. Oksana ALLAIRE et Mme. Floriane PIVETAU pour leurs implications dans la réalisation de ce mémoire.

Je remercie M. Mohamed HALIMI et M. Thomas GAUTHRON pour la proposition du sujet du mémoire et leurs conseils.

Un grand merci à M. Charles PARTINGTON, M. Hugo HAMMERER et M. Romain TOESCA pour leurs disponibilités et leurs conseils dans la modélisation.

Je tiens à remercier également toute l'équipe produit Multirisque Professionnelle pour leur aide.

Je tiens à remercier l'ensemble des enseignants de l'EURIA pour la formation de qualité au cours de ces trois années et particulièrement M. Jean-Marc DERRIEN pour l'encadrement au cours de l'alternance.

Enfin, je tiens à remercier chaleureusement ma famille pour le soutien infaillible durant cette période.

# Table des matières

**Note de synthèse** **iii**

**Executive summary** **viii**

**Remerciements** **xiii**

**Introduction** **1**

**1 Contexte et enjeux** **2**
1.1 AXA France : vision et fonctionnement . . . . . 2
1.2 Présentation du marché . . . . . 2
1.2.1 Définitions . . . . . 2
1.2.2 Le marché de l'agriculture . . . . . 3
1.3 Présentation du produit Multirisque Agricole (MRA) . . . . . 5
1.3.1 Contrat MRA . . . . . 5
1.3.2 Les garanties couvertes . . . . . 6
1.3.3 Distinction entre MRA et Assurance récolte . . . . . 7
1.4 Enjeux du mémoire . . . . . 8
1.4.1 Le tarif . . . . . 8
1.4.2 Présentation de la méthode actuelle de tarification . . . . . 9
1.4.3 Open Data . . . . . 11

**2 Présentation de la base de données** **13**
2.1 Base par images de contrats . . . . . 13
2.1.1 Base des contrats . . . . . 13
2.1.2 Base des sinistres . . . . . 14
2.1.3 Fusion des deux bases . . . . . 16
2.2 Statistiques descriptives . . . . . 17
2.2.1 Analyse univariée . . . . . 18
2.2.2 Étude des corrélations . . . . . 21
2.3 Les données en *Open Data* . . . . . 22
2.3.1 Base des centres de secours incendie . . . . . 22
2.3.2 Base de données météorologiques NOAA . . . . . 24
2.3.3 Base de criminalité ONDRP . . . . . 26

2.3.4 Base Valeur foncière DVF . . . 28
2.3.5 Base des années de construction des logements . . . 28
2.3.6 Base nationale des bâtiments (BNDB) . . . 29

**3 Traitement de la base de données 31**
3.1 Traitement des valeurs extrêmes . . . 31
3.1.1 Contexte . . . 31
3.1.2 Théorie mathématique du choix du seuil . . . 31
3.1.3 Application à nos données . . . 34
3.2 Traitement des sinistres en cours . . . 39
3.2.1 Présentation de la méthode . . . 39
3.2.2 Application . . . 40
3.3 Traitement des valeurs manquantes . . . 41
3.3.1 Théorie . . . 41
3.3.2 Application . . . 42

**4 Théorie sur la tarification d'un contrat d'assurance 43**
4.1 Modèle collectif . . . 43
4.2 Théorie sur les modèles linéaires généralisés . . . 44
4.2.1 Ajustement des lois . . . 45
4.2.2 Sélection des variables . . . 47
4.2.3 Application . . . 48
4.3 Théorie des modèles d'apprentissage statistique . . . 51
4.3.1 Arbres de décision . . . 51
4.3.2 *Gradient Boosting Machine* : GBM . . . 53
4.4 Évaluation des modèles . . . 55
4.4.1 Métriques d'évaluation des modèles . . . 55
4.4.2 Validation croisée . . . 57

**5 Modélisation de la fréquence et du coût moyen 59**
5.1 Contexte . . . 59
5.2 Modèles de fréquence . . . 60
5.2.1 Modélisation GLM . . . 60
5.2.2 Modélisation par apprentissage statistique . . . 67
5.3 Modèles de coût moyen . . . 69

**6 Apport de l'*Open Data* et zonier MRA 73**
6.1 *Open Data* . . . 73
6.1.1 Modèles de fréquence . . . 73
6.1.2 Coût moyen . . . 77
6.2 Construction des zoniers . . . 77
6.2.1 Contexte : Intérêt des nouveaux zoniers . . . 77
6.2.2 Approche bayesienne . . . 78
6.2.3 Application . . . 81

**7 Synthèse de la modélisation** **86**
7.1 Prime pure incendie . . . . . 86
7.1.1 Habitation . . . . . 86
7.1.2 Exploitation . . . . . 87
7.2 Prime pure TGN . . . . . 88
7.2.1 Habitation . . . . . 88
7.2.2 Exploitation . . . . . 90

**Conclusion** **90**

**A Test de Kolmogorov-Smirnov** **93**

**B Indice de Moran** **94**

**C *Grid search* XGBoost** **95**

**D Résidus quantiles GLM log-normale** **96**

**Bibliographie** **103**

# Introduction

Le secteur agricole en France est un secteur avec une importante valeur ajoutée à assurer. D'après une enquête réalisée en 2021 par l'INSEE, la production et la valeur ajoutée de l'agriculture en France était de 81,2 milliards d'euros. Il paraît alors nécessaire d'assurer la pérennité de l'activité des acteurs du marché agricole pour maintenir un niveau de productivité de la branche agricole dans l'économie nationale.

Le marché de l'assurance agricole étant un marché en situation d'oligopole, la proposition d'un tarif attractif pour capter et garder les assurés devient une clé de subsistance. Face aux enjeux liés aux différentes activités agricoles, un niveau de segmentation tarifaire minimum est nécessaire pour avoir une couverture et une prime adaptée au risque.

Dans une politique d'amélioration de l'offre agricole d'AXA France, une revue de la mécanique de tarification a été suggérée. La modification de la structure de tarification s'effectuera principalement par la mise en place d'un modèle de prime pure par garantie. L'un des intérêts de la mise en place d'une prime pure est qu'elle permettra de mesurer le montant uniquement nécessaire pour couvrir la sinistralité de l'assuré. De plus, la prime pure aura pour avantage de mieux segmenter la tarification des nouveaux assurés puisqu'il n'existait pas de modèle de prime pure au sein de la branche agricole d'AXA France.

Ce mémoire présentera l'étude effectuée pour la mise en place de la prime pure pour deux des garanties les plus importantes du tarif global multirisque agricole : les garanties incendie et tempête grêle neige. Cette étude se décomposera en sept chapitres. Dans un premier temps, une présentation détaillée du marché agricole, du produit multirisque agricole et des enjeux du mémoire sera effectuée. Le deuxième et le troisième chapitre seront dédiés à la construction et au traitement de la base de modélisation. Ensuite, une quatrième partie exposera la théorie sur les outils nécessaires à la tarification d'un contrat d'assurance en IARD. Dans une phase d'application, les chapitres 5 et 6 présenteront les résultats des différents modèles obtenus pour la construction de la prime pure. Enfin, le dernier chapitre sera consacré à l'évaluation de la prime pure modélisée sur les deux garanties.

# Chapitre 1

# Contexte et enjeux

## 1.1 AXA France : vision et fonctionnement

AXA France, dans sa vision d'agir pour le progrès humain, tout en protégeant ce qui compte, crée des produits sur mesure pour maîtriser au mieux les risques auxquels ses assurés sont confrontés. Étant aujourd'hui l'une des plus grandes marques d'assurance au monde, AXA France se doit de développer des produits d'assurance avec le plus haut niveau d'expertise qui soit. Pour ce faire, AXA France s'est organisé en différentes entités opérationnelles qui sont : AXA IARD Particuliers et Entreprises, AXA Santé et Collectives et AXA Epargne, Retaite et Prévoyance individuelle, dans le but de créer des produits d'assurance couvrant quasiment tous les domaines.
Plus spécifiquement au sein de l'entité opérationnelle AXA Particuliers et IARD Entreprises, la branche multirisque agricole a pour objectif de proposer la meilleure couverture possible aux agriculteurs afin d'assurer la pérennité de leurs activités. Cette couverture spécifique des agriculteurs est liée à la particularité de leur activité.

## 1.2 Présentation du marché

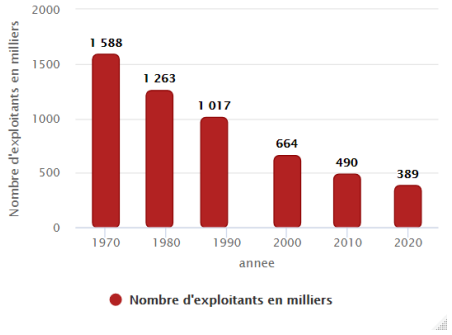
### 1.2.1 Définitions

— **Exploitation agricole** :
D'après l'INSEE , une exploitation agricole est une unité de production de produits agricoles gérée de manière indépendante.
Pour qu'une exploitation agricole soit reconnue, elle doit disposer d'un minimum de caractéristiques liées à sa superficie et sa production, qu'elle soit agricole ou animale.
Ces caractéristiques minimales sont les suivantes :
— Une superficie agricole utilisée au moins égale à un hectare ;
— Ou une superficie en cultures spécialisées au moins égale à 20 ares ;
— Ou une activité suffisante de production agricole, estimée en cheptel, surface cultivée ou volume de production.

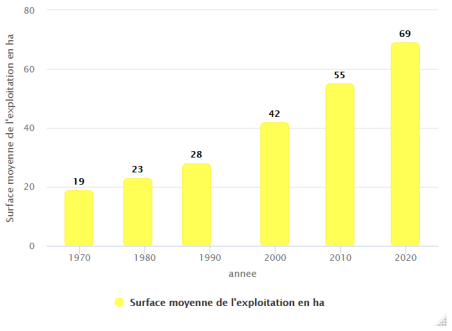
— **Exploitant agricole** :
Un exploitant agricole est une personne physique qui met en valeur une exploitation agricole. Il est celui qui a la plus grande responsabilité sur l'exploitation et assure sa gestion quotidienne. Il peut être propriétaire de l'exploitation ou locataire.

### 1.2.2 Le marché de l'agriculture

Le monde de l'agriculture en France fait face à une situation assez paradoxale. Le constat est le suivant : le nombre d'exploitants agricoles est en perpétuelle décroissance (1.1-a) tandis que la surface moyenne d'exploitation par exploitant ne fait que croître (1.1-b). D'après le recensement général de l'agriculture de 2020, il y aurait actuellement environ 389 000 exploitants agricoles en France contre 490 000 en 2010, ce qui traduit une baisse de 21% sur 10 ans. En revanche, comme énoncé plus haut, la surface moyenne des exploitations agricoles ne fait qu'augmenter, elle est passée de 55 hectares à 69 hectares en France métropolitaine, soit une augmentation de 25% sur 10 ans.




(a) Nombre d'exploitants agricoles en milliers en France




(b) Surface moyenne des exploitations agricoles en France

FIGURE 1.1 – Source : [AGRESTE, 2022]

Ce paradoxe pourrait s'expliquer en s'appuyant sur le graphique 1.2, d'une part, par la croissance du nombre de grandes exploitations (exploitations dégageant plus de 250 000 euros par an de production brute standard [AGRESTE, 2021]) au fil de ces dix dernières années. Leur nombre a évolué de 3,4% entre 2010 et 2020. Ces exploitations détiennent en moyenne 136 hectares de surface d'exploitation en 2020. D'autre part, le nombre de petites et moyennes exploitations (exploitations avec une production brute standard inférieure à 250 000 euros par an [AGRESTE, 2021]) ne fait que décroître au cours de ces dernières années.
L'hypothèse la plus probable qui expliquerait ce phénomène est l'absorption des moyennes exploitations par les grandes exploitations. Cette raison a été évoquée comme raison principale de la diminution des moyennes exploitations observée entre le recensement de 2000

et celui de 2010[AGRESTE, 2010a].

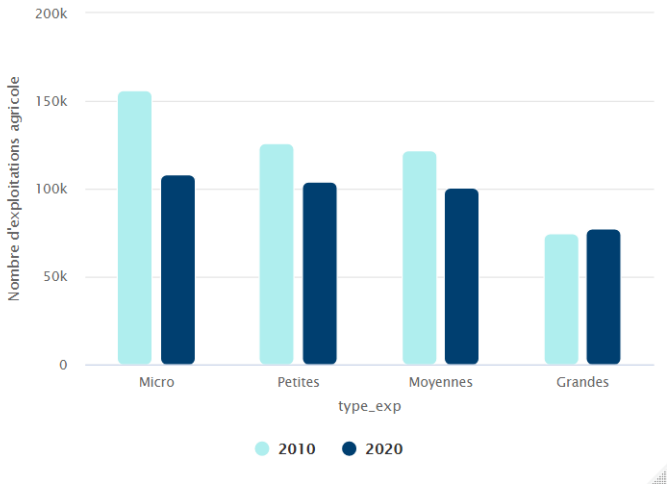



FIGURE 1.2 – Évolutions des différents types d'exploitations source : AGRESTE

De plus, dans un contexte de gains de productivité et de modernisation des techniques de production, la main d'œuvre humaine est remplacée par des machines plus performantes mais qui engagent plus de coûts de production. De même, la proportion de personnes se tournant vers le métier d'exploitant agricole ne fait que diminuer au fil des années. En 2019, la part des agriculteurs exploitants dans l'emploi (au sens du Bureau International du Travail (BIT)) était de 1,5% en France hors Mayotte [INSEE, 2020].

Par ailleurs, au regard des données de France Assureurs de 2020[France Assureurs, 2020], le marché de l'assurance multirisque agricole est assez rentable. En effet, les sinistres valent environ deux tiers du montant des cotisations entre 2016 et 2020 (tableau 1.1).

| Année | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|
| Cotisations : (M€) | 790 | 804 | 818 | 828 | 837 |
| Charges des sinistres : (M€) | 434 | 523 | 507 | 546 | 527 |
| Ratio S/C | 55% | 65% | 62% | 66% | 63% |

TABLE 1.1 – Ratio S/C par année : FRANCE ASSUREURS

En outre, le marché de l'assurance multirisque agricole est détenu majoritairement par très peu d'acteurs. Parmi ceux-ci, AXA France est classé troisième en termes de part de marché, après le leader Groupama et Crédit Agricole. Dans ce contexte concurrentiel, AXA France développe son empreinte sur le marché en proposant une assurance sur mesure avec un contrat multirisque agricole adapté aux clients.

Enfin, la pandémie du COVID-19 n'a eu que très peu d'impact sur le secteur agricole. Au cours des trois confinements, les exploitants agricoles n'ont pas arrêté leur activité. Selon une étude menée par l'institut IFOP [IFOP, 2020] (institut d'études opinion et marketing) en décembre 2020, l'ensemble des exploitants interrogés disent avoir eu la même charge de travail que l'année précédente (2019). Pour corroborer cet avis, des articles de la revue scientifique *Cahiers Agricultures (Volumes 29,30, 2020-2021)*[Cahiers Agricultures, 2021] montrent que la production agricole a été peu perturbée, même pendant les périodes de confinement. De plus, les contraintes logistiques pour le transport des productions agricoles n'ont eu d'impact qu'au début de la crise sanitaire.

## 1.3 Présentation du produit Multirisque Agricole (MRA)

### 1.3.1 Contrat MRA

D'après le Code des Assurances, *« Les assurances de dommages obligent l'assureur à indemniser l'assuré des conséquences d'un sinistre sur son patrimoine. Elles regroupent les assurances de choses et les assurances de responsabilité »* [Code des assurances, 2022]. Un sinistre est la réalisation d'un évènement aléatoire prévu dans le contrat d'assurance et statistiquement prévisible qui affecte les biens ou l'intégrité physique d'une personne, un lieu ou une entité [Boursedescredits, 2022]. Ainsi, il est assez raisonnable pour des agriculteurs qui mènent une activité exposée à de multiples aléas, climatiques et humains, de souscrire à un contrat d'assurance pour leur activité.

La souscription à un contrat d'assurance repose sur le principe du cycle inversé de production. Prenons l'exemple des agriculteurs pour expliquer ce principe fondamental en assurance IARD. D'une part, l'agriculteur paye une cotisation à l'assureur bien avant qu'un sinistre ne survienne sans connaître le montant du sinistre qui pourrait survenir. D'autre part, l'assureur reçoit la cotisation de l'agriculteur sans avoir effectué aucune prestation et s'engage à l'indemniser à la survenance d'un sinistre. Une telle situation où le montant de la cotisation est payé avant la prestation, ne s'observe qu'en assurance et est appelée le cycle inversé de production (figure 1.3).

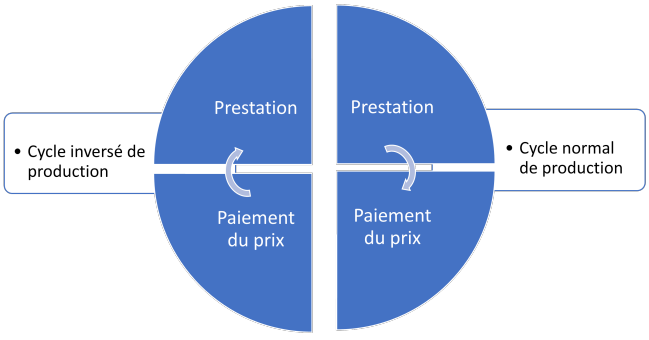



FIGURE 1.3 – Cycle de production

### 1.3.2 Les garanties couvertes

Un contrat Multirisque Agricole (MRA) a pour but d'assurer l'ensemble des biens et responsabilités professionnels et privés des exploitants en activité ou non. Ce contrat est destiné aux propriétaires et locataires dont l'activité ou les biens sont à vocation agricole avec des exploitations enregistrées en nom propre ou sous forme sociétaire (GAEC, SCEA, EARL, SA, etc.). Les principales activités agricoles couvertes par ce périmètre sont les suivantes :

- La polyculture : elle correspond à la production de végétaux (céréales, betteraves, pommes de terres, etc.) ;
- L'élevage traditionnel : il correspond à l'élevage d'animaux en troupeaux. Le troupeau est constitué d'un ensemble d'animaux de même espèce dont l'élevage représente l'activité principale d'une exploitation ;
- L'élevage intensif : il nécessite des bâtiments et installations spécialisés destinés à recevoir des lots ou bandes importants d'animaux selon un protocole zootechnique et sanitaire prédéterminé, notamment des techniques d'alimentation industrialisées, faisant appel à des aliments produits ou transformés à l'extérieur de l'exploitation assurée ;
- La viticulture : elle s'étend de la production de vins d'appellation régionale aux grands crus.

Une des particularités du contrat multirisque agricole est sa capacité à couvrir deux risques différents qui sont l'habitation et l'exploitation de l'agriculteur. Ce contrat propose des garanties à la fois communes et spécifiques à l'habitation et l'exploitation.
Les garanties proposées sont les suivantes :
**Évènements garantis pour le risque Habitation :**

| **Formule multirisque** | **Formule Minimum** |
|---|---|
| Incendie / Explosions et Risques divers | Incendie / Explosions et Risques divers |
| Évènements climatiques | Évènements climatiques |
| Attentats et actes de terrorisme | Attentats et actes de terrorisme |
| Dommages électriques | Catastrophes naturelles |
| Catastrophes naturelles | |
| Dégâts des eaux | |
| Bris de glaces | |

Table 1.2 – Garanties risque habitation

A la souscription d'un contrat MRA deux formules de couvertures sont proposées au souscripteur pour l'assurance de son habitation :

- la formule minimum : elle correspond à une couverture minimale avec les garanties de base ;
- la formule multirisque : elle correspond à une formule plus complète avec plus de garanties ;

Une extension « Perte de denrées en congélateurs » est proposée en option quelque soit

la formule de l'assurance habitation choisie.
Pour une couverture adaptée du risque, l'offre proposée doit correspondre aux besoins de l'assuré. C'est pourquoi le contrat MRA propose en plus des garanties nécessaires à la couverture de toutes les activités, des garanties spécifiques. Ces garanties sont de deux types : automatiques ou facultatives comme l'indique le tableau 1.3.
**Évènements garantis pour le risque Exploitation :**

| **Garanties automatiques** | **Garanties facultatives** |
|---|---|
| **Toutes activités confondues** | |
| Incendie / Explosions et Risques divers<br>Evénements climatiques<br>Attentats et actes de terrorisme<br>Dommages électriques<br>Catastrophes naturelles | Dégâts des eaux<br>Bris de glaces<br>Vol / Vandalisme<br>Perte de liquides<br>Bris de machines<br>Marchandises transportées. |
| **Garanties spécifiques par activité** | |
| Elevage traditionnel laitier | Perte de lait en tanks réfrigérés (Contenant et Contenu) |
| Elevage intensif | Accident d'élevage (mort des animaux) |
| Viticulture | Coulage (Contenant et Contenu)<br>Bris de bouteilles |
| Arboriculture | Chambres froides |
| Polyculture | Séchage |

TABLE 1.3 – Garanties risque exploitation

### 1.3.3 Distinction entre MRA et Assurance récolte

Dans un contrat MRA l'agriculteur a la possibilité d'inclure ou non la couverture de son habitation, car son habitation peut déjà avoir une assurance multirisque habitation chez AXA France ou chez un autre assureur.
Il est également important de faire la distinction entre un contrat MRA et l'assurance récolte. Un contrat MRA est axé sur les bâtiments d'exploitation et d'habitation de l'agriculteur (l'assuré). Le contrat MRA n'a pas vocation de couvrir la récolte en terre, qui est prise en charge par l'assurance récolte. Néanmoins, pour les produits d'origine végétale non récoltés, l'assureur s'engage à un dédommagement en cas de sinistre d'un montant équivalent à une année de production [1].
En outre, un contrat MRA couvre le contenu agricole récolté, le mobilier, la valeur des espèces animales (en cas de décès) et le matériel agricole de l'exploitation. Le matériel agricole non automoteur, comme le tracteur, n'est pas couvert par le contrat MRA, mais plutôt couvert par un contrat matériel agricole (MatAgri) d'AXA France.

1. A l'exception de l'évènement grêle qui est assuré par Swiss grêle

## 1.4 Enjeux du mémoire

### 1.4.1 Le tarif

Ce mémoire a pour objectif de mettre en place un modèle de prime pure du produit multirisque agricole d'AXA France. La prime pure ou prime technique correspond au montant attendu des sinistres d'un assuré sur une période. Actuellement, il n'existe pas de modèle de prime pure, la tarification du produit MRA est basée sur l'avis d'experts, en fonction de la sinistralité et des tarifs proposés par les assureurs concurrents. La création d'un modèle de prime pure s'inscrit dans une politique d'ajustement tarifaire du produit MRA. En effet, depuis 2017 des mesures tarifaires fortes ont été mises en place afin d'améliorer le tarif proposé. Ces mesures visaient une réduction du tarif affaire nouvelle et portefeuille, jugé trop cher par certains assurés.

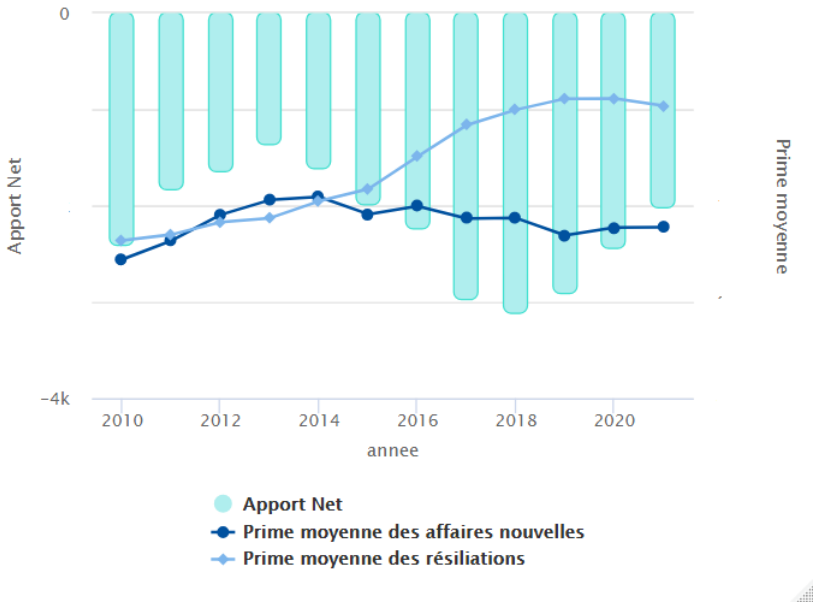



FIGURE 1.4 – Évolution de l'apport net

Le graphique 1.4 permet d'observer l'amélioration de l'apport net [2] depuis 2018 qui résulte des mesures tarifaires. Cependant, même s'il s'améliore, l'apport net ne redevient pas positif avec l'usage de toutes ces mesures. Un autre point à observer sur ce graphique est la tendance à la baisse du tarif affaire nouvelle à partir de 2017. Cette dernière s'observe également sur la prime des résiliations à partir de 2019.
L'une des faiblesses de ces mesures est qu'elles sont appliquées sur une mécanique tarifaire qui n'a pas été actualisée, d'où la nécessité d'un modèle de prime pure pour mesurer le risque réel de chaque assuré .

2. Différence entre le nombre de souscriptions et le nombre de résiliations.

La prime est la principale ressource financière d'une compagnie d'assurance IARD. Elle doit donc être calculée avec la plus grande attention. La prime représente le coût du transfert de risque de l'assuré vers l'assureur. Bien évidemment, l'assureur veut que cet échange soit un « business » rentable. De ce fait, il paraît primordial d'avoir un tarif qui permet à la fois à l'assureur d'être rentable sur le long terme et qui soit abordable pour l'assuré, afin que celui-ci souscrive un contrat chez l'assureur plutôt que chez un concurrent. Pour réaliser cet arbitrage dans le marché concurrentiel de l'assurance, les assureurs appliquent deux principes : la mutualisation et la segmentation.
La mutualisation est le principe selon lequel, pour un groupe d'assurés exposés au même risque, les primes collectées aux assurés (sinistrés et non sinistrés) servent à couvrir les sinistres survenus aux assurés sinistrés du groupe. Ce principe permet à l'assureur de ne pas faire peser le coût du risque sur une seule prime provenant d'un assuré et d'ajuster son tarif afin qu'il soit attractif.
Le second aspect clé de la tarification, et non des moindres, est la technique de segmentation. Par cette technique l'assureur cherche à identifier des classes de risque, afin d'ajuster le tarif de chaque assuré en fonction de la classe dans laquelle il se trouve. Ce principe permet à l'assureur de réduire le phénomène antisélection. Un moyen d'améliorer la segmentation consiste à créer des zones de risques homogènes en termes de risque géographique. Cette segmentation est capturée au sein d'un zonier qui fait une correspondance entre une zone et un niveau de risque.
Par exemple, un assuré se trouvant dans une zone fortement exposée au vol aura un tarif plus élevé qu'un autre dans une zone moins exposée.

Ce mémoire a trois enjeux majeurs :
- **Comprendre au mieux le produit multirisque agricole d'AXA France** ;
- **Mettre en place un modèle de prime pure pour le produit MRA d'AXA France** ;
- **Mettre en place un nouveau zonier**.

### 1.4.2 Présentation de la méthode actuelle de tarification

Comprendre la mécanique de tarification actuelle est une étape fondamentale pour créer une nouvelle structure de tarification, surtout quand il s'agit de modélisation. Cette étape consiste à faire un état des lieux de la méthode de tarification actuelle, qui peut servir de proxy pour identifier les variables qui seraient plus ou moins pertinentes dans la modélisation de la prime pure. Par soucis de confidentialité, cette section ne présentera que les éléments composant la structure du tarif.

Pour rappel, un contrat MRA a la possibilité de couvrir l'habitation et l'exploitation de l'assuré. Le tarif final est la somme des tarifs des risques exploitation et habitation. Toutefois, l'assuré a la possibilité de ne pas assurer son habitation avec son contrat MRA, dans ce cas le tarif habitation est nul.

1. **Tarification de l'habitation**
   La tarification habitation correspond à la somme des tarifs des différents bâtiments d'habitation de l'agriculteur. Il convient de noter que le contrat MRA peut couvrir au maximum 6 bâtiments d'habitation. Le tarif habitation d'un bâtiment dépend principalement de quatre informations liées à son habitat :
   — **Type de propriété de l'habitat** : le tarif lié à l'habitat varie en fonction de la qualité de l'assuré sur l'habitat. Trois modalités de réponse sont disponibles pour cette variable : locataire, propriétaire, ou propriétaire non occupant du bâtiment d'habitation. Dans le cas où l'assuré a la modalité « propriétaire non occupant », cela signifie que l'habitation est habitée par un exploitant locataire.
   — **Nombre de pièces** : il s'agit du nombre de pièces habitables dans le bâtiment d'habitation.
   — **Contenu de l'habitation** : il s'agit de la valeur maximale des biens mobiliers assurés au sein du bâtiment. Ce montant est le montant de base qui servira à l'allocation des contenus associés à chaque garantie. Par exemple, le contenu vol est $x\%$ du contenu et ce contenu correspond à la valeur maximale assurable en cas de vol.
   — **Taux d'objets précieux** : il s'agit de la proportion d'objets précieux dans un bâtiment d'habitation.
   — **Nombre de congélateurs**.

2. **Tarification de l'exploitation** :
   Le principe est le même que pour la tarification de l'habitation. Toutefois, la tarification des bâtiments d'exploitation relève d'un niveau de complexité supérieur, en raison des interactions entre les variables tarifaires. Le contrat MRA peut couvrir plusieurs bâtiments d'exploitation [3] ainsi que les terres correspondantes. Lorsque plusieurs bâtiments d'exploitation se trouvent sur la même exploitation, le tarif exploitation est alors égal à la somme des tarifs des différents bâtiments.
   — **Activité** : il s'agit de l'activité agricole de l'exploitant. Chaque activité présente un risque particulier, d'où l'existence d'un ajustement du tarif en fonction de l'activité.
   — **Zone tarifaire** : il s'agit de la zone où se trouve le risque. Le zonier dépend de l'activité et du département.
   — **Type de propriété de l'exploitation** : indique la qualité en laquelle agit le souscripteur :
      — Propriétaire ;
      — Propriétaire non exploitant avec locataire : il n'exploite pas et il n'occupe pas la ferme ;
      — Locataire ;
      — Métayer : exploitant qui fait valoir la terre et partage les récoltes avec le propriétaire ;

3. Au maximum 32 bâtiments peuvent être couverts par un contrat MRA.

  - Propriétaire non exploitant occupant : il n'exploite pas mais il occupe la ferme.

  L'assuré peut être propriétaire d'un bâtiment d'exploitation et locataire d'un autre.
- **Surface de l'exploitation** : il s'agit de la surface de l'exploitation sur laquelle se trouve le bâtiment d'exploitation.
- **Type de bâtiment** : le type de bâtiment permet de savoir si le bâtiment d'exploitation est un bâtiment moderne ou traditionnel en fonction des standards de construction définis par les experts.
- **Niveau d'amélioration et d'isolation du bâtiment** : un ajustement tarifaire est effectué en fonction des améliorations, en termes de sécurité ou de matériel de production, réalisées dans le bâtiment d'exploitation. Il en est de même pour le type d'isolation et de chauffage installé dans le bâtiment.

En somme, cette méthode de calcul de la prime commerciale peut être optimisée en utilisant un modèle qui prendrait en compte plus de variables ou qui permettrait d'avoir une meilleure segmentation. Raison pour laquelle ce mémoire vise à modéliser une prime pure MRA, et éventuellement améliorer la qualité de segmentation du modèle avec l'*Open Data*.

### 1.4.3 Open Data

La question de l'utilisation des données en *Open Data* a suscité notre intérêt pour une amélioration de capacité de segmentation des modèles. Aujourd'hui, une grande masse de données est récoltée et mise à disposition du grand public sur internet. Cette disponibilité des données a été favorisée par la loi Lemaire du 7 octobre 2016 [LEGIFRANCE, 2016] qui s'articule autour de quatre points :

- la circulation des données et du savoir ;
- la protection des droits dans la société numérique ;
- l'accès au numérique ;
- les dispositions relatives à l'Outre-mer.

Cette loi a pour objectif de permettre à tous les citoyens d'accéder à la donnée et de créer une opportunité de développement et de croissance pour la France.

Au regard de cette masse de données disponibles et des garanties couvertes, il se pourrait que certaines informations récoltées dans le parcours de souscription soient disponibles en *Open Data*. De plus, il est probable que certains sinistres soient expliqués par des variables en *Open Data*. Par conséquent, si cette étude s'avère concluante, il serait éventuellement possible de réduire le parcours de souscription et/ou d'ajouter des variables qui permettraient d'avoir une prime pure mieux segmentée. En effet, certaines variables exogènes liées à la géolocalisation, au climat et aux infrastructures s'imposent à nos risques et pourraient avoir un effet marginal sur la sinistralité. L'influence de ces variables pourrait être d'autant plus significative avec une précision plus élevée de la maille d'observation. Par exemple, il serait grossier d'attribuer le niveau de sécurité d'une région à toutes les

communes de la région du fait du caractère hétérogène de la délinquance d'un quartier à un autre.
Cependant, un aspect non négligeable de l'*Open Data* est le risque de pérennité de ces variables externes dans le temps. En effet, l'absence ou la modification future de l'une de ces variables qui aurait été introduite dans la modélisation impliquerait la réalisation d'une nouvelle modélisation. En outre, un tel évènement induirait un biais de comparabilité de la prime modélisée avant et après cet évènement. En somme, un arbitrage pérennité-pertinence s'avère nécessaire avant toute intégration de variable disponible en *Open Data* dans la structure finale de prime pure.

**A retenir** :
- **Risques couverts** : bâtiments d'habitation et d'exploitation ainsi que leur contenu, cultures non récoltées sur un an ;
- **Contexte actuel** : apport net négatif, tarif jugé trop cher par certains profils de risque ;
- **Tarification actuelle** : prime commerciale globale basée sur l'avis d'experts et selon les tendances du marché ;
- **Solution suggérée** : mise en place d'une prime pure par garantie ;
- **Amélioration de la segmentation** : intégration des données en *Open data* et création d'un nouveau zonier.

# Chapitre 2

# Présentation de la base de données

Le cadre général de l'étude étant présenté, la première étape de la phase pratique consistera à construire une base de modélisation. Cette étape constitue le socle de l'étude et mérite donc une attention particulière.

## 2.1 Base par images de contrats

### 2.1.1 Base des contrats

Un contrat d'assurance peut être considéré comme un être humain. Celui-ci a une date d'affaire nouvelle qui correspond à sa date de naissance et une date de résiliation qui correspond à sa date de décès. Comme un être humain, il peut évoluer dans le temps. Dans ce cas, on parle d'avenant au contrat, ou de remplacement, pour signifier que les caractéristiques du contrat ont changé. Différentes causes peuvent déclencher un avenant au contrat, et donc un remplacement, comme par exemple un déménagement (modification de l'adresse du risque), une hausse ou une baisse de la surface d'exploitation, l'ajout ou la suppression d'options du contrat, etc. Il est nécessaire d'enregistrer la date de chaque changement, appelée date de remplacement. Il peut y avoir plusieurs remplacements au cours d'une même année. Par conséquent, il est crucial de construire une base de données avec toutes les images, ou photographies de tous les remplacements, du contrat pour une bonne modélisation.

En pratique, une extraction avec une vision mensuelle de la base de données annuelle est effectuée pour les douze mois de l'année. L'étape suivante consiste à regrouper toutes les visions de chaque contrat et de garder les images différentes de chaque risque, c'est-à-dire de chaque contrat, sur une année.

$$DTDEBUT = \max(DTFAN, 1^{er}\textit{janvier de l'année}, DTFRP)$$

$$DTFIN = \min(DTFRS, 31\textit{Décembre de l'année})$$

$$Exposition_{image} = \frac{DTFIN_{image} - DTDEBUT_{image}}{365}$$

Avec :

— DTFAN : Date d'effet de l'affaire nouvelle
— DTFRP : Date d'effet du remplacement
— DTFIN : Date de fin de l'image
— DTDEBUT : Date de début de l'image
— DTFRS : Date de résiliation de l'image

Le mécanisme est le suivant :

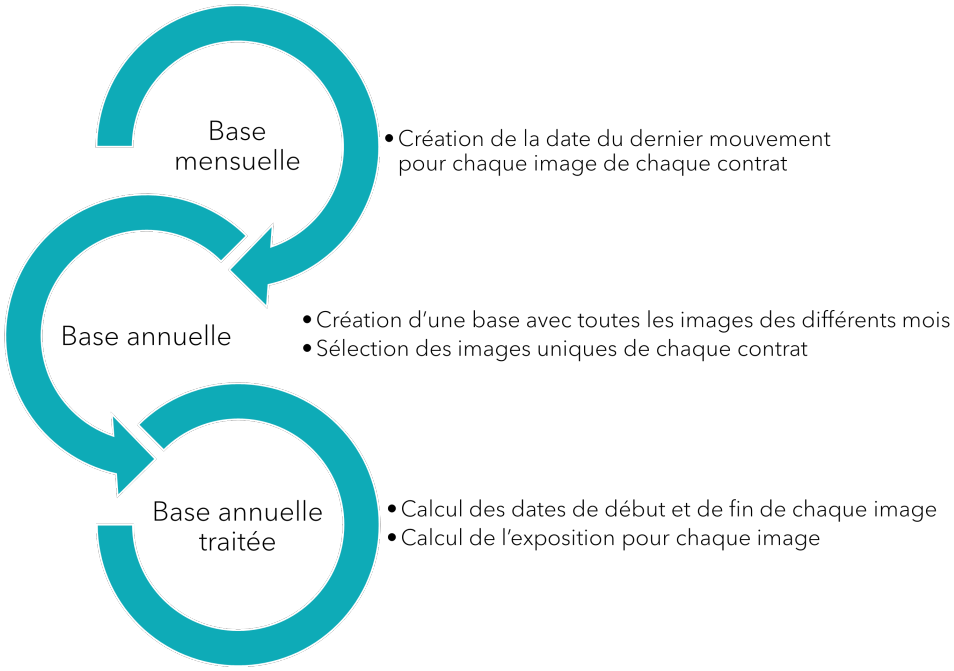



FIGURE 2.1 – Étapes de construction de la base des contrats

Pour la réalisation du mémoire, les bases des contrats de 2010 à 2021 sont utilisées.

### 2.1.2 Base des sinistres

La construction de la base des sinistres est une étape clé de la modélisation. Cette base contient l'ensemble des informations liées à chaque sinistre. L'une des étapes importantes de sa construction est l'association sinistre-garanties. En effet, un sinistre peut être lié à plusieurs garanties ; par exemple, un sinistre dégât des eaux peut engendrer un dommage électrique.
L'obtention d'un modèle robuste passe par une association correcte entre image de risque et sinistre. En pratique, les clés de jointure sont le numéro de contrat, la date de survenance du sinistre, ainsi que la date de début et de fin de l'image de risque.

Le processus est décrit de manière suivante :

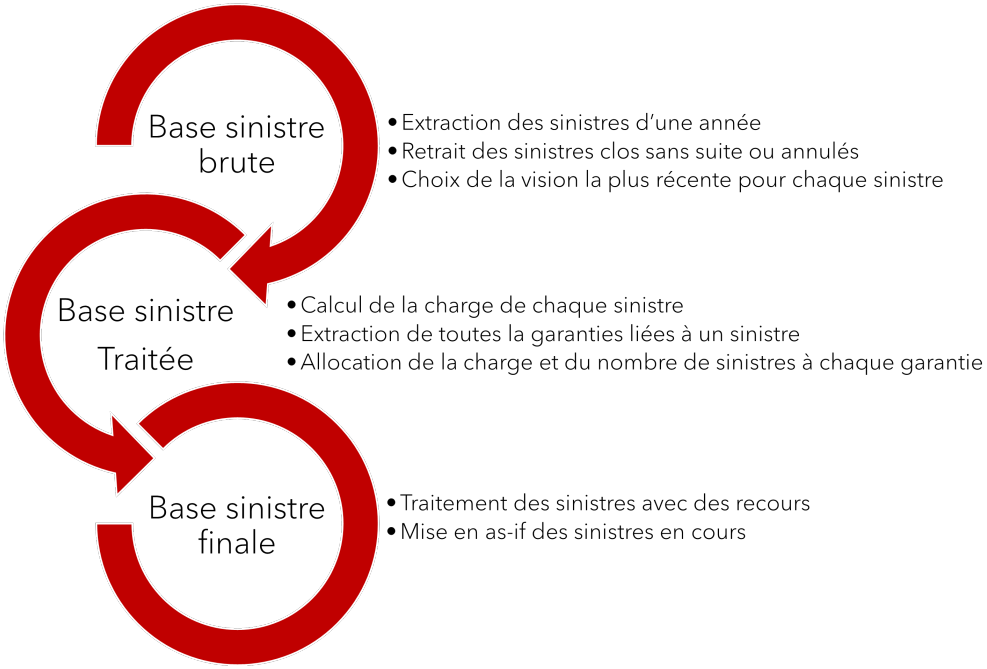



FIGURE 2.2 – Étapes de construction de la base des sinistres

***Mise en « as-if »***

La mise en « as-if » est une étape de la construction de la base de modélisation, qui consiste à mettre à jour les montants des sinistres comme s'ils avaient eu lieu aujourd'hui. Cette actualisation du montant de la charge des sinistres se fait en tenant compte de l'inflation du marché. Pour mesurer cette inflation, le proxy utilisé au sein de la branche multirisque agricole d'AXA France est l'indice du coût de la construction de la Fédération française du bâtiment (FFB)[Goodassur, 2022]. Cet indice est produit trimestriellement depuis le $1^{er}$ janvier 1941.

Il est logique de penser qu'un sinistre qui a eu lieu en 2012 n'a plus la même valeur en 2021. L'utilisation de l'indice se fait alors de la manière suivante : pour un coût de sinistre de 10000 €, un indice au $1^{er}$ trimestre de l'année 2012 de 901 et un indice au $1^{er}$ trimestre de l'année 2021 de 1022,3, on a :

$$Sinistre\ actualisé = 10000 \times \frac{1022,3}{901} = 11346,28\ €$$

En pratique, l'indice utilisé est la moyenne géométrique [1] sur les 4 trimestres de l'année,

1. Définition de la moyenne géométrique : $\overline{x} = \sqrt[n]{x_1 \times x_2 \times ... \times x_n}$
L'utilisation de la moyenne géométrique est mieux adaptée aux calculs de taux et d'indices moyens, car elle correspond au taux $\overline{x} annuel constant donnant le même taux final$.

de 2010 à 2021.

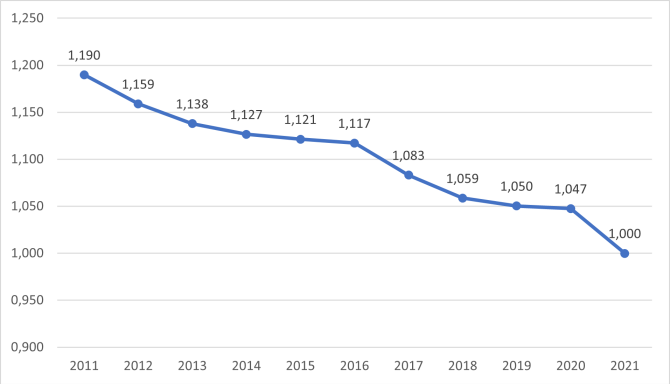



FIGURE 2.3 – Évolution de facteurs d'actualisation

### 2.1.3 Fusion des deux bases

La fusion entre la base des sinistres et la base des contrats se fait par association entre chaque image de risque et les sinistres survenus entre la date de début et de fin de l'image. L'étape suivante consiste à agréger les sinistres par garantie pour chaque image, pour avoir une ligne par image.

Une habitation abrite des êtres humains avec comme contenu du mobilier particulier et de l'électroménager. En revanche, pour l'exploitation, l'assurance porte sur les récoltes conservées dans les bâtiments, le bétail, le matériel agricole, etc. Il est important de distinguer ces différents types de risques. Trois risques sont alors extraits de chaque image :

— Un risque habitation lié aux garanties habitation, si l'habitation est assurée ;
— Un risque exploitation lié aux garanties exploitation ;
— Un risque commun lié aux garanties communes à l'habitation et l'exploitation.

Pour résumer le schéma est le suivant :


FIGURE 2.4 – Création de la base de modélisation

## 2.2 Statistiques descriptives

Face à la contrainte de temps, le choix des garanties à modéliser est fait en fonction de la volumétrie des données et du poids de celles-ci dans la prime pure observée. Partant de l'approche fréquence-coût pour la modélisation de la prime pure, la prime pure observée se définit de la manière suivante :

$$\text{Prime pure observée} = \frac{\text{Coût du sinistre}}{Exposition}$$

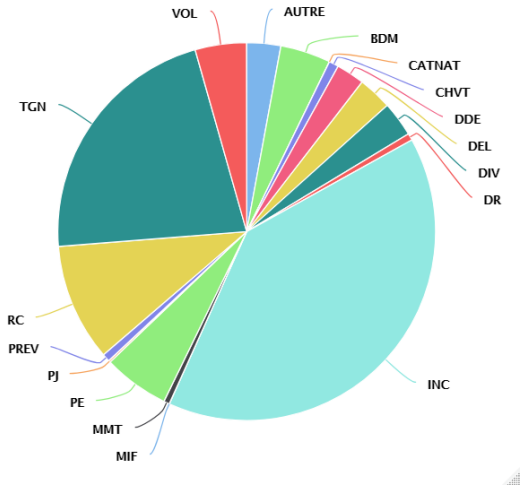



FIGURE 2.5 – Poids des primes pures observées par garantie

Le graphique 2.5 montre que les garanties [2] les plus coûteuses sont l' incendie (38%) et la tempête grêle neige (20%), soit 58% de la prime pure observée globale. Il s'agit de deux garanties obligatoires. En outre, la volumétrie en termes de sinistralité est favorable à la mise en place d'un modèle et est supérieure à notre seuil de 7000 sinistres défini à dire d'expert. Une présentation du contenu de ces deux garanties permet de mieux les appréhender :

— **Garantie Incendie** :
Les évènements couverts par cette garantie sont l'incendie, l'explosion, la chute directe de la foudre, le choc d'un véhicule terrestre identifié, le choc ou la chute d'un appareil de navigation aérienne ou spatiale ou d'objets tombants. Outre ces évènements, sont également couverts :
— L'émission soudaine de fumées provenant du fonctionnement défectueux d'un appareil ou de l'incendie d'un bâtiment voisin ;
— L'électrocution et la fulguration des animaux ;

2. N.B : *Autre* regroupe l'ensemble des garanties annexes

- L'asphyxie des animaux consécutive à un incendie.

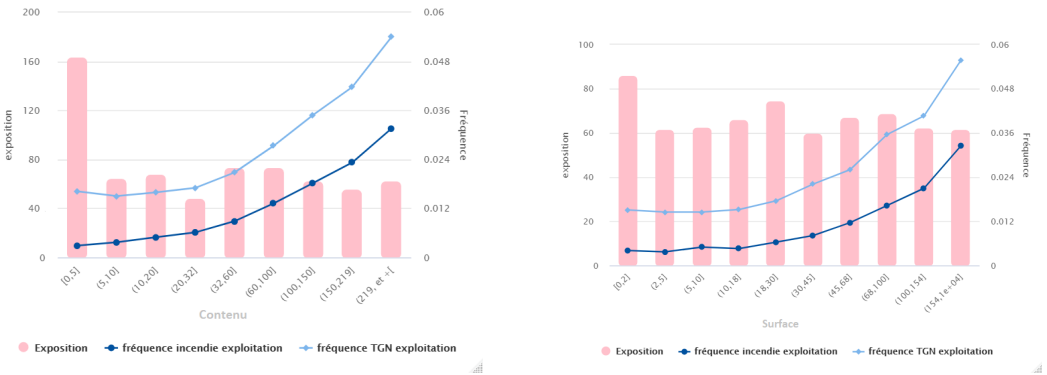
- **Garantie tempête grêle neige** :
  - De l'action directe du vent ou du choc d'un corps renversé ou projeté par le vent ;
  - De la chute de la grêle sur les toitures ;
  - Du poids de neige ou de la glace accumulée sur les toitures ;
  - Des effets du gel sur les canalisations et appareils de chauffage situés à l'intérieur des bâtiments d'habitation et de bureaux ;
  - Des intempéries (pluie, neige, grêle) qui pénètrent à l'intérieur d'un bâtiment endommagé pendant les 72 heures qui suivent l'heure du dommage.

L'analyse descriptive des données est nécessaire pour comprendre le risque à modéliser. Cette analyse permet d'identifier les différents liens entre les variables de la base de modélisation. De plus, ces statistiques descriptives serviront à vérifier la cohérence du modèle construit.

Pour la réalisation de l'étude, les variables quantitatives ont été discrétisées. L'intérêt de la discrétisation est d'avoir une prime pure segmentée par profil d'assuré. Par ailleurs, la discrétisation des variables quantitatives favorise la prise en compte des effets non linéaires entre les variables dans la modélisation.
Dans ce mémoire, les variables quantitatives ont été discrétisées en quantile. Cette discrétisation permet d'avoir assez d'exposition pour chacune des classes crées.

### 2.2.1 Analyse univariée

- **Exploitation**




(a) Fréquence de sinistres selon le contenu de l'exploitation en milliers d'€

(b) Fréquence de sinistres selon la surface de l'exploitation en ha

FIGURE 2.6 – Fréquence de sinistres exploitation : contenu et surface

Les graphiques 2.6 présentent la fréquence de sinistres relatifs au risque exploitation en

fonction du contenu de l'exploitation et de la surface totale de l'exploitation. L'analyse de ces graphiques permet de mettre en évidence la croissance de la fréquence de sinistres avec la surface et le contenu de l'exploitation. Cette tendance observée est la même quelque soit la garantie (l'incendie est représentée en bleu foncé et la tempête-grêle-neige en bleu clair).
En théorie un contenu important correspond à une grande surface. Cependant, cela n'est pas toujours le cas, il existe de petites exploitations avec de grands contenus et vice-versa. Cette remarque peut s'observer à partir du graphe 2.7, qui représente la distribution de chaque groupe de surface en fonction du contenu en pourcentage.

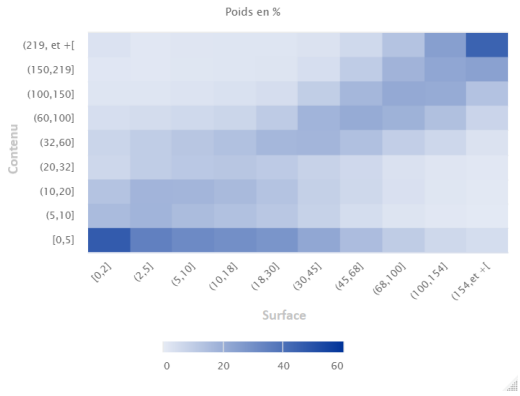
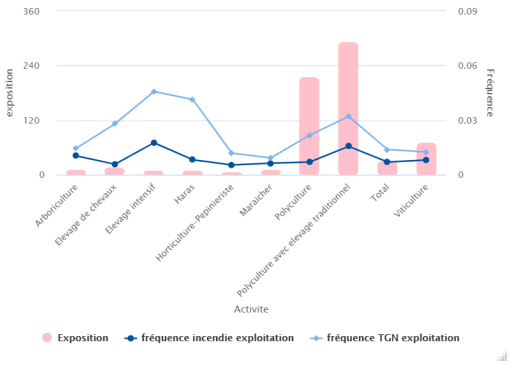
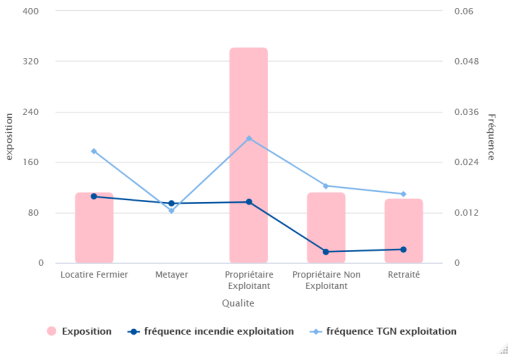



FIGURE 2.7 – Répartition des images de risque selon le contenu et la surface de l'exploitation




(a) Fréquence de sinistres selon l'activité




(b) Fréquence de sinistres selon la qualité de l'exploitant

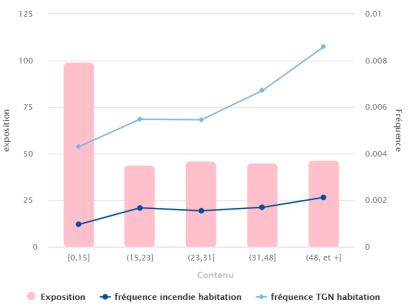
FIGURE 2.8 – Fréquence de sinistres exploitation : activité et qualité sur l'exploitation

Au regard des graphiques 2.8, l'élevage est l'une des activités avec le plus de risque, quelle que soit la garantie. Toutefois, l'élevage intensif semble présenter plus de risque que l'élevage traditionnel.

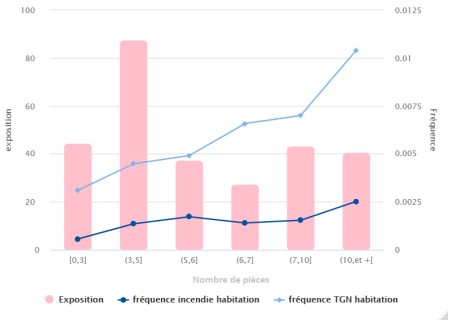
Concernant la qualité de l'exploitant, même si la fréquence est plus élevée pour les locataires, la variation globale de la fréquence de sinistres est beaucoup moins importante.

— **Habitation**

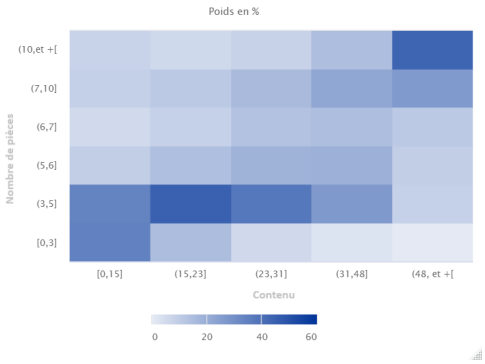
Après avoir analysé la fréquence de sinistres sur l'exploitation, l'étude de la fréquence des sinistres est réalisée sur le risque habitation.



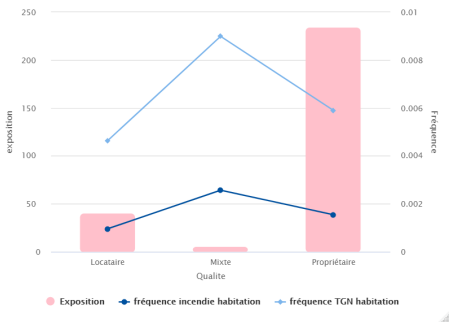

(a) Fréquence de sinistres selon le contenu de l'habitation en milliers d'€




(b) Fréquence de sinistres selon le nombre de pièces




(c) Répartition des images de risque en fonction du contenu et du nombre de pièces




(d) Fréquence de sinistres selon la qualité sur l'habitation

FIGURE 2.9 – Fréquence de sinistres habitation : contenu, nombre de pièces, qualité

Les graphiques 2.9 permettent d'observer une croissance de la fréquence de sinistres avec le contenu de l'habitation et le nombre de pièces. Comme pour l'exploitation, la croissance du contenu de l'habitation en fonction du nombre de pièces n'est pas toujours observée. L'analyse de la fréquence par rapport à la qualité de l'assuré sur l'habitation montre que les assurés qui sont à la fois propriétaires et locataires (qualité mixte) ont une fréquence de sinistres plus importante.

### 2.2.2 Étude des corrélations

L'étude des corrélations est une étape essentielle lors de la construction d'un modèle linéaire généralisé. En effet, la présence de variables corrélées peut biaiser l'effet individuel d'une variable explicative sur la variable réponse. Lors de l'interprétation du modèle, il est intéressant de visualiser uniquement l'impact d'une variable sur la prédiction sans modifier la valeur des autres variables explicatives. Cette analyse devient difficile en présence de variables corrélées. Par ailleurs, l'estimation des coefficients des variables corrélées est moins précise, car le pouvoir de prédiction individuel de ces variables est réduit.

Une métrique pour étudier les corrélations entre les variables qualitatives est le $V$ de Cramer. Il permet de mesurer l'indépendance entre deux variables qualitatives. Le $V$ de cramer consiste à faire le rapport entre la statistique de test du $\chi^2$ mesurée sur les données et sa valeur maximale théorique $\chi^2_{max} = n \times (min(l, c) - 1)$, avec $n$ le nombre d'observations, $l$ et $c$ respectivement le nombre de lignes et de colonnes du tableau de contingence des deux variables d'étude.

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

Plus le $V$ de Cramer est proche de 0, plus l'indépendance entre les variables est élevée. En revanche, plus il est proche de 1, plus la dépendance entre les deux variables est élevée. En pratique, les experts considèrent qu'il est préférable de garder dans la modélisation que les couples de variables avec un $V$ de Cramer inférieur ou égal à $0,4$. Ces variables sont ainsi considérées comme non corrélées. Toutefois, ce choix se fait en tenant compte de l'information apportée par chaque variable.

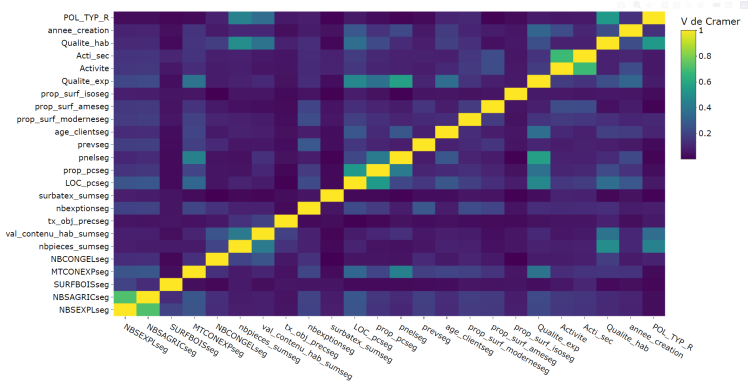



FIGURE 2.10 – Corrélations : $V$ de Cramer

A partir de ces corrélations 2.10, il a été décidé de conserver la variable « *surface agricole utile* » (NBSAGRICseg) et de retirer la variable « *surface totale de l'exploitation* » (NBSEXPLseg) ; étant donné que la surface agricole utile représente généralement plus de 90% de la surface de l'exploitation.

Le choix entre la variable « *qualité d'assuré* » (Qualite_exp) et les variables « *pourcentage de surface en fonction de la qualité* » (LOC_pcseg, prop_pcseg et pnelseg) sera effectué en fonction des modèles à construire.
Bien qu'une corrélation significative soit observée entre l'activité de l'exploitant et son activité secondaire ; cette dernière apporte une précision sur le niveau de risque couvert (précisément en élevage). La variable « *activité secondaire* » (Acti_sec) est conservée pour améliorer la segmentation du risque.

## 2.3 Les données en *Open Data*

Dans le but d'améliorer la qualité de segmentation du modèle obtenu à partir des données internes, un regard a été porté sur les données en *Open Data*. Dans ce mémoire, l'approche tarifaire proposée est une modélisation par garantie. La modélisation de deux garanties sera étudiée : incendie et tempête-grêle-neige. Les données *Open Data* qui seront utilisées pour la modélisation de ces garanties sont exposées dans la suite de cette section.

### 2.3.1 Base des centres de secours incendie

Le but en utilisant cette base est de calculer la distance entre l'adresse du risque et le point de secours incendie le plus proche. Cette variable géographique pourrait avoir une influence sur le coût des sinistres incendie et permettrait d'avoir une segmentation plus fine en termes de sévérité de la sinistralité incendie.

Le calcul de la distance au centre de secours incendie le plus proche se fait en deux étapes. La première étape consiste à construire une base géolocalisée des centres de secours incendie. Pour réaliser cette première étape de géolocalisation, la définition de la maille de recherche optimale en termes de temps de calcul et de nombre de résultats est capitale. La plupart des endroits dans le monde sont référencés par une *box*. Une *box* est une délimitation d'une zone avec des longitudes et latitudes minimales et maximales. A partir de cette *box*, il est possible de vérifier l'existence d'un centre d'intérêt (habitation, bâtiment administratif, stade, etc.) dans un endroit et d'avoir les coordonnées géographiques du centre d'intérêt. Il est possible de récupérer les coordonnées d'une *box* avec le nom de la zone concernée en précisant le pays.
En pratique, les coordonnées de la *box* sont obtenues avec le package *osmData* de R, plus précisément la fonction *getbb* avec comme paramètre principal le nom de l'endroit recherché.Un exemple de géocodage pour la ville de Brest est montré sur la figure 2.11.

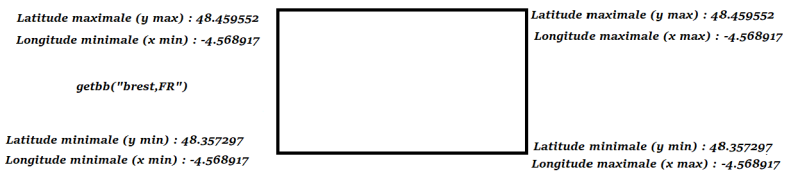



FIGURE 2.11 – Application de la fonction *getbb* sur la Brest

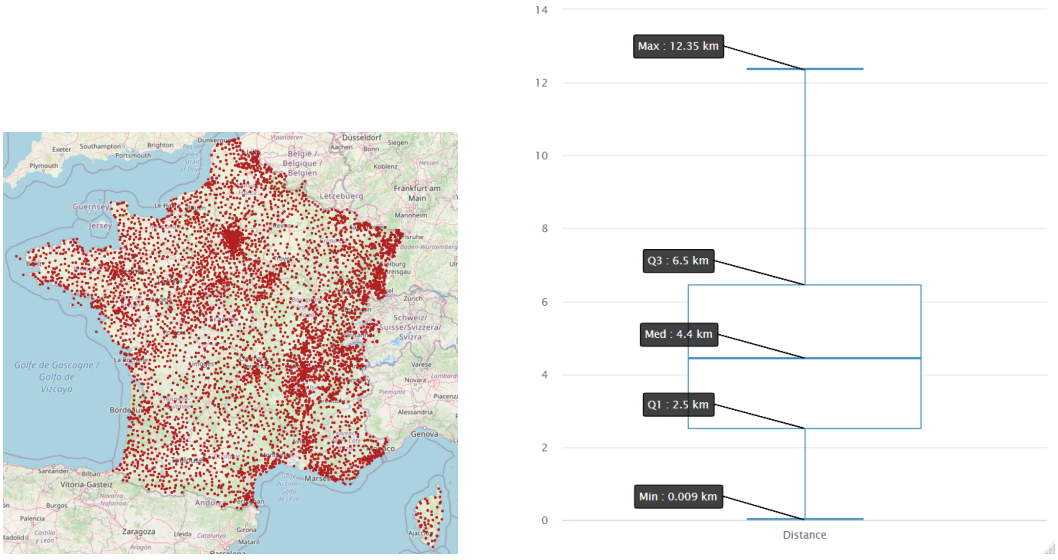
Les coordonnées des centres d'intérêt dans chaque *box* sont obtenues en utilisant la fonction *geo_ amenity* du package *nominatimlite* de R. Cette fonction [3] interroge le site internet *openstreet map* pour récupérer les centres d'intérêt dans la *box*. Le type de centre d'intérêt est défini avec le paramètre *amenity* ; pour les centres incendies *amenity= "fire_ station"*. Ensuite, en rédigeant un algorithme qui inclut cette fonction et la liste des communes de France, une base géolocalisée des centres incendies et de secours est obtenue. La base obtenue contient 10117 centres incendie et de secours en 2022.

L'étape suivante consiste à utiliser la base obtenue pour calculer la distance entre chaque risque et le centre d'incendie le plus proche. Pour le calcul, chaque risque est géolocalisé en coordonnées (x,y) Lambert 93 [4] de même que chaque centre de secours incendie de la base *Open Data*. Enfin, l'application de la fonction *pointDistance* du package *raster* combinée à d'autres fonctions permettent d'avoir la distance minimale entre l'adresse du risque et celle du centre incendie.

> ***Focus sur la projection Lambert 93***
> Pour mieux appréhender la distance calculée, il est nécessaire de définir le système de coordonnées (référentiel) utilisé. Pour cette projection, la terre est modélisée sous forme d'un ellipsoïde, en occurrence sous le système de coordonnées RGF93 [5]. Dans ce système, l'expression des coordonnées est tridimensionnelle : longitude, latitude et hauteur ellipsoïdale. Par ailleurs, la projection cartographique qui minimise les altérations de projection (linéaire et surfacique) dans le système RGF93 est la projection conique Lambert 93. La combinaison du système RGF93 et de la projection Lambert 93 (bidimensionnelle) fournit des coordonnées en $(x, y)$ en mètre pour tous les points de la France ; ce qui autorise l'utilisation d'une distance euclidienne [6] dans le cadre d'une projection plane. La fonction *pointDistance* peut être paramétrée pour préciser qu'il s'agit d'une projection plane.

3. Le nombre de résultats est limité à 50 par requête.
4. Projection officielle de la carte de France
5. Réseau Géodésique Français
6. Définition de la distance euclidienne :

$$Distance(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$




(a) Centres d'incendie et de secours en France

(b) Distance du risque au centre de secours

FIGURE 2.12 – Répartition des centres d'incendie et de secours

Les graphiques 2.12 permettent d'observer la distribution des centres d'incendie et de secours en France Métropolitaine. De plus, trois quart des assurés de la base de modélisation sont situés à mois de 6,5 kilomètres d'un centre d'incendie et de secours.

### 2.3.2 Base de données météorologiques NOAA

La NOAA (National Oceanic and Atmospheric Administration) est une agence américaine responsable de l'étude de l'océan et de l'atmosphère. Cette agence met à disposition en *Open Data* des données météorologiques récoltées par leurs différentes stations présentes dans le monde. Dans le cadre de ce mémoire, un filtre a été effectué sur les stations disponibles en France et dont les données sont disponibles entre 2009 et 2022.

La construction de cette base de données météorologiques s'est faite principalement grâce au package *rnoaa*, qui permet de déterminer les stations disponibles en France en fonction de la date de disponibilité des données. Par la suite, les données de chaque station sont téléchargées [7] puis lues avec la fonction *isd_read* de ce package.
La base constituée recense l'ensemble des données journalières de 177 stations en France entre 2009 et 2020. Les variables météorologiques dans cette base sont :

— La vitesse du vent en mètre par seconde (en m/s) ;
— La hauteur des nuages en mètre (en m) ;

7. Un algorithme a été écrit pour automatiser le téléchargement.

— La température en degrés celsius (en °C) ;
— La température de point de rosée en degrés celsius (en °C) ;
— La pression atmosphérique : correspond à la pression de l'air au niveau de la mer en hectopascal (en hPa) ;
— L'humidité relative : rapport sans unité entre la quantité d'humidité atmosphérique présente par rapport à la quantité qui serait présente si l'air était saturé ;
— La précipitation en millimètre (en mm).

A partir de chacune des variables initiales, de nouvelles variables sont obtenues en calculant le maximum, le minimum, la moyenne, la médiane, les quantiles 5%, 75% et 95% des données journalières. Une analyse des corrélations et de pertinence de ces variables sera faite avant leur intégration dans la modélisation.

***Interpolation des données***

Ne disposant pas de l'information météorologique à la maille adresse, ni à la maille Insee, mais plutôt à une maille station, une alternative est d'effectuer une attribution pondérée de chaque variable par zone Insee.
Soit une commune $c_i$, et soit $s_i$, $i \in [|1, n|]$, une station météorologique de France, pour laquelle la variable $m_i$ est mesurée. On note $d_i$ la distance entre la commune $c_i$ et la station $s_i$.
Soit $f$ une fonction à valeurs réelles telle que $\lim_{x \to +\infty} f(x) = 0$. Soit $M_i^c$ la valeur de la variable composite pour la commune $c_i$, alors

$$M_i^c = \frac{\sum_{i=1}^{n} m_i f(d_i)}{\sum_{i=1}^{n} f(d_i)}$$

L'objectif est de pénaliser les stations éloignées et d'attribuer une information agrégée pour l'ensemble des stations proches d'une commune. Ainsi, le risque d'un mauvais mixte de données météorologiques est minimisé par commune avec cette pondération par distance. Pour le choix de la fonction $f$, un exemple classique est l'ensemble des fonctions du type $x^{-p}$ avec $p > 0$ et $x \neq 0$ (p=3 dans ce mémoire).
La carte 2.13 permet d'observer la distribution des stations météorologiques.

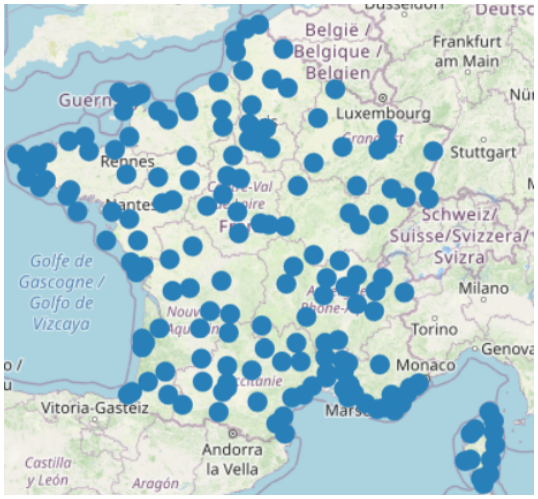



FIGURE 2.13 – Répartition des stations météorologiques

### 2.3.3 Base de criminalité ONDRP

L'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP) diffuse des données sur les crimes et délits enregistrés par les services de police et les unités de la gendarmerie nationale. Les données sont disponibles pour l'ensemble des départements en France depuis 2000.
L'utilisation de cette base de données permettrait d'avoir un modèle qui tient compte de l'information sur la criminalité environnante. Conformément aux risques couverts par le contrat MRA et des garanties à modéliser, il parait judicieux d'effectuer un filtre sur l'ensemble des crimes enregistrés par l'ONDRP. Les crimes retenus sont les suivants :

| |
|---|
| Vols simples sur exploitations agricoles |
| Incendies volontaires de biens privés |

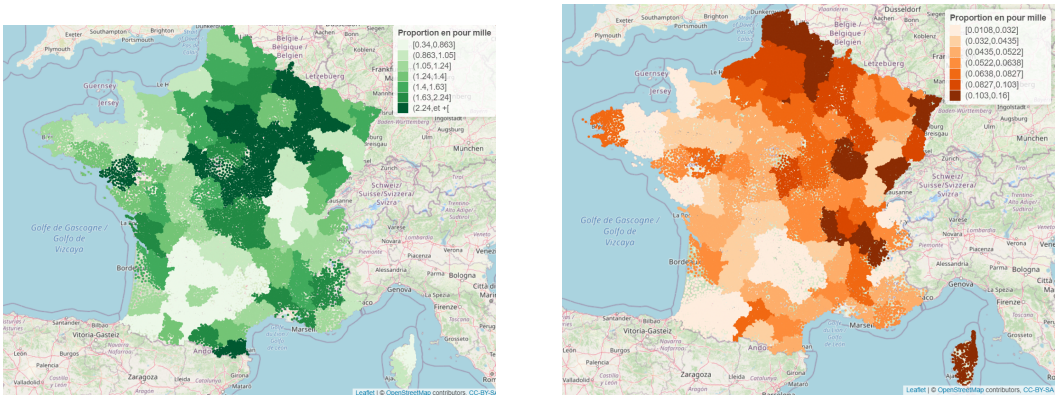
TABLE 2.1 – Crimes sélectionnés

Du fait de l'hétérogénéité du nombre de bâtiments et d'exploitations par département, il paraît plus pertinent de normaliser le nombre de crimes. Pour la variable « vols simples sur exploitations agricoles », la variable utilisée pour normaliser est le nombre d'exploitations agricoles par département. Concernant la variable « incendies volontaires de biens privés», la variable utilisée pour normaliser est le nombre de résidences principales. N'ayant pas accès à la donnée de chaque année pour normaliser ces variables, il a été décidé de faire des estimations. Le nombre de résidences principales en 2018, 2013 et 2008 issu de la base INSEE du recensement 2018 a été utilisé comme proxy pour le nombre de résidences principales. Pour affiner ce proxy, un taux de croissance global entre chaque recensement a été calculé. Puis, à partir de ces taux globaux, un taux de croissance annuel est calculé pour une interpolation du nombre de résidences principales sur la période d'étude de 2010-2018 pour chaque département. Pour l'estimation des années 2019, 2020

et 2021 (2020 et 2021 affectés par le COVID), les taux de croissance du nombre de résidences principales nationaux ont été utilisés.
Concernant le nombre d'exploitations, la donnée disponible est le nombre d'exploitations par commune issue du recensement agricole de 2010. A partir de cette donnée et du nombre total d'exploitations agricoles en 2010 et 2020, une estimation du nombre d'exploitations pour chaque année a été effectuée. La méthode d'estimation est la même que pour le nombre de bâtiments décrite ci-dessus.
Le tableau 2.2 présente les taux de croissance annuels utilisés pour l'estimation du nombre de logements et d'exploitations.

| Taux de croissance annuels | | | | | |
|---|---|---|---|---|---|
| Logement | | | | | Exploitation |
| 2008-2013 | 2013-2018 | 2019 | 2020 | 2021 | 2010-2020 |
| 0,81% | 0,86% | 1% | 0,9% | 0,7% | -2,3% |

TABLE 2.2 – Taux de croissance annuels




(a) Proportion de vols sur exploitations agricoles pour mille exploitations agricoles (2021)

(b) Proportion d'incendies volentaires de biens privés pour mille logements (2021)

FIGURE 2.14 – Répartition des crimes par département

Les deux cartes 2.14 représentent la distribution de la proportion des deux crimes normalisée par département sur l'ensemble du territoire, et montrent que les départements du nord et le centre de la France sont plus impactés par la criminalité que les départements du sud-ouest.
L'intégration de ces variables dans la base de modélisation s'effectuent en utilisant comme clé de jointure le département et l'année de survenance des crimes.

### 2.3.4 Base Valeur foncière DVF

La base de Demande de Valeur Foncière (DVF) géolocalisée est une base qui recense les informations sur les transactions immobilières en France. Cette base est obtenue à partir des actes notariés et des informations cadastrales. Cette base servira au calcul du prix au mètre carré des logements par code Insee.

L'intérêt au travers de cette variable est de capter un effet géographique lié à une caractéristique socio-économique. La distribution de cette variable en fonction de la zone Insee est présentée sur la carte 2.15. L'Ile de France, le sud-est et le littoral sont les zones où les logements valent le plus cher.

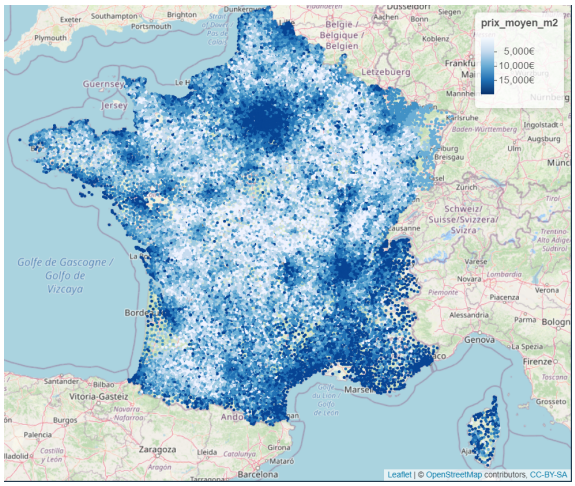



FIGURE 2.15 – Prix au mètre carré des logements

Pour les zones Insee avec des valeurs manquantes, une imputation par la moyenne pondérée des prix au mètre carré des zones Insee les plus proches en termes de distance a été faite.

### 2.3.5 Base des années de construction des logements

La base de données contenant les années de construction des logements est issue des tableaux détaillés des logements construits avant 2016 du recensement général de la population de 2018. Ces tableaux résument l'information sur le nombre de maisons, appartements et autres types d'habitations par période de construction pour chaque code Insee.
A partir de cette base, la proportion de logements construits a été calculée en fonction des périodes suivantes :

— Avant 1919 ;
— De 1919 à 1945 ;

— De 1946 à 1970 ;
— De 1971 à 1990 ;
— De 1991 à 2005 ;
— De 2006 à 2015.

Cette variable pourra permettre d'identifier s'il existe un risque associé aux zones avec de vieilles constructions ou des récentes.
La distribution de la proportion des maisons construites entre 2006 et 2015 est présentée sur cette carte 2.16 :

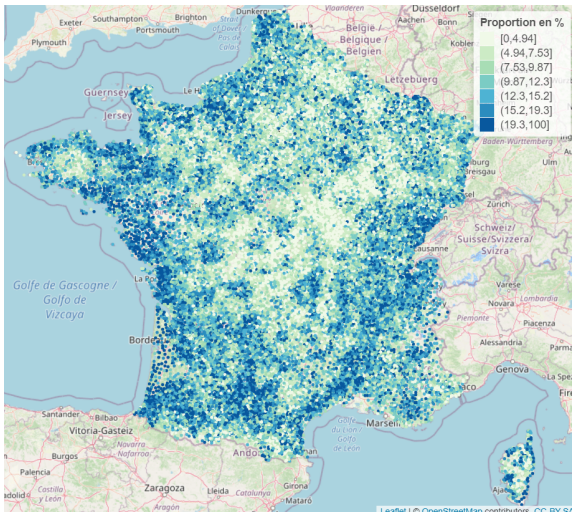



FIGURE 2.16 – Proportion de maisons construites entre 2006 et 2015

### 2.3.6 Base nationale des bâtiments (BNDB)

La Base de Données Nationale des Bâtiments (BNDB) est un projet visant à mettre à la disposition du public le maximum d'informations associées à l'ensemble des bâtiments en France métropolitaine. La BNDB est issue du croisement géospatial d'une vingtaine de bases de données d'organismes publics.

Pour sa version v0.6.2, la BNDB contient 4 tables pour 438 variables liées aux bâtiments et téléchargeables sur *Datagouv*. Les variables présentes dans cette base sont par exemple :
— l'adresse du bâtiment ;
— les coordonnées géographiques du bâtiment : longitude, latitude ;
— l'année de construction ;
— l'altitude au sol du bâtiment ;
— le nombre de logements du bâtiment.

L'inconvénient principal de cette base est la taille du fichier de données (36 giga) ce qui rend sa manipulation très difficile. La manipulation de ce fichier de données a été infruc-

tueuse, malgré le recours à une expertise interne. De plus, une étude a été réalisée par un prestataire externe sur le taux de correspondance entre les bâtiments du portefeuille AXA France et les bâtiments de la BNDB. Deux villes françaises avec des proportions élevées de bâtiments correctement géolocalisés ont été choisies. Le taux de correspondance obtenu avoisinait 75%. Ce résultat est en partie lié à la différence de générateur de coordonnées géographiques (géocodeur) entre la BNDB et le portefeuille AXA France. Par ailleurs, pour ces bâtiments, il n'y avait pas une totale concordance entre les variables présentes dans le portefeuille AXA France et celles renseignées dans la BNDB.
C'est pourquoi cette piste a été abandonnée par souci de cohérence d'informations.

**N.B** : Dans le cadre d'une modélisation d'une prime pure affaire nouvelle, la jointure des données météorologiques et des données sur la criminalité à la base de modélisation s'effectuera en utilisant les données de l'année $n-1$. En effet, lors de la souscription du contrat d'assurance, l'information sur la criminalité, et sur la météo globale de l'année en cours est inconnue. Pour éviter d'avoir un modèle biaisé et non opérationnel la jointure a été faite avec les données de 2009 à 2020 pour correspondre à la période d'étude 2010-2021.

**A retenir** :
- **Période d'étude** : 2010 - 2021 ;
- **Base de modélisation** : base par images des contrats ;
- **Mise en *as-if*** : indice FFB ;
- **Garantie à modéliser** : incendie (38% de la prime pure observée) et tempête grêle neige (20% de la prime pure observée) ;
- **Tendances principales sur l'exploitation** : croissance de la fréquence de sinistres en fonction de la valeur du contenu de l'exploitation et de la surface ;
- **Tendances principales de l'habitation** : croissance de la fréquence de sinistres en fonction du nombre de pièces et de la valeur du contenu de l'habitation ;
- **Données *Open data*** : distance au centre de secours incendie, données météorologiques, données sur la criminalité, prix moyen au mètre carré, données sur l'année de construction des logements.

# Chapitre 3

# Préparation de la base de données

Les résultats d'un modèle sont généralement influencés par la qualité des données utilisées. Il est dès lors nécessaire de traiter les données afin de détecter les éventuelles anomalies ou transformations à effectuer.

## 3.1 Traitement des valeurs extrêmes

### 3.1.1 Contexte

Comme l'indique le titre, cette section aura pour but d'étudier les valeurs extrêmes des charges des sinistres. En général, les sinistres sont séparés en deux catégories : les attritionnels et les graves. Un sinistre est considéré comme grave du fait de sa rareté et du coût de l'indemnisation très élevé qui lui est associé ; tandis qu'un sinistre attritionnel est plutôt caractérisé par sa fréquence assez élevée et un montant d'indemnisation « usuel ». Les notions d' « usuel » et de « très élevé » évoquées dans la phrase précédente prennent leurs sens au travers de la théorie des valeurs extrêmes qui permet de déterminer la frontière (le seuil) entre ces deux notions.

### 3.1.2 Théorie mathématique du choix du seuil

Actuellement, le seuil à partir duquel un sinistre est considéré comme grave est commun à toutes les garanties du produit MRA. Ce seuil de 150 000 € peut être affiné afin d'être plus adapté à chaque garantie. Par exemple, la charge maximale observée sur la garantie vol sur la période d'étude est inférieure à 110 00 euros ; d'où l'intérêt d'avoir un seuil plus juste pour une meilleure modélisation de la charge de sinistre. Dans le cadre de ce mémoire, un focus sera fait sur trois méthodes de détermination du seuil : la fonction d'excès moyen, l'estimateur de Hill et l'estimateur par la méthode de Gerstengarbe.

Dans un premier temps, la loi du maximum $M_n$ sera présentée. Soit $(X_n)_{n\geq 1}$, une suite de $n$ variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de même loi de probabilité et $F$ la fonction de répartition telle que $F(x) = P(X \leq x)$ avec

$M_n = \max(X_1, X_2, X_3, .., X_n)$. Alors :

$$\begin{aligned} P(M_n \leq x) &= P(max(X_1, X_2, X_3, .., X_n) \leq x) \\ &= P(\bigcap_{i=1}^{n} X_i \leq x) \\ &= F(x)^n \end{aligned}$$

L'inconvénient de ce résultat est que la fonction $F$ est généralement inconnue. Toutefois, l'objectif étant de déterminer une loi asymptotique de $F$, soit $x^F$ le point extrêmal de $F$ défini par : $x^F = sup\{x \in \mathbb{R}, F(x) < 1\}$. Alors $M_n \underset{n\to\infty}{\overset{\mathcal{P}}{\to}} x^F$ et sa distribution asymptotique est dégénérée. Une alternative a été proposée avec le théorème de Fisher-Tippet pour avoir une distribution non dégénérée [1].

**Théorème** : Fisher-Tippet
Soit $(X_n)_{n\geq 1}$ une suite de $n$ variables aléatoires i.i.d. et de fonction de répartition $F$ définie par $F(x) = P(X \leq x)$.
S'il existe deux suites de réels $(a_n \in R_+^*; b_n \in \mathbb{R}; n \geq 1)$ et une distribution non dégénérée $G$ telles que :

$$\begin{aligned} \lim_{n\to+\infty} \mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \lim_{n\to+\infty} F^n(a_n x + b_n) \\ &= G(x), \qquad \forall x \in \mathbb{R}, \end{aligned}$$

alors $\exists \mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ tels que $G$ soit une *GEV (distribution des extrêmes généralisée)*.
Soit $G_{\mu,\sigma,\xi}$ la fonction de répartition d'une $GEV$ :

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} exp\left(-\left(1 + \xi(\frac{x-\mu}{\sigma})\right)^{\frac{-1}{\xi}}\right), & si\ \xi \neq 0 \\ exp\left(-exp(-\frac{x-\mu}{\sigma})\right), & si\ \xi = 0 \end{cases}$$

Avec $\xi$ le paramètre de forme, $\mu$ le paramètre de position, et $\sigma$ le paramètre d'échelle. Plus $\xi$ est grand, plus le poids des extrêmes dans la distribution est important. Ce paramètre $\xi$ est l'une des variables clés dans la notion de domaine d'attraction par rapport à l'une des trois lois décrites ci-dessous :

Domaine d'attraction $(\xi = 0)$ : queue à décroissance exponentielle

$$G_{\mu,\sigma}(x) = exp\left(-exp(-\frac{x-\mu}{\sigma})\right), \qquad x \in \mathbb{R}$$

1. Loi non dégénérée : la variance de la loi est non nulle

Domaine d'attraction de Frechet ($\xi > 0$) : queue de distribution épaisse et à décroissance lente

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} exp\left(-\left(1+\xi(\frac{x-\mu}{\sigma})\right)^{\frac{-1}{\xi}}\right), & x > \mu \\ 0, & x \leq \mu \end{cases}$$

Domaine d'attraction de Weibull ($\xi < 0$) : queue de distribution bornée

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} exp\left(-\left(1-|\xi|(\frac{x-\mu}{\sigma})\right)^{\frac{1}{|\xi|}}\right), & x > \mu \\ 1, & x \leq \mu \end{cases}$$

***Fonction d'excès moyen***

Soit $X$ une variable aléatoire de fonction de répartition $F$. On appelle fonction d'excès moyen de seuil $u$, la fonction $e(u)$ définie par :

$$e(u) = E[X - u | X > u]$$

En particulier :
Si $Y = X - u^* | X > u^* \sim G_{0,\sigma_{u^*},\xi}$ alors la fonction $e(u)$ est linéaire en $u$ pour $u > u^*$ et $\xi < 1$.

$$e(u) = E[X - u | X > u] = \frac{\sigma_{u^*} + \xi u}{1 - \xi}$$

*Estimateur empirique*
Soit $(X_1, ..., X_n)$ des variables aléatoires i.i.d., l'estimateur empirique de $e(u)$ est donné par :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n}(I_{X_i > u})} = \frac{\sum_{i=1}^{n}(X_i - u)^+}{N_u}$$

Avec $N_u$ le nombre d'observations dépassant le seuil $u$.
En pratique le seuil $u^*$ se détermine graphiquement. La fonction $e(u)$ étant linéaire pour $u > u^*$, le but est de rechercher le seuil à partir duquel le graphique $(u, e(u))$ devient linéaire.

***Estimateur de Hill***

Cet estimateur est généralement utilisé dans le cas où $\xi > 0$. Il n'est utilisable que pour le domaine de Fréchet et assure un bon équilibre biais-variance. Soient $(X_1, ..., X_n)$ des

variables aléatoires i.i.d. et $(X_{1,n}, ..., X_{n,n})$ la statistique d'ordre associée. L'estimateur est défini de la manière suivante :

$$\overset{\wedge Hill}{\xi}_{k,n} = \frac{1}{k}\sum_{j=1}^{k} log(X_{n-j+1,n}) - log(X_{n-k,n})$$

C'est un estimateur consistant et asymptotiquement normal mais très sensible aux valeurs de $k$. Le graphique de l'estimateur de Hill représente la valeur de l'estimateur en fonction de l'indice $k$ de la statistique d'ordre. L'indice $k$ servira à déterminer l'ordre à partir duquel se forment les extrêmes. Cet ordre correspond au plus petit indice du plateau (zone de stabilité de l'estimateur), où l'estimateur semble robuste. La valeur correspondante à cet ordre peut être choisie comme le seuil $u$.

***Estimation par la méthode de Gerstengarbe***

La méthode de Gerstengarbe, développée par Gerstengarbe et Werner, est basée sur le test statistique non paramétrique de Mann-Kendall. Ce test vise à déterminer, pour une série de données, le point de changement brusque de la tendance. Ce point de changement est considéré comme point de départ de la région des sinistres extrêmes. En pratique, ce changement s'observe lorsqu'il y a une modification de la tendance des écarts consécutifs des sinistres.

Soit $(x_i)_{i\in[|1,N|]}$ la série ordonnée de la charge de sinistres tel que $x_1 < x_2 < ... < x_N$. L'écart de sinistres est défini par $\Delta_i = x_i - x_{i-1}$, pour $i \in [|2, N|]$ avec $N$ le nombre de sinistres. La cible dans cette méthode est le point $x_i^*$, à partir duquel est constatée une modification de la tendance des $\Delta_i$ liée au passage de la zone des sinistres attritionnels à la zone des sinistres extrêmes.
Le point $x_i^*$ est le point d'intersection des deux séries statistiques suivantes :

$$U_i = \frac{\sum_{k=1}^{i} n_k - \frac{i(i-1)}{4}}{\frac{\sqrt{i(i+1)(2i+5)}}{72}} \qquad \tilde{U_i} = \frac{\sum_{k=1}^{i} \tilde{n_k} - \frac{i(i-1)}{4}}{\frac{\sqrt{i(i+1)(2i+5)}}{72}}$$

$n_k = \sum_{j=1}^{k} 1_{(\Delta_j < \Delta_k)}$ et $\tilde{n_k} = \sum_{j=1}^{k} 1_{(\Delta_{n-j} < \Delta_{n-k})}$.

### 3.1.3 Application à nos données

Dans ce mémoire, l'application de la théorie des valeurs extrêmes servira à déterminer le seuil à partir duquel se forment les extrêmes pour écrêter et mutualiser la charge de sinistres. Bien que cette méthode soit moins robuste, elle est la plus adaptée à notre contexte du fait du faible volume de sinistres [2].

2. Moins de 7000 sinistres par année

L'écrêtement des sinistres est une méthode qui consiste à déduire de la charge de chaque sinistre grave la part qui est supérieure au seuil des extrêmes.

$$Charge\ écrêtée = \begin{cases} Charge\ du\ sinistre & \text{si } Charge\ du\ sinistre \leq seuil \\ Charge\ du\ sinistre - seuil & \text{sinon} \end{cases}$$

Le reste des montants des sinistres, appelé la sur-crête, est mutualisé sur l'ensemble des sinistres. Cette répartition s'effectue en calculant la valeur de mutualisation $VM$ :

$$VM = \frac{\sum_{j=1}^{N_g} C_j - N_g * seuil_{grave}}{N}$$

Avec $N_g$ le nombre de sinistres graves , $N$ le nombre de sinistres, $C_j$ le coût du j-ième sinistre grave.
Finalement, la charge mutualisée est la suivante :

$$Charge\ mutualisée = \begin{cases} Charge\ écrêtée + VM & \text{si } Charge\ écrêtée < seuil \\ seuil + VM & \text{sinon} \end{cases}$$

Cette méthode est la moins robuste mais plus appropriée à une modélisation de la charge des sinistres dans un contexte de faiblesse de volume de sinistres extrêmes.

### La garantie incendie

La garantie incendie est subdivisée en deux sous garanties : incendie habitation et incendie exploitation. Une comparaison de la distribution de la charge de sinistres de ces garanties est faite à partir d'un graphique quantile-quantile (*QQ-plot*) en figure 3.1.

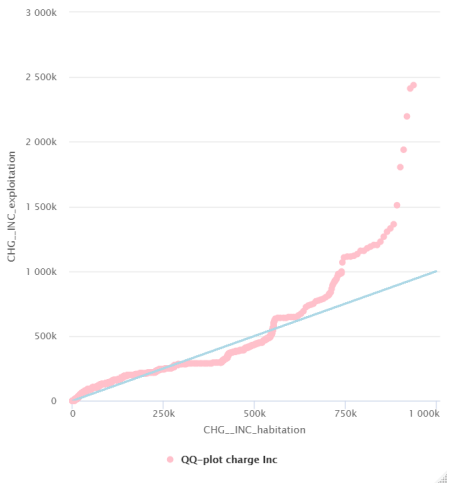
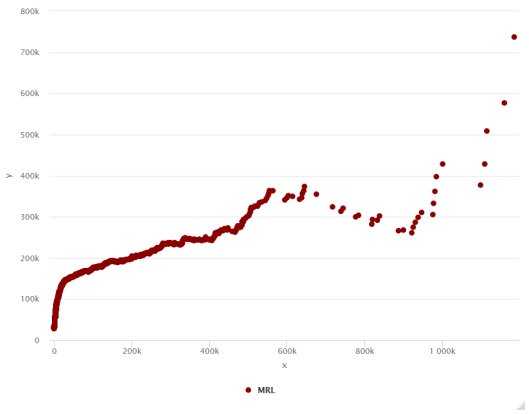


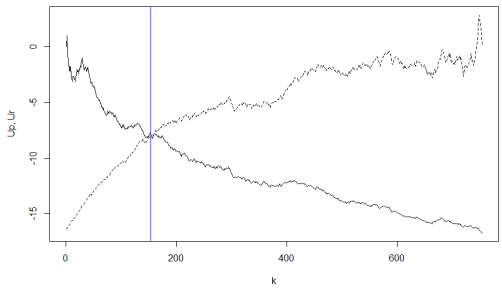

FIGURE 3.1 – *QQ-plot* charge incendie exploitation - habitation

L'analyse du graphique 3.1 montre dans un premier temps que la distribution de la charge attritionnelle est très proche entre les deux risques. Ensuite, ce *QQ-plot* permet de remarquer que la charge incendie exploitation présente plus de sinistres atypiques qu'en risque habitation. Pour étayer l'hypothèse de proximité des distributions, l'hypothèse $H_0$ du test de Kolmogorov-Smirnov a été acceptée au seuil de 5%. Cela signifie que, pour un risque de se tromper de 5%, les lois de probabilité des charges de sinistres habitation et exploitation auraient la même fonction de répartition.
Outre cet aspect de test, il est préférable de choisir un seuil des extrêmes commun pour l'incendie habitation et exploitation du fait de la faible volumétrie des sinistres incendie habitation, au risque d'avoir un seuil biaisé.
Néanmoins, la variable « type de risque » devra être intégrée dans la modélisation pour un ajustement de la charge. Cette variable permettra de tenir compte du fait qu'il y a des contrats MRA sans risque habitation.




(a) Fonction d'excès moyen


(b) Estimateur Hill




(c) Estimateur de Gerstengarbe

FIGURE 3.2 – Détermination du seuil incendie

L'application des méthodes théoriques décrites plus haut permet de déterminer un seuil de sinistres extrêmes théorique, qu'il faudra confronter aux contraintes business. Dans

un premier temps, l'analyse des résultats théoriques s'avère nécessaire pour prendre une décision commerciale avec le plus de recul possible.

La fonction d'excès moyen 3.2-a propose trois seuils : un premier à 30 558€, un deuxième 644 860€ et un dernier 921 931€. En effet, en ces points la fonction d'excès moyen présente des tendances linéaires et maximise localement l'approximation par une droite de la fonction d'excès moyen en tenant compte du critère du $R^2$.

| Seuil | 30 558€ | 644 860€ | 921 931€ |
|---|---|---|---|
| Poids des extrêmes dans la charge | 88,09% | 14,40% | 9,70% |
| Proportion des extrêmes | 14,28% | 0,41% | 0,24% |
| $R^2$ ajustement linéaire local | 0,98 | 0,92 | 0,81 |

TABLE 3.1 – Synthèse des résultats - Fonction d'excès moyen - Garantie incendie

En outre, le graphique de Hill 3.2-b ne présente pas de stabilité, même en augmentant le nombre de statistique d'ordre à calculer. Cette tendance ne permet pas de définir un ordre à partir duquel se forment les extrêmes.

Enfin, en appliquant la dernière méthode théorique exposée, c'est-à-dire la méthode de Gerstengarbe, le seuil qui est proposé est un seuil 310 000 € en prenant comme point de départ de détection de modification de la tendance le quantile 0,9. Ce point de départ a été choisi pour qu'il ne détecte pas une modification de la tendance parmi les sinistres attritionnels mais plutôt au niveau de la queue de distribution. Il faut noter que cette estimation dépend fortement du point de départ. En prenant un point de départ plus extrême, l'estimation n'est plus la même.

Face à tous ces résultats théoriques, se confronte une contrainte commerciale qui impose une correspondance entre le poids des extrêmes dans la charge et la proportion de sinistres extrêmes. De manière plus explicite, il faudrait que le seuil choisi définisse un contexte où la proportion maximale de sinistres extrêmes soit autour de 5% et la part dans la charge totale soit autour de 30%. Le seuil permettant de s'en approcher le plus est celui déterminé avec la méthode de Gerstengarbe, soit 310 000€, avec une proportion de 2% et un poids dans la sinistralité totale de 39%. Ce seuil a été donc retenu dans ce mémoire.

### La garantie tempête grêle neige (TGN)

De manière analogue à la garantie incendie, on effectue une première étude sur la distribution de la charge de la sinistralité entre la sinistralité habitation et la sinistralité exploitation. Cette analyse est nécessaire car l'impact des évènements climatiques dépend fortement des matériaux de construction du bâtiment.

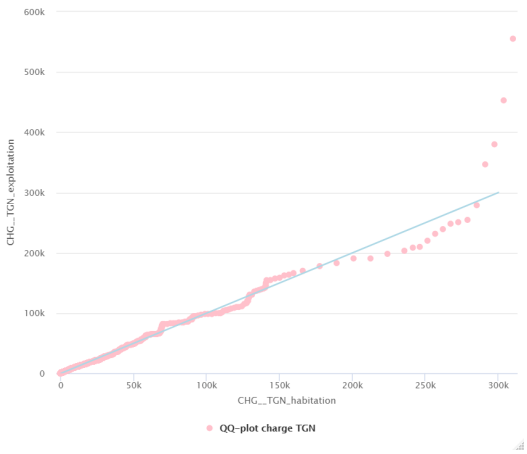



FIGURE 3.3 – *QQ-plot* charge TGN exploitation - habitation

Cependant, après un test de Kolmogorov-Smirnov et une analyse du *QQ-plot* 3.3, il n'y a pas de différence statistiquement significative entre les lois des distributions des charges des sinistralités habitation et exploitation. Ce résultat n'est pas tout à fait décevant dans la mesure où les deux risques peuvent être situés à la même adresse et peuvent donc être soumis à la même intensité de l'évènement climatique ; bien que l'impact peut être différent. En outre, les seuils obtenus en faisant une distinction entre l'habitation et l'exploitation ne sont pas significativement différents du seuil obtenu sans distinction. Les résultats sans distinction de l'habitation et l'exploitation seront présentés.

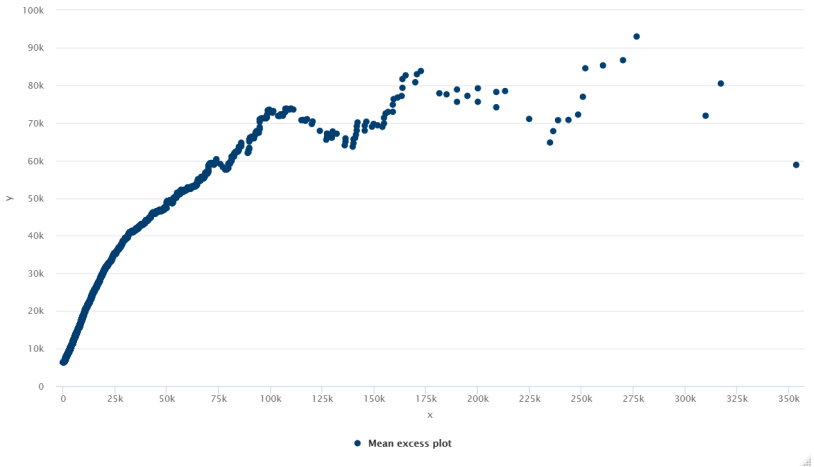
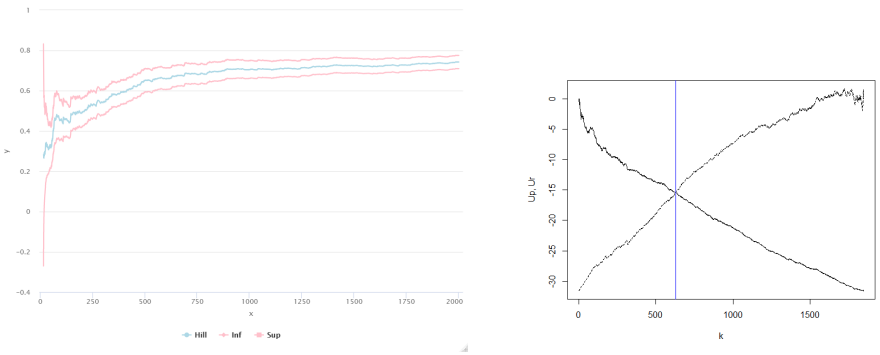



FIGURE 3.4 – Fonction d'excès moyen TGN




(a) Estimateur Hill (b) Estimateur de Gerstengarbe

FIGURE 3.5 – Détermination du seuil TGN

L'usage des différentes méthodes théoriques ont permis de choisir les seuils suivants :

— 31 790 € pour la fonction d'excès moyen (3.4) ;
— 31 690 € avec l'estimateur de Hill pour la statistique d'ordre 565 (3.5-a) ;
— 29 238 € avec l'estimation par la méthode de Gerstengarbe (3.5-b).

Finalement, le seuil de 31 790 € avec la méthode d'excès moyen a été retenu avec un poids des extrêmes dans la charge de 34% et une proportion de 3%.

## 3.2 Traitement des sinistres en cours

La survenance d'un sinistre entraîne une indemnisation de l'assuré par l'assureur si ce sinistre respecte les termes du contrat. Un dossier associé au sinistre est ouvert et une visite d'expertise est effectuée pour vérifier si l'évènement survenu est inclus dans la couverture du contrat. A la suite de cette visite d'expertise deux cas de figure se présentent à l'assuré :

— le sinistre n'est pas inclus dans les termes du contrat, celui-ci passe alors au statut « clos sans suite » ;
— le sinistre est couvert par le contrat, celui-ci passe alors au statut « en cours » puis au statut « clos avec suite » au moment de l'indemnisation totale du sinistre.

Dans le but d'avoir un modèle de prime pure par garantie avec le moins de biais possible, il est nécessaire d'avoir une estimation de la charge totale des sinistres en cours. La méthode qui sera appliquée pour l'estimation de la charge finale des sinistres est la méthode de Chain-Ladder. Il s'agit de la méthode utilisée en interne pour la projection des sinistres en cours.

### 3.2.1 Présentation de la méthode

L'estimation de la charge finale des sinistres en cours avec la méthode de Chain-Ladder passe par la construction du triangle d'écoulement de la charge de sinistres par

année de survenance et année de développement. Par souci de robustesse, il est nécessaire d'avoir un historique assez profond sur la sinistralité.

Considérons les notations suivantes :
- $i$ : l'année de survenance du sinistre ;
- $j$ : l'année de développement du sinistre ;
- $C_{ij}$ : la charge des sinistres survenus en $i$ et évalués en $j$.

Le triangle de liquidation cumulé est :

| | Année de développement | | | | | | |
|---|---|---|---|---|---|---|---|
| Année de survenance | $AD_0$ | $AD_1$ | ... | $AD_j$ | ... | $AD_{m-1}$ | $AD_m$ |
| $AS_0$ | $C_{0,0}$ | $C_{0,1}$ | ... | $C_{0,j}$ | ... | $C_{0,m-1}$ | $C_{0,m}$ |
| $AS_1$ | $C_{1,0}$ | $C_{1,1}$ | ... | $C_{1,j}$ | ... | $C_{1,m-1}$ | |
| ... | ... | ... | ... | ... | ... | | |
| $AS_i$ | $C_{i,0}$ | $C_{i,1}$ | ... | $C_{i,j}$ | | | |
| ... | ... | ... | ... | | | | |
| $AS_{n-1}$ | $C_{n-1,0}$ | $C_{n-1,1}$ | | | | | |
| $AS_n$ | $C_{n,0}$ | | | | | | |

TABLE 3.2 – Triangle de liquidation cumulé

La méthode de Chain-Ladder repose sur l'hypothèse de constance des facteurs de développement pour toutes les années de survenances, c'est-à-dire $\frac{C_{i,j+1}}{C_{i,j}} = f_j$ pour tout $i$. L'estimateur des facteurs de développement de Chain-Ladder est le suivant :

$$\hat{f}_j = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}}$$

Ainsi pour tout $j > i$ :

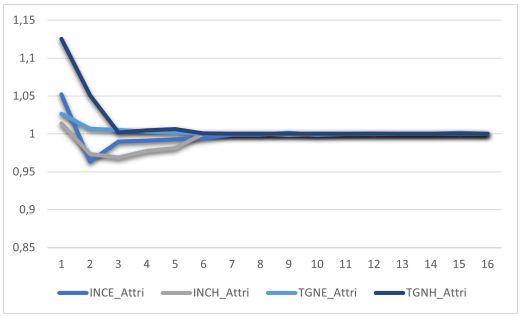
$$\hat{C}_{i,j+1} = \hat{f}_j \times C_{i,j}$$

Il est alors possible de compléter le triangle de liquidation cumulé. Notre objectif étant d'utiliser les facteurs de développement de Chain-Ladder pour calculer la charge à l'ultime des sinistres individuels.

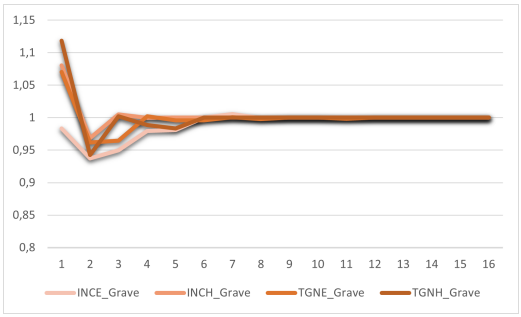
### 3.2.2 Application

La proportion de sinistres en cours dans la base de modélisation est de 5%. Ces sinistres seront projetés à l'ultime en utilisant les facteurs de Chain-Ladder définis précédemment. En pratique, une profondeur de 16 ans (2005-2021) a été utilisée pour la

construction des triangles de règlements. L'analyse de l'écoulement de la charge (graphique 3.6) permet d'observer qu'il y a une surestimation des coûts d'ouverture par rapport à la charge finale et la présence de recours éventuels. Néanmoins, l'analyse des triangles de règlements et des facteurs de développements révèlent que les sinistres TGN sont réglés au bout de 3 ans pour les sinistres attritionnels et 4 ans pour les sinistres graves. Concernant les sinistres incendie attritionnels et graves, il y a une stabilisation des facteurs au bout de 5 ans, ce qui serait liée à une charge élevée des sinistres.




(a) Facteurs Chain-Ladder - Sinistres attritionnels




(b) Facteurs Chain-Ladder - Sinistres graves

FIGURE 3.6 – Facteurs de Chain-Ladder

## 3.3 Traitement des valeurs manquantes

### 3.3.1 Théorie

L'analyse des valeurs manquantes est une étape particulière dans la création de la base de modélisation. Cette analyse participe à l'amélioration de la qualité du modèle fourni en limitant les biais dus à l'absence d'informations. Pour mieux appréhender la problématique des valeurs manquantes, la présentation des différentes types de valeurs manquantes est nécessaire. Elles sont de trois types :

— **Missing Completely At Random (MCAR) :**
Ce type correspond aux absences de données obtenues de manière totalement aléatoire. Cette absence n'est aucunement liée à une variable présente dans l'ensemble de données. Ainsi, la probabilité d'absence de données est la même pour toutes les observations et ne dépend que de paramètres externes à la donnée manquante. Le biais en ne tenant pas compte de ce type de données manquantes est négligeable si la proportion de ce type de donnée est faible dans la base ; cela au détriment d'une perte de précision occasionnelle.

— **Missing At Random (MAR) :**
Pour ce type de valeurs manquantes, l'absence d'information n'est pas complètement aléatoire mais plutôt liée à la modalité d'une ou de plusieurs autres variables de la base de modélisation. Le biais apporté par ce type de valeur manquante est non négligeable dans un contexte de sur-représentation des valeurs manquantes. Un traitement adéquat est nécessaire pour limiter ce biais dans la modélisation.

— **Missing Not At Random (MNAR)**
Une valeur manquante est de type MNAR pour une variable si la présence de cette valeur manquante est totalement causée par la variable en question. Ce type de valeurs manquantes contribue fortement à la perte en précision du modèle et au biais dans l'estimation des paramètres.

### 3.3.2 Application

La problématique des valeurs manquantes est présente en MRA. Elle est principalement causée par la spécificité des différentes activités.
Par exemple, pour un agriculteur qui ne fait pas d'élevage, toutes les variables spécifiques à l'activité d'élevage sont manquantes. Les valeurs manquantes ainsi obtenues sont du type MNAR. Le traitement effectué a donc consisté à transformer les *"NA"* en *"Non concerné"*, afin d'ajuster un coefficient à cette modalité lors de la modélisation.
Pour ce qui est des valeurs manquantes de type MCAR, leur présence dans la base de données était liée dans un premier temps à l'absence des caractéristiques des contrats liés à l'ancienne gamme du produit MRA. En effet, les caractéristiques de ces contrats sont enregistrées dans une base avec un encodage méconnu de l'expertise actuelle. Ceux-ci représentaient 7% de la base de modélisation. Ils ont été retirés de la base de modélisation, puisqu'il n'y a plus de souscription de l'ancienne gamme depuis plus de vingt ans.
Le second groupe de variables avec des valeurs manquantes de types MCAR est constitué de l'âge de l'exploitant et l'année de création de l'exploitation. Après analyse, 1% des observations avaient l'année de création manquante et moins de 4% des observations avaient l'âge du client non renseigné. Par souci de simplification, l'imputation sur l'âge a été faite par une moyenne en fonction de l'activité et la qualité. Le mode [3] de l'année de création par activité et selon la qualité a été utilisé pour l'imputation de la variable année de création.

**A retenir** :
— **Choix du seuil** : seuil commun choisi pour l'habitation et l'exploitation ;
— **Seuil incendie** : 310 000 € ;
— **Seuil Tempête grêle neige** : 31 790 € ;
— **Traitement des sinistres en cours** : méthode de Chain-Ladder ;
— **Traitement des valeurs manquantes** : imputation par la moyenne ou le mode.

3. La modalité avec le plus d'occurrences

# Chapitre 4

# Théorie sur la tarification d'un contrat d'assurance

L'objectif de ce chapitre est de poser les bases théoriques nécessaires à la mise en place et l'évaluation d'un modèle de prime pure. La première partie présentera la théorie relative aux modèles linéaires généralisés. La seconde partie de ce chapitre sera consacrée à l'exposition de quelques modèles d'apprentissage statistique. Enfin, une dernière section sera dédiée à la présentation des différents outils d'évaluation des modèles.

## 4.1 Modèle collectif

L'une des approches les plus classiques pour la modélisation de la prime pure est le modèle collectif. Dans cette approche, la prime pure est définie comme l'espérance de la charge des sinistres survenus sur une période (généralement l'année en assurance IARD). Soient :

- $S$ la charge totale de sinistres au cours d'un exercice ;
- $X_j$ la charge du $j^{\text{ième}}$ sinistre au cours de l'exercice : variable aléatoire à valeurs dans $\mathbb{R}_+$ ;
- $N$ le nombre de sinistres au cours de l'exercice : variable aléatoire à valeurs dans $\mathbb{N}$.

D'où :

$$S = \sum_{j=1}^{N} X_j$$

Les deux hypothèses clés de cette modélisation sont les suivantes :

- L'indépendance entre $N$ et $X_j$ pour tout $j$ ;
- $(X_j)_{j\geq 1}$ est une suite de variables aléatoires indépendantes et identiquement distribuées.

Par application de ces hypothèses, la prime pure se décompose de la manière suivante en passant par l'espérance conditionnelle :

$$\textit{Prime pure} = \mathbb{E}(S) = \mathbb{E}[\mathbb{E}(X|N)] = \mathbb{E}[\sum_{j=1}^{N}\mathbb{E}(X_j|N)] = \mathbb{E}[N \times \mathbb{E}(X_j)] = \mathbb{E}(N) \times \mathbb{E}(X_j)$$

Empiriquement, l'estimation est la suivante :

$$\begin{aligned}\textit{Prime pure} &= \textit{fréquence} \times \textit{coût moyen des sinistres} \\ &= \frac{\textit{Nombre de sinistres}}{\textit{Exposition}} \times \frac{\textit{Charge totale des sinistres}}{\textit{Nombre de sinistres}} \\ &= \frac{\textit{Charge totale des sinistres}}{\textit{Exposition}}\end{aligned}$$

Cette dernière égalité correspond à la prime pure observée sur les données.

## 4.2 Théorie sur les modèles linéaires généralisés

Les modèles linéaires généralisés sont les modèles les plus couramment rencontrés en tarification en assurance IARD. Ces modèles ont été présentés de manière complète par Mc Cullagh et Nelder (1989) et sont prisés pour leur interprétabilité.
Les modèles linéaires généralisés permettent d'établir une relation linéaire entre un ensemble de variables appelées prédicteurs $X = (X_1, ..., X_p)$ et une variable réponse $Y$. L'établissement de cette relation linéaire se fait par le biais d'une fonction de lien $g$ supposée monotone et différentiable au moins une fois telle que $g(\mathbb{E}[Y|X]) = X\beta$, avec $\beta$ la matrice des coefficients associée à $X$.
Un modèle linéaire généralisé est caractérisé par trois composantes :

- *La composante aléatoire* : il s'agit de la loi de probabilité appartenant à la famille de lois exponentielles associées à la distribution de la variable Y ;
- *La composante déterministe* : c'est la matrice des prédicteurs X, l'ensemble des variables explicatives ;
- *La fonction de lien* : c'est la fonction qui établit la relation fonctionnelle entre la composante aléatoire et la composante déterministe.

### La famille de loi exponentielle

Dans le cadre d'un modèle linéaire généralisé, la loi de la distribution de la variable réponse Y doit appartenir à la famille de lois exponentielles. Cette propriété permet la modélisation de variables dont le support n'est pas nécessairement $\mathbb{R}$ tout entier mais des supports réduits tel que $\mathbb{R}_+$ ou $\mathbb{N}$.
Une loi de probabilité appartient à la famille des lois exponentielles si la fonction densité de probabilité qui lui est associée est de la forme :

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

Avec :

— $a(\phi)$ : une fonction non nulle ;
— $b(\theta)$ : une fonction de classe $C^2$ et convexe ;
— $c(y, \phi)$ : une fonction ne dépendant pas de $\theta$ ;
— $\phi$ : un paramètre de dispersion ;
— $\theta$ : un paramètre canonique.

Les relations suivantes sont vérifiées par l'espérance et la variance d'une variable aléatoire de la famille de lois exponentielles :

$$\mathbb{E}(Y) = \mu = b^{'}(\theta) \text{ et } \mathbb{V}(Y) = \sigma^2 = b^{''}(\theta) \times a(\phi)$$

Le tableau ci-dessous présente quelques lois de familles exponentielles fréquemment rencontrées en actuariat :

| Loi et Support | $a(\phi)$ | $b(\theta)$ | $c(y, \phi)$ | $g(x)$ | Expression |
|---|---|---|---|---|---|
| Normale $(\mu, \sigma^2) : \mathbb{R}$ | $\sigma^2$ | $\frac{\theta^2}{2}$ | $-\frac{1}{2}(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2))$ | $x$ | $\hat{y_i} = \hat{\beta_0} + \sum_{j=1}^{p} \hat{\beta_j} * x_{i,j}$ |
| Bernoulli $(p) : 0, 1$ | 1 | $ln(1 + e^{\theta})$ | 0 | $ln(\frac{x}{1-x})$ | $\hat{y_i} = (1+exp(-[\hat{\beta_0} + \sum_{j=1}^{p} \hat{\beta_j} * x_{i,j}]))^{-1}$ |
| Poisson$(\lambda)$ : $\mathbb{N}$ | 1 | $e^{\theta}$ | $-ln(y!)$ | $ln(x)$ | $\hat{y_i} = exp(\hat{\beta_0} + \sum_{j=1}^{p} \hat{\beta_j} * x_{i,j})$ |
| Gamma $(\alpha, \lambda) : \mathbb{R}_+$ | $\frac{1}{\alpha}$ | $-ln(-\theta)$ | $(\frac{1}{\phi} - 1)ln(y) - ln(\Gamma(\frac{1}{\phi}))$ | $\frac{1}{x}$ | $\hat{y_i} = (\hat{\beta_0} + \sum_{j=1}^{p} \hat{\beta_j} * x_{i,j})^{-1}$ |

TABLE 4.1 – Exemples de lois appartenant à la famille exponentielle

Les estimateurs des coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ sont obtenus par la méthode du maximum de vraisemblance. La log-vraisemblance s'exprime de la manière suivante :

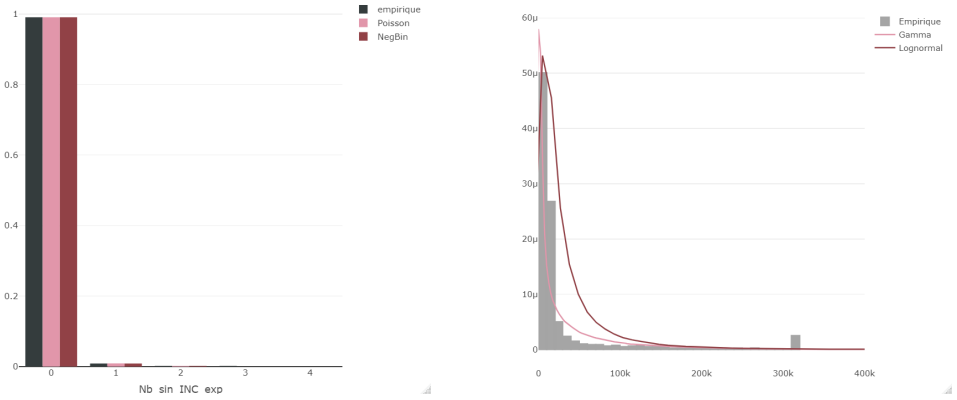
$$log(L(\beta)) = \sum_{i=1}^{n} \left[ \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right]$$

Ces estimateurs s'obtiennent finalement après résolution numérique des p équations différentielles de la fonction de log-vraisemblance.
N.B : En général, la fonction $a$ est définie par $a(\phi) = \phi$

### 4.2.1 Ajustement des lois

Après avoir présenté la théorie sur les modèles linéaires généralisés, il faut choisir laquelle des lois appliquer pour la modélisation. Ce choix s'effectue en vérifiant qu'il y ait une bonne adéquation entre la loi théorique et loi empirique sur nos données.
Dans le cadre d'une modélisation fréquence - coût, l'ajustement des lois théoriques se fera sur le nombre de sinistres et sur le coût moyen des sinistres.

### Garantie incendie




(a) Ajustement de lois - nombre de sinistres exploitation

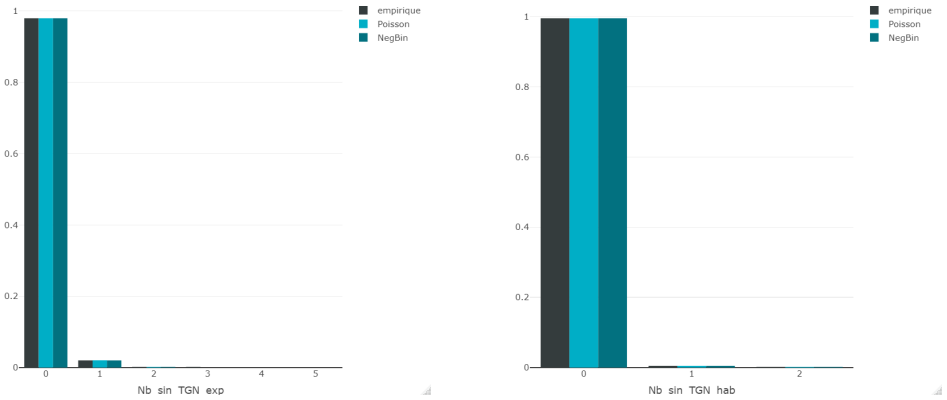
(b) Ajustement de lois - coût moyen

FIGURE 4.1 – Ajustement de lois - Garantie incendie

Pour la modélisation du nombre de sinistres incendie exploitation, la loi de Poisson et la loi binomiale négative s'ajustent bien à nos données (4.1-a). De plus, $\hat{\mathbb{E}}[N] \approx \hat{\mathbb{V}}[N]$ avec $N$ le nombre de sinistres. Ainsi, il paraît légitime d'utiliser une loi de Poisson pour la modélisation.
Au regard du graphique 4.1-b sur la densité empirique du coût moyen écrêté mutualisé, il semble judicieux d'utiliser une loi gamma pour la modélisation. En effet, la loi gamma sous-estime les coûts moyens faibles mais décrit assez correctement la tendance observée sur l'ensemble de la distribution ; tandis que l'ajustement d'une loi log-normale entraînerait une surestimation du coût moyen dans l'ensemble.

### Garantie TGN

Comme sur la garantie incendie, il y a une bonne adéquation entre la loi empirique et la loi de Poisson sur le nombre de sinistres habitation et exploitation (figure 4.2). Cette loi est alors retenue pour la modélisation de la fréquence du risque habitation et exploitation pour la garantie tempête grêle neige.




(a) Ajustement de lois - nombre de sinistres exploitation

(b) Ajustement de lois - nombre de sinistres habitation

FIGURE 4.2 – Ajustement de lois - Garantie TGN

Pour ce qui est du coût moyen des deux risques (exploitation et habitation), la loi gamma et la loi log-normale s'ajustent bien à nos données, comme observé sur la figure 4.3. Toutefois, la loi gamma propose une meilleure estimation du coût moyen des petits sinistres. Celle-ci est alors retenue pour la modélisation.

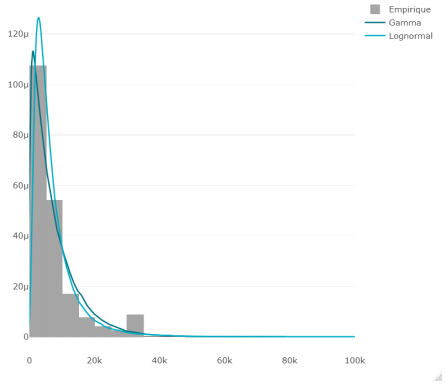



FIGURE 4.3 – Ajustement de lois sur le coût moyen - Garantie TGN

### 4.2.2 Sélection des variables

La sélection des variables explicatives est une phase de la modélisation qui vise à choisir les variables qui permettraient de prédire au mieux la valeur prise par la variable réponse $Y$. Il existe une panoplie de méthodes permettant d'avoir un modèle parcimonieux, telles que la sélection pas à pas ou la régression pénalisée.

**Régularisation**

La régularisation, ou pénalisation, vise à "encadrer" la valeur prise par les coefficients $\hat{\beta}_i$ dans l'estimation par la méthode de vraisemblance. Cette régularisation s'effectue en ajoutant un terme de pénalisation dans l'expression de la log-vraisemblance dans le cadre d'un GLM. L'ajout de ce terme permet de limiter les risques de sur-apprentissage du modèle. Par ailleurs, la régularisation permet également de réduire le nombre de variables utilisées dans le modèle en mettant des coefficients $\hat{\beta}_j$ à 0 pour faciliter l'interprétabilité du modèle.
Trois types de régularisation sont généralement rencontrés en tarification : LASSO (Least Absolute Shrinkage and Selection Operator), Ridge et Elastic Net. Cette dernière est une combinaison des deux premiers types de régularisation.

**Régularisation LASSO**

La régularisation de type LASSO force la nullité des coefficients des variables explicatives les moins pertinentes dans la modélisation. Cette régularisation utilise la norme $L^1$ des coefficients dans le terme de régularisation au cours de la maximisation de la log-vraisemblance. Les $\hat{\beta}_{LASSO}$ sont obtenus en maximisant l'expression :

$$\sum_{i=1}^{n}\left[\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i,\phi)\right] - \lambda\sum_{j=1}^{p}|\beta_j| = log(L(\beta)) - \lambda||\beta||_1$$

Le choix de la valeur du paramètre $\lambda$ joue un rôle clé dans la régularisation. Ce paramètre contrôle l'intensité de la pénalisation. En pratique, plus la valeur de $\lambda$ choisie est grande, plus le nombre de coefficients $\hat{\beta}_i$ égaux à 0 augmente. En théorie, le choix d'un $\lambda$ infini conduit à l'obtention d'un modèle avec un vecteur $\hat{\beta}$ nul. Par ailleurs, la valeur de $\lambda$ permet d'ajuster le compromis biais/variance, dans la mesure où la variance d'un modèle a souvent tendance à augmenter avec un grand nombre de variables.

### 4.2.3 Application

Avant de commencer la sélection de variables à proprement dite, il paraît intéressant de savoir s'il serait correct de modéliser le risque exploitation et le risque habitation par un seul modèle.

Pour répondre à cette interrogation, une sélection de variables a été effectuée sur les modèles de fréquence et de coût avec la variable « *type de risque* »(habitation ou exploitation). L'idée est de savoir si la variable « *type de risque* » aura un coefficient nul dans la modélisation. La régression de LASSO a été utilisée pour la sélection de variables sur ces modèles. Cette sélection de variable a été optimisée par validation croisée, pour obtenir un $\lambda$ optimal qui minimise la déviance du modèle. Au terme de la sélection de variables, la variable « *type de risque* » a été retenue comme significative dans la modélisation de la fréquence des deux garanties étudiées.

Pour aller plus loin dans cette étude, une sélection de variables sur les modèles de fréquence des deux risques habitation et exploitation est testée individuellement. Ce test a permis de vérifier qu'il y a une différence significative entre les variables sélectionnées pour les deux risques. Les variables liées à l'exploitation avaient des coefficients nuls ou moins importants en valeur absolue dans le modèle de fréquence associé au risque habitation.
Pour terminer, une analyse des interactions entre la variable « type de risque » et quelques variables a été faite. Cette étude a permis d'observer qu'une variable peut être discriminante pour l'exploitation et non pour l'habitation. Par exemple (graphique 4.4), la valeur relative au contenu de l'exploitation est une variable discriminante sur l'exploitation ; mais elle ne l'est pas sur l'habitation. Ainsi, il est préférable de modéliser la fréquence de sinistres de l'exploitation et de l'habitation séparément pour ne pas avoir un modèle biaisé.

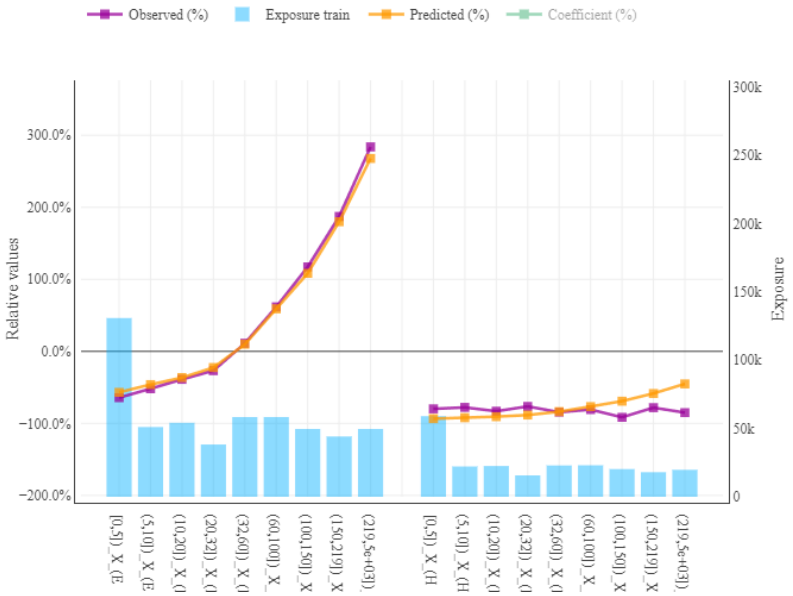
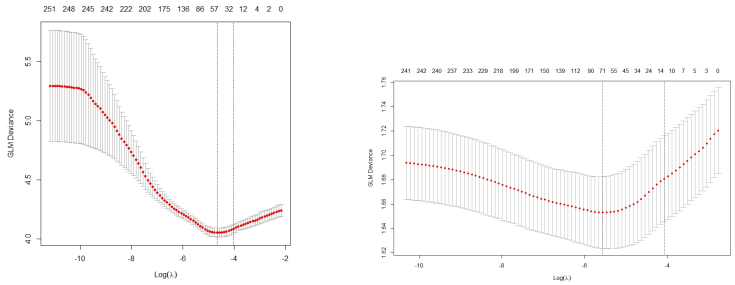



FIGURE 4.4 – Interaction des variables « type de risque » et « contenu de l'exploitation» incendie (exploitation à gauche, habitation à droite)

L'étape suivante consiste à vérifier s'il est nécessaire de réaliser une modélisation du coût moyen en distinguant l'habitation de l'exploitation. Cette étape servira à confirmer ou rejeter l'idée d'une unique modélisation, évoquée dans la section relative aux valeurs extrêmes 3.1.2.
En appliquant une régression pénalisée de LASSO sur le coût moyen global incendie, avec la valeur $\lambda_{min}$ [1] 0,009419, la variable « *type de risque* » a un coefficient nul. En utilisant la même démarche sur la garantie TGN, avec la valeur $\lambda_{min}$ 0,003826, la variable « *type*

1. Valeur minimale de $\lambda$ qui minimise la déviance par validation croisée

*de risque* » a un coefficient quasi nul (<0,008).




(a) Déviance coût moyen incendie $\lambda$ Lasso

(b) Déviance coût moyen TGN $\lambda$ Lasso

FIGURE 4.5 – Déviance du GLM gamma en fonction de $\lambda$

La déviance globale du modèle ne serait pas améliorée significativement par l'ajout de cette variable. Ceci serait probablement lié à la faiblesse du volume de sinistres habitation. Par conséquent, étant donné le manque d'impact de cette variable et pour éviter de complexifier la structure de modélisation en vue d'une éventuelle implémentation tarifaire, il semble raisonnable d'utiliser un seul modèle de coût moyen à partir des résultats obtenus.
Néanmoins, pour avoir un modèle plus précis sur l'habitation, il serait potentiellement intéressant de vérifier - de manière analogue a ce qui a été fait pour la fréquence - que les mêmes variables influent sur le coût moyen exploitation et le coût moyen habitation, ainsi que de vérifier que la segmentation optimale déterminée au global, principalement influencée par les tendances observées sur l'exploitation, soit aussi optimale pour l'habitation. Cet exercice est cependant moins réalisable sur la sévérité que sur la fréquence étant donné le faible nombre de sinistres habitation. D'autre part, étant donnée la faiblesse de la fréquence habitation, une segmentation sous-optimale de la sévérité sur le périmètre de l'habitation n'aura pas d'effet néfaste sur la puissance globale du modèle et ne risque donc pas de conduire à une erreur tarifaire ou une anti-sélection majeure.

La sélection de variables proprement dite s'effectue en deux temps sur chaque modèle. Une première sélection est réalisée en utilisant l'analyse des corrélations entre les variables et le score d'importance des variables dans la prédiction de la variable cible. Le score d'importance est proportionnel à la valeur de $\lambda_{LASSO}$ qui permet d'annuler tous les coefficients de la variable.

Le deuxième niveau de sélection de variables est effectué en testant la nullité des coefficients des différentes variables en entrée du modèle avec l'outil de tarification Akur8. Pour les variables catégorielles, la différence entre le coefficient de chaque modalité et 0 est testée. Concernant les variables ordinales, le test est basé sur la différence des coefficients des modalités voisines (en d'autres termes, si leur différence vaut 0). Cette

méthode peut aussi s'exprimer sous la forme d'une équation Lagrangienne, et peut être interprétée comme une régression pénalisée de LASSO.
En dernier lieu, une grille de modèles est proposée en fonction du nombre de variables et du Gini du modèle (graphique 4.6). Le modèle choisi est celui qui a la meilleure qualité de segmentation et le minimum de variables.


FIGURE 4.6 – Sélection des modèles en fonction du Gini

## 4.3 Théorie des modèles d'apprentissage statistique

### 4.3.1 Arbres de décision

Faisant partie des méthodes d'apprentissage statistique, les arbres de décision permettent d'expliquer et de prédire la variable réponse Y en partitionnant l'espace des variables explicatives. Les arbres de décision sont très prisés pour leur facilité d'interprétation et pour leur capacité de classification et de régression avec les modèles CART (Classification And Regression Tree, Breiman 1984).
La structure hiérarchique de l'arbre de décision est la suivante :

- La racine : le sommet de l'arbre, qui contient l'ensemble des observations en entrée du modèle ;
- Les noeuds : les différentes étapes de partitionnement des observations ;
- Les branches : les différentes règles de partitionnement pour chaque noeud de l'arbre ;
- Les feuilles : les différents groupes d'observations homogènes.

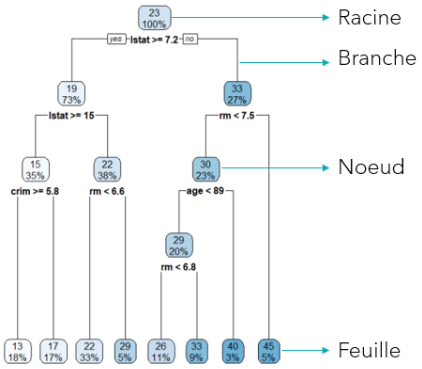



FIGURE 4.7 – Exemple d'arbre de décision : source [R-bloggers, 2021]

La construction de l'arbre de décision pour une régression se décompose au travers des étapes clés suivantes :

— **Le partitionnement** :
Pour la création de groupes homogènes, l'arbre segmente l'ensemble des variables explicatives afin de minimiser l'erreur de prédiction.
Soit $(X_j)_{j\in[|1,p|]}$ les variables quantitatives explicatives du modèle et $V_j$ l'ensemble des valeurs prises par la variable $X_j$. L'objectif est de déterminer pour chaque noeud la variable $X_{j^*}$ et le seuil $s^*$ à valeurs dans $V_j$ qui minimise la fonction de coût $C(j,s)$ telle que :

$$C(j^*,s^*) = \underset{(j,s)}{argmin} \sum_{x_i\in\{X|X_j<s\}} (y_i - \overline{y}\{X|X_j < s\})^2 + \sum_{x_i\in\{X|X_j\geq s\}} (y_i - \overline{y}\{X|X_j \geq s\})^2$$

Pour une classification, le critère de partitionnement à utiliser est l'indice de Gini :

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - p_{mk})$$

Où $\hat{p}_{mk}$ est la proportion d'observations de la classe $k$ dans l'espace de partitionnement (ensemble) $m$. Le critère à minimiser est l'indice de Gini de chaque ensemble afin d'obtenir les sous-arbres les plus homogènes possible.

— **La règle d'arrêt**
Pour avoir un modèle généralisable, il faut ajuster la profondeur de l'arbre pour éviter un sur-apprentissage. Ainsi, il est nécessaire de définir une règle d'arrêt pour stopper la progression de l'arbre. Deux critères d'arrêt anticipé sont généralement utilisés :
— Le nombre d'observations minimum par feuille ;
— La profondeur maximale de l'arbre.

— **L'élagage**
L'élagage consiste à sélectionner le meilleur sous-arbre en élaguant les branches de l'arbre maximal. L'idée est de trouver un arbre intermédiaire entre l'arbre maximal et l'arbre avec uniquement la racine. Cet arbre intermédiaire optimal est choisi de telle sorte qu'il ait une capacité de prédiction proche de l'arbre maximal tout en évitant au mieux le sur-apprentissage. Dans le cas de la régression, cet arbre est sélectionné à partir d'une suite d'arbres élagués de l'arbre maximal qui minimise le critère suivant :

$$Crit_\alpha(T) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{T,i})^2 + \alpha|T|$$

Avec $\alpha$ le paramètre de pénalisation et $|T|$ le nombre de feuilles de l'arbre intermédiaire T. Parmi ces arbres, l'arbre intermédiaire sélectionné est celui qui a l'erreur de prédiction la plus faible soit en utilisant un échantillon test ou par validation croisée (cf. paragraphe 4.4.2).

— **La variable réponse**
A la fin de la construction de l'arbre, chaque feuille contient une réponse, prédiction de la variable réponse (régression), pour le groupe homogène obtenu :

$$\hat{y}_{Feuille_i} = \frac{1}{card(Feuille_i)}\sum_{j \in Feuille_i} y_j$$

### 4.3.2 *Gradient Boosting Machine* : GBM

Gradient Boosting est un algorithme d'apprentissage supervisé basé sur l'agrégation de modèles. La mécanique sous-jacente à cet algorithme consiste à regrouper des modèles simples avec une faible qualité de prédiction pour obtenir une meilleure prédiction.
En pratique, l'algorithme fonctionne de manière itérative en utilisant le résultat du modèle $i$ pour ajuster celui du modèle $i+1$ jusqu'à ce que la condition d'arrêt soit atteinte. Dans le cadre d'une régression, ce sont les résidus obtenus à partir du modèle $i$ qui servent d'input au modèle $i+1$ avec un ajustement par minimisation de la fonction de coût globale par descente du gradient. Pour une classification, l'ajustement se fait par actualisation des poids des différents individus [2] entre le modèle $i$ et $i+1$ avec une descente du gradient.
En général, les modèles faibles utilisés sont les arbres de décision. L'algorithme dans le cadre d'une régression est le suivant :

2. Les poids sont tous égaux initialement (à t=0).

**Algorithm 1** Gradient Boosting Machine

Initialisation du modèle avec une valeur constante :

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

**for** m=1 à M **do**

1. Calcul des résidus du modèle $i$ :

$$r_m^i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{Pour } i = 1, \ldots, n$$

2. Entraînement de l'arbre de régression $T_m$ aux données $\{(x_i, r_m^i)\}_{i=1}^n$

3.Calcul de $\gamma_m$ par minimisation avec descente du gradient :

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma T_m(x_i)\right)$$

.

4.Mise à jour du modèle : $F_m(x) = F_{m-1}(x) + \gamma_m T_m(x)$

**end for**

Résultat : $F_M(x)$

Dans ce mémoire, l'algorithme XGBoost a été utilisé comme implémentation du Gradient Boosting. La particularité du XGBoost est qu'il est optimisé pour effectuer plus rapidement les calculs nécessaires à l'application du Gradient Boosting en traitant les données par blocs compressés. Cette mécanique permet, d'une part, à l'algorithme d'être plus rapide dans le tri des données et, d'autre part, un traitement en parallèle de ceux-ci. Les paramètres de ce modèle sont les suivants :

- *nrounds* : le nombre d'itérations dans le processus d'agrégation ;
- max_depth : le nombre de noeuds maximal de chaque arbre ;
- eta : la vitesse d'apprentissage des arbres ;
- colsample_bytree : la proportion de variables utilisées lors de la construction d'un arbre ;
- gamma : le minimum de la fonction de perte requis pour créer une nouvelle partition sur le noeud d'un arbre ;
- min_child_weight : la somme minimale de poids nécessaire pour créer une nouvelle partition. Elle peut être considérée comme le nombre minimum d'observations nécessaires dans chaque feuille dans le cadre d'une régression linéaire.
- subsample : la proportion d'observations utilisées pour l'entraînement des arbres.

## 4.4 Évaluation des modèles

Les métriques d'évaluation des modèles servent à mesurer la qualité de précision et de segmentation des modèles. Ces métriques permettent de juger la robustesse des modèles.

### 4.4.1 Métriques d'évaluation des modèles

#### Déviance

La déviance est une statistique qui mesure la qualité d'ajustement du modèle calibré. L'idée est de comparer la vraisemblance du modèle saturé, qui correspond à un modèle avec autant de variables que d'observations, et celle du modèle calibré.
L'inconvénient de cette statistique est qu'elle ne tient pas compte du sur-apprentissage. La sensibilité de cette statistique avec l'ajout d'une variable au modèle dépend de la pertinence.

$$Déviance = -2(L_m - L_s)$$

avec $L_m$ et $L_s$ les valeurs de la fonction du maximum de vraisemblance maximisée respectivement du modèle étudié et du modèle saturé.
Dans une comparaison de modèle, le modèle avec la plus petite déviance est retenu.

#### Indice de Gini

Initialement construit pour évaluer la répartition des richesses dans une population, l'indice de Gini est une métrique qui peut être utilisée en tarification pour mesurer la qualité de segmentation du modèle de prime pure. Cet indice est calculé à partir de la courbe de Lorenz 4.8.

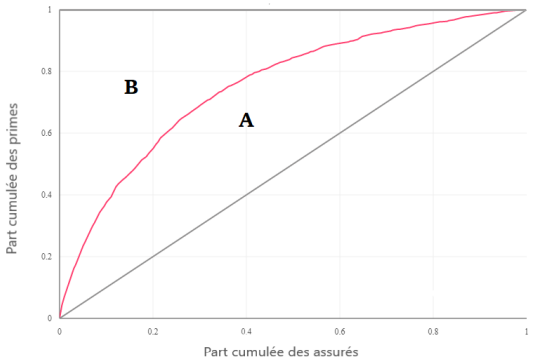



FIGURE 4.8 – Courbe de Lorenz

$$Indice\ de\ Gini = \frac{aire\ A}{aire\ A + aire\ B}$$

En tarification, un indice de Gini nul correspond à une situation de mutualisation égale du risque. Plus l'indice de Gini est élevé, plus la capacité de segmentation du modèle est élevée.

### RMSE : Root Mean Squared Error

Comme son nom l'indique, la RMSE mesure l'erreur quadratique moyenne d'un modèle. C'est un indicateur de mesure de la précision globale du modèle. L'ordre de grandeur de la RMSE dépend du type de données à prédire.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

### Erreur globale

L'erreur globale du modèle permet de mesurer la qualité d'ajustement globale du modèle.

$$Erreur\ globale = \frac{\Sigma\, Valeurs\ prédites - \Sigma\, Valeurs\ observées}{\Sigma\, Valeurs\ observées}$$

### La *lift curve*

La *lift curve* est une métrique graphique qui permet d'évaluer l'adéquation de la tendance des valeurs prédites aux valeurs observées. La *lift curve* permet de voir si la moyenne de la valeur prédite s'éloigne de la valeur moyenne observée par groupe d'assurés. Ces groupes sont obtenus en regroupant par quantile [3] les assurés triés préalablement par ordre croissant de la valeur prédite.

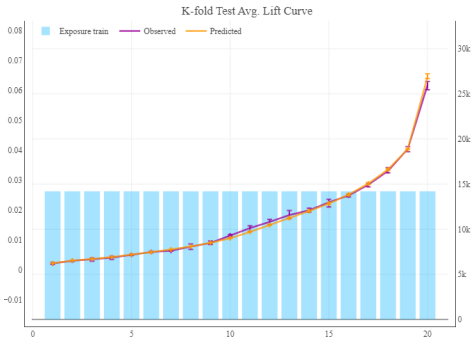



FIGURE 4.9 – Lift curve

### Les résidus quantiles

Les résidus quantiles servent à évaluer la qualité de l'ajustement de la loi théorique utilisée pour la modélisation. Ils sont obtenus en "transformant" la distribution du modèle en

3. Généralement, les regroupements sont réalisés par quantile de 5%.

distribution gaussienne centrée réduite.
Les résidus quantiles sont définis par :

$$r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\phi}\}$$

Avec :

- $y$ le vecteur des réponses de loi $\mathcal{P}(\mu, \phi)$ avec $\mu_i = E[y_i]$ et $\phi$ un paramètre commun à tous les $y_i$ supposés indépendants ;
- $F(y)$ la fonction de distribution continue de loi $\mathcal{P}(\mu, \phi)$ et $F(y_i; \mu_i, \phi)$ distribuées uniformément sur $[0, 1]$ ;
- $\Phi$ la fonction de répartition de la loi normale centrée réduite.

Si les résidus quantiles sont distribués de manière gaussienne (graphique 4.10) selon l'axe des ordonnées, alors la distribution choisie pour la modélisation s'ajuste bien aux données.

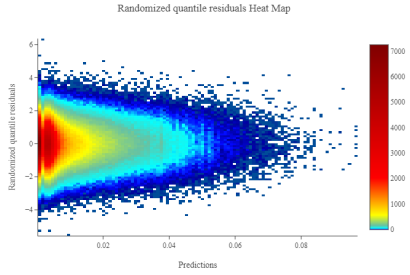
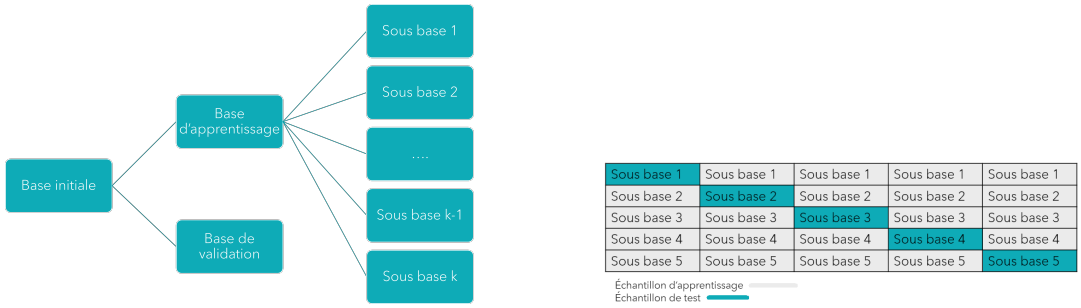



FIGURE 4.10 – Résidus quantiles

### 4.4.2 Validation croisée

La validation croisée est une technique qui vise à séparer la base de donnée initiale en deux sous-échantillons tirés aléatoirement. Un premier échantillon, appelé base d'apprentissage, sert à calibrer les coefficients du modèle. Il représente généralement 80% de la base initiale. Le second échantillon, appelé base de validation, permet d'évaluer la capacité de généralisation du modèle obtenu.

Pour optimiser la qualité d'ajustement des coefficients obtenus sur la base d'apprentissage, la méthode *k-folds* est appliquée. Cette méthode consiste à calibrer $k$ fois le modèle en partitionnant la base d'apprentissage en $k$ échantillons de même taille. Le principe est d'utiliser les $k - 1$ échantillons pour ajuster les coefficients et de tester le modèle obtenu sur le k-ième échantillon. Cette opération est itérée $k$ fois afin que chacun des $k$ échantillons serve de base de test. Les coefficients finaux sont la moyenne des différents coefficients obtenus sur les sous-échantillons.




(a) Validation croisée (b) k-*folds*

FIGURE 4.11 – Structure de la validation croisée

**A retenir** :

— **Approche de modélisation** : fréquence-coût moyen ;
— **Modélisation de la fréquence** : besoin de segmentation entre les deux risques ;
  — Habitation : GLM Poisson ;
  — Exploitation : GLM Poisson ;
— **Modélisation du coût moyen** : intérêt de la segmentation des deux risques négligeable par rapport à la complexité opérationnelle, unique GLM gamma ;
— **Sélection des variables** : régression Lasso et test de nullité des coefficients.

# Chapitre 5

# Modélisation de la fréquence et du coût moyen

Suite à la présentation des différentes théories des modèles, ce chapitre est dédié à leur mise en application dans l'objectif de modéliser la fréquence et le coût moyen des sinistres.

## 5.1 Contexte

Pour rappel, le caractère optionnel de l'assurance de l'habitation dans le contrat MRA conduit à une diminution de la proportion de contrat avec l'option habitation.

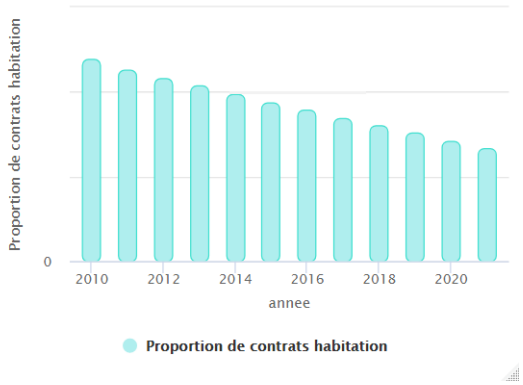



FIGURE 5.1 – Proportion de contrats habitation

Cette diminution de la proportion de contrats habitation a un impact sur la réalisation d'un modèle stable. Pour tenir compte de la particularité des deux risques, la structure de la modélisation est la suivante :

$$Prime\ pure = (\mathbb{1}_{Exploitation} * fréquence_{Exploitation} + \mathbb{1}_{Habitation} * fréquence_{Habitation}) * coût\ moyen\ global$$

Cette structure avec un coût moyen global se justifie, d'une part, par la faiblesse de la volumétrie de sinistres pour établir un modèle de coût stable pour les deux risques ; et, d'autre part, en utilisant les résultats de la section sur la sélection des variables 4.2.3.

## 5.2 Modèles de fréquence

### 5.2.1 Modélisation GLM

La modélisation de la fréquence de sinistres des deux garanties s'effectuera avec un modèle linéaire généralisé de Poisson.

**Incendie**

— **Habitation**
La difficulté majeure dans la réalisation du modèle de fréquence de sinistres habitation réside dans la faiblesse du nombre de sinistres sur l'habitation. En effet, la fréquence de sinistres sur cette garantie est de l'ordre 0,1%. En essayant de calibrer un modèle de fréquence sur ces données, les résultats obtenus sont les suivants :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson | 0,0427 | 31,98% | 0,0427 | 18,47 % |

TABLE 5.1 – Résultats de la modélisation fréquence incendie habitation

Le constat est assez clair (5.1) : le modèle a sur-appris sur les données d'apprentissage. L'écart entre le Gini calculé sur la base d'apprentissage et celui calculé sur la base de validation est très grand, ce qui signifie que le modèle n'a pas une bonne qualité de segmentation et qu'il est instable. Par conséquent, ce modèle ne réussira pas à faire de prédictions correctes pour la fréquence de sinistralité des assurés.

Face à cette problématique, une solution envisageable est de calculer une fréquence empirique par groupe de risque. Les deux variables qui seront utilisées pour la création des groupes de risque seront le nombre de pièces et la valeur du contenu de l'habitation. Ces deux variables sont celles qui sont ressorties significatives lors de la modélisation. De plus, ces deux variables décrivent au mieux la fréquence des sinistres au regard des statistiques descriptives.
Finalement, en croisant ces deux variables, la table de fréquence de sinistres obtenue est la suivante :

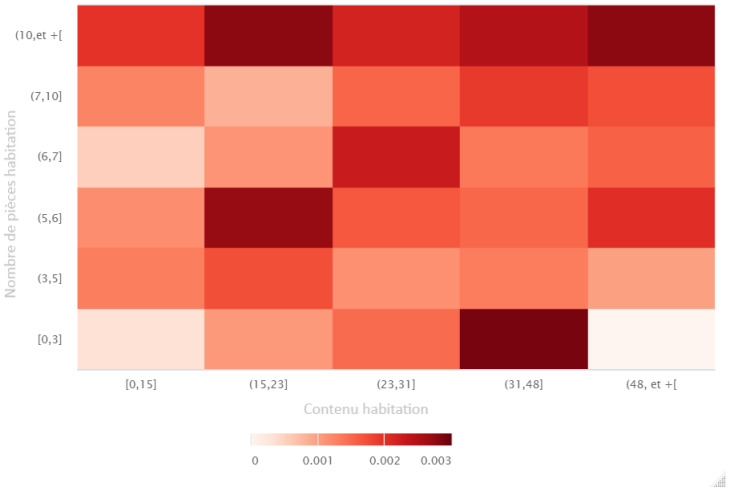



FIGURE 5.2 – Fréquence de sinistres habitation

L'analyse du graphique 5.2 montre la tendance à la hausse de la fréquence avec le nombre de pièces et la valeur du contenu de l'habitation. En se basant sur l'historique des sinistres, il apparaît que les grandes habitations (plus de 10 pièces) présentent plus de risque incendie. Cette analyse stricte est corrigée par la valeur du contenu habitation. En utilisant le croisement du nombre de pièces avec le contenu de l'habitation, il est possible d'observer que les petites habitations (0 à 3 pièces) avec de grands contenus présentent une fréquence de sinistres élevée. Enfin, le croisement du nombre de pièces et de la valeur du contenu permet d'observer que les habitations de taille moyenne (5 à 7 pièces) qui ont un contenu moyen (15 à 31) présentent un risque incendie élevé.
En somme, cette table servira de substitut au modèle GLM instable.

— **Exploitation**
Les résultats de la modélisation de la fréquence de sinistres incendie sur l'exploitation sont les suivants :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson | 0,118 | 49,08% | 0,118 | 48,44% |

TABLE 5.2 – Résultats de la modélisation fréquence incendie exploitation

Au vu des résultats affichés dans le tableau 5.2, il possible de dire que le modèle sur la fréquence présente de bonnes performances. Les valeurs de l'indice de Gini et de la RMSE sont assez stables sur la base d'apprentissage et la base de test. Ces performances permettent d'affirmer que le modèle réussit à segmenter le risque et qu'il est généralisable.
Pour aller plus loin dans la validation du modèle, l'analyse de la courbe de Lorenz, de la *lift curve* et des résidus peut être effectuée.

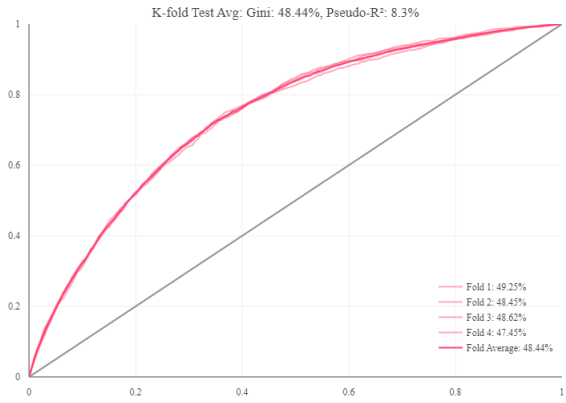



FIGURE 5.3 – Courbes de Lorenz sur base de validation : *k-folds*

Les différentes courbes de Lorenz (5.3) permettent d'affirmer que la qualité de segmentation du modèle reste stable sur les sous-échantillons de la base de validation ; ce qui confirme l'absence de sur-apprentissage et la capacité du modèle à prédire de bons résultats sur de nouvelles données.
L'analyse des résidus quantiles (5.4) du modèle permet de valider la pertinence du choix de la loi utilisée pour la modélisation et de la fonction de lien. Si le modèle est bien ajusté, les résidus quantiles doivent avoir une tendance normale et être compris entre -2 et 2 ; ce qui est bien le cas dans la modélisation de la fréquence incendie exploitation.

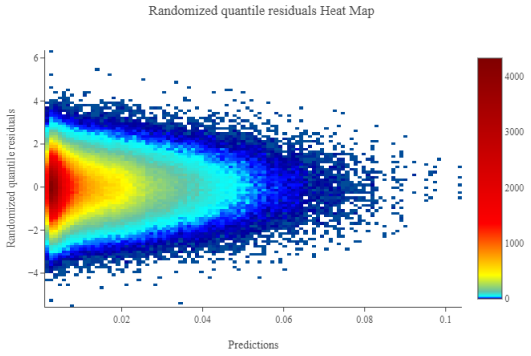



FIGURE 5.4 – Résidus quantiles : fréquence incendie exploitation

Enfin, le dernier outil de validation utilisé est la *lift curve*, qui met en confrontation la courbe des quantiles sur la fréquence prédite et celle de la fréquence observée. La performance du modèle est jugée par la proximité de la courbe de la fréquence prédite et observée.

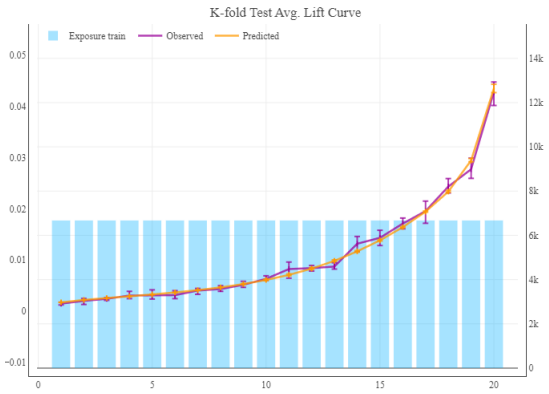



FIGURE 5.5 – *lift curve* : fréquence incendie exploitation

La proximité des courbes de fréquence prédite et observée (5.5) est satisfaisante. Le modèle prédit assez correctement la fréquence observée.
L'étape suivante concerne la compréhension du modèle. Les variables significatives du modèle sont les suivantes :

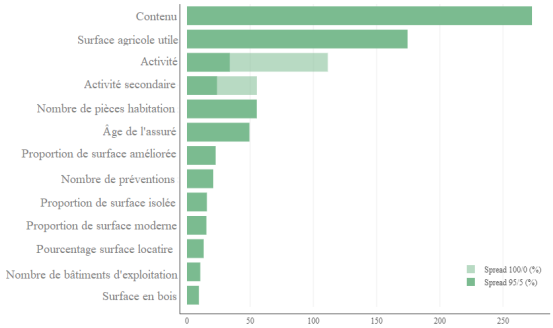


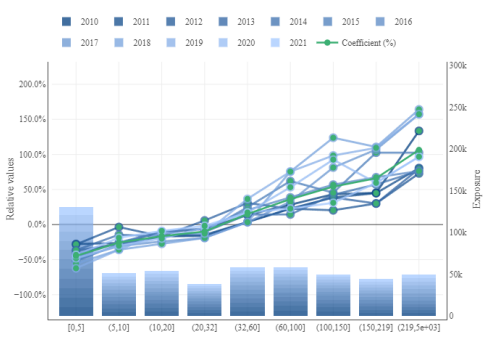

FIGURE 5.6 – Variables significatives : fréquence incendie exploitation

Les variables ainsi présentées en figure 5.6 sont classées par ordre d'importance dans la prédiction. L'importance des variables est mesurée à partir du *spread* de la variable.

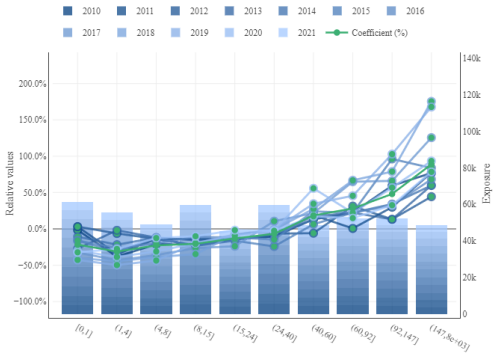
$$Spread = \frac{max(Coefficient)}{min(Coefficient)} - 1$$

A partir de cet indicateur, il est possible d'identifier les variables avec une évolution significative des coefficients. Deux valeurs de *spread* sont calculées : le *spread* 100/0(%), qui tient compte de tous les coefficients dans la formule du *spread*, et le *spread* 95/5(%), qui est calculé en enlevant 5% des plus grands coefficients et des plus faibles.

**Résumé de la modélisation :**
Le modèle décrit assez parfaitement les tendances observées dans la section statistiques descriptives. Outre ces tendances déjà décrites, les assurés exerçant une activité secondaire de production de viande bovine et de récolte présentent une fréquence de sinistres supérieure à la tendance générale. Cet effet est capté par la variable activité secondaire, d'où son intérêt. Par ailleurs, la fréquence de sinistres diminue avec l'âge de l'assuré, grâce à l'expérience qu'il a acquise. De plus, le risque incendie est faible sur les exploitations où plus de 90% de la surface de l'exploitation est moderne. En outre, les assurés ayant une habitation en plus de l'exploitation sont plus risqués que ceux qui ont uniquement l'exploitation. Enfin, les assurés avec une surface en bois importante semblent avoir une fréquence de sinistres incendie plus élevée, car le bois un matériau très inflammable.
Pour conclure, il est possible de regarder la stabilité temporelle des coefficients.




(a) Stabilité temporelle : fréquence incendie - contenu exploitation




(b) Stabilité temporelle : fréquence incendie - surface agricole utile

FIGURE 5.7 – Stabilité temporelle : fréquence incendie

La stabilité temporelle (5.7) est acceptable pour les variables « contenu de l'exploitation » et « surface agricole utile », qui sont les plus importantes du modèle. Bien que ce niveau de stabilité ne soit pas parfait, il est correct relativement au nombre de sinistres.

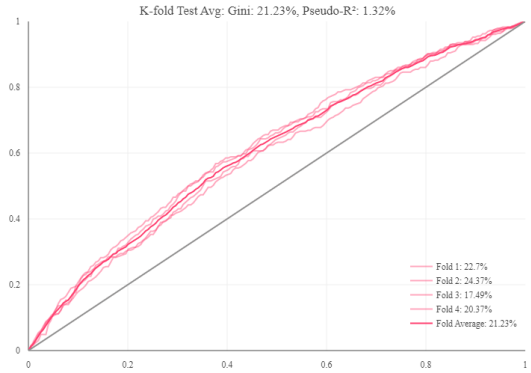
## TGN

L'un des avantages de cette garantie est la possibilité de réaliser un modèle de fréquence sur le risque habitation.

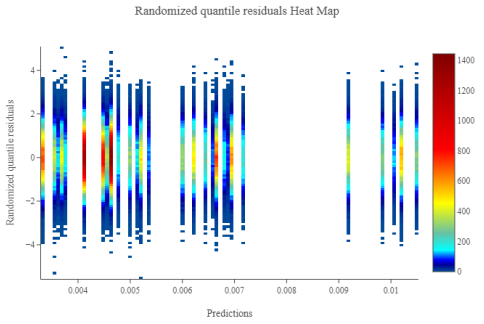
— **Habitation**
Le modèle obtenu après la calibration d'un GLM Poisson sur nos données présente les résultats suivants :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson | 0,089 | 21,71 % | 0,089 | 21,23 % |

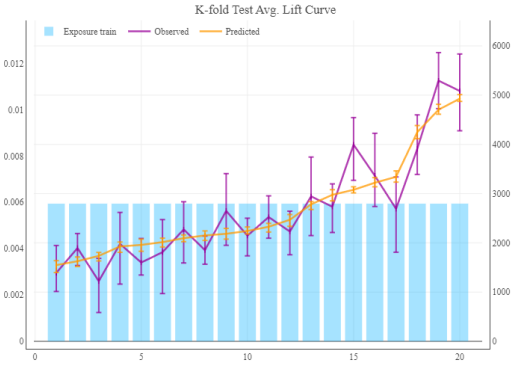
TABLE 5.3 – Résultats de la modélisation fréquence TGN habitation



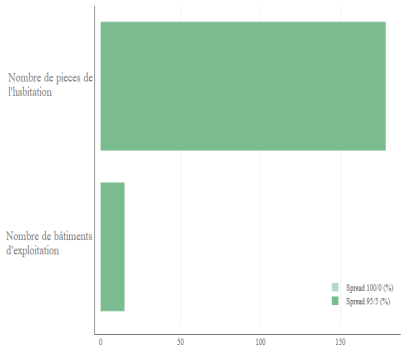

(a) Courbes de Lorenz sur base de validation : TGN habitation k-folds




(b) Résidus quantiles : fréquence TGN habitation




(c) *lift curve* : fréquence TGN habitation




(d) Variables significatives : fréquence TGN habitation

FIGURE 5.8 – Graphiques des résultats : modélisation fréquence TGN habitation

Comme observé dans le tableau 5.3, les résultats de ce modèle sont plutôt satisfaisants. L'indice Gini est assez stable entre la base d'apprentissage et la base de test. Toutefois, cet indicateur présente quelques écarts sur les sous-échantillons de la base de validation. La tendance normale observée sur les résidus quantiles (cf. figure 5.8) est limite acceptable. Néanmoins, la fréquence prédite décrit assez correctement la tendance de la fréquence observée même si des écarts entre elles sont visibles.

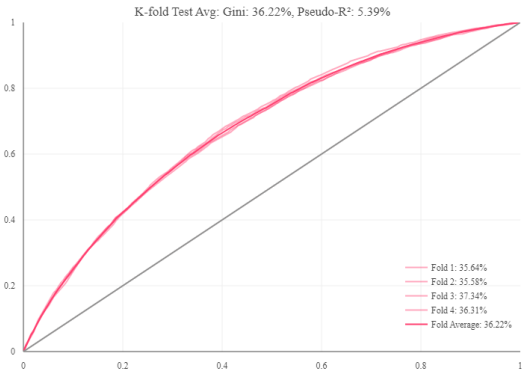
**Résumé de la modélisation :**
Le niveau de risque sur l'habitation est essentiellement identifié par le nombre de

pièces et de bâtiments d'exploitation. Le niveau de risque est croissant en fonction de ces deux caractéristiques. Cela s'expliquerait dans un premier temps par le fait qu'une grande surface soit plus exposée aux phénomènes climatiques. Ensuite, il est possible que le risque soit accentué pour les habitations situées au même endroit que des exploitations ayant plus de bâtiments d'exploitation.

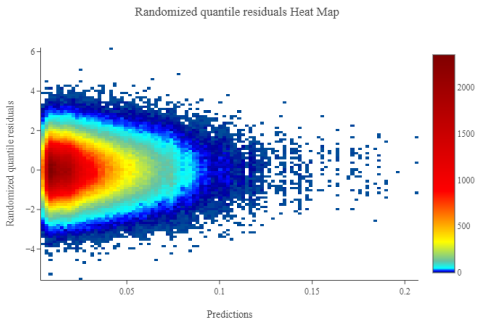
— **Exploitation**
Les performances du modèle sont les suivantes :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson | 0,193 | 36,59 % | 0,193 | 36,22 % |

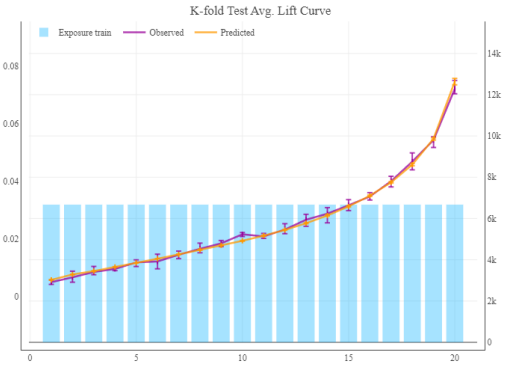
TABLE 5.4 – Résultats de la modélisation fréquence TGN exploitation



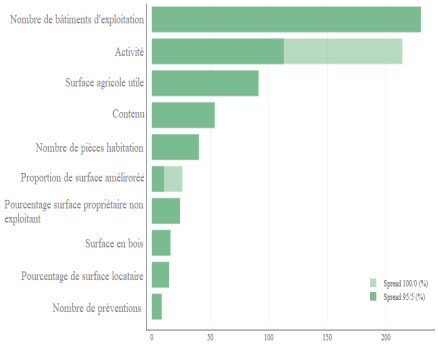

(a) Courbes de Lorenz sur base de validation : TGN exploitation k-folds




(b) Résidus quantiles : fréquence TGN exploitation




(c) *lift curve* : fréquence TGN exploitation




(d) Variables significatives : fréquence TGN exploitation

FIGURE 5.9 – Graphiques des résultats : modélisation fréquence TGN exploitation

La qualité du modèle de prédiction est très satisfaisante au regard de chacun des

critères de validation du modèle, présentés dans le tableau 5.4 et sur la figure 5.9.

**Résumé de la modélisation :**
Il est intéressant d'observer que la variable la plus significative du modèle n'est plus le contenu de l'exploitation (graphique 5.6) mais plutôt le nombre de bâtiments d'exploitation. Cela pourrait s'expliquer par le fait que la majeure partie des évènements liés à cette garantie ne sont couverts que sur le bâtiment d'exploitation. La fréquence de sinistres est d'autant plus importante que le nombre de bâtiments augmente. Par ailleurs, les assurés avec un pourcentage important de surface non exploitée sembleraient être plus risqués.

### 5.2.2 Modélisation par apprentissage statistique

L'intérêt dans cette section est de challenger les modèles de fréquence obtenus sur l'exploitation par des modèles d'apprentissage statistique. Le modèle utilisé pour le challenge est l'*eXtrême Gradient Boosting* (XGBoost).
Les paramètres optimaux (5.5) obtenus par validation croisée, avec comme critère le critère de stabilité de l'indice de Gini sur la base d'apprentissage et de validation, sont les suivants :

| Parmaètres | Incendie exploitation | TGN exploitation |
|---|---|---|
| nrounds | 150 | 200 |
| max_depth | 3 | 4 |
| colsample_bytree | 0,8 | 0,8 |
| eta | 0,050 | 0,050 |
| gamma | 0 | 0 |
| min_child_weight | 10 | 10 |
| subsample | 0,9 | 0,9 |

Table 5.5 – Paramètres XGBoost

#### Incendie exploitation

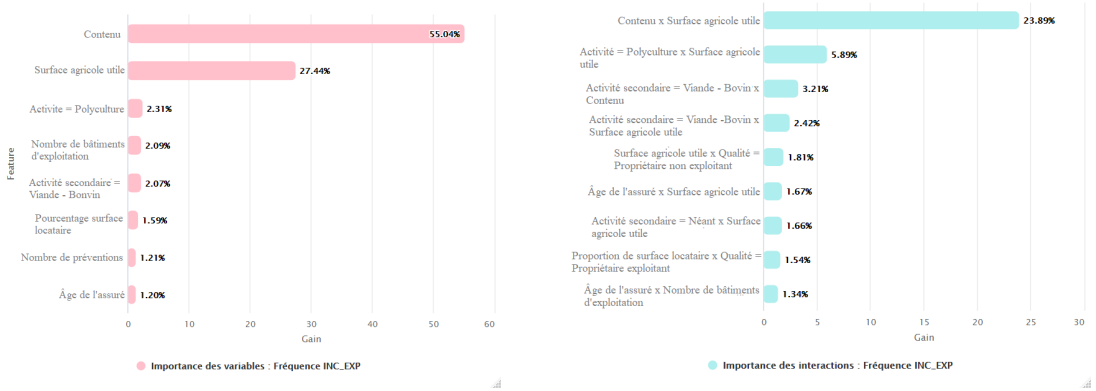
Les résultats de la calibration du modèle XGBoost avec les paramètres optimisés sur la fréquence incendie sont les suivants :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| XGBoost poisson | 0,099 | 45,41% | 0.098 | 44,23% |
| GLM Poisson | 0,118 | 49,08% | 0,118 | 48,44% |

Table 5.6 – Résultats de la modélisation : XGBoost fréquence incendie exploitation

Le modèle XGBoost obtenu présente une RMSE plus faible que le GLM (tableau 5.6). Toutefois, le modèle XGBoost a une capacité de segmentation bien inférieure au GLM. Pour mieux comprendre le modèle XGBoost une étude sur l'importance des variables est nécessaire.

*Importance des variables et interactions*




(a) Importance des variables (b) Importance des interactions

FIGURE 5.10 – XGBoost fréquence incendie exploitation

L'ordre de l'importance des variables dans la prédiction du modèle XGBoost (5.10-a) n'est pas en désaccord avec celui du GLM. Les variables présentées sont celles qui ont un niveau de contribution dans la prédiction supérieur à 1%. De plus, l'analyse des interactions (5.10-b) permet d'observer les combinaisons de variables qui contribuent le plus à l'amélioration de la prédiction. Il apparaît que la surface agricole utile, le contenu de l'exploitation, l'activité de polyculture et l'activité secondaire d'élevage de bovins contribuent le plus dans la prédiction.
En somme, la logique de prédiction du modèle XGBoost ne s'éloigne pas de celle du GLM.

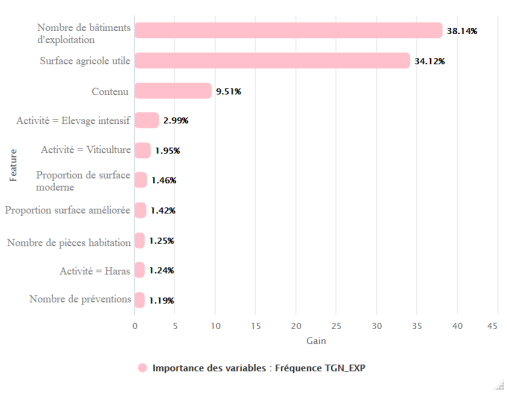
**TGN**

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| XGBoost poisson | 0,145 | 37,23% | 0,145 | 36,92% |
| GLM Poisson | 0,193 | 36,59 % | 0,193 | 36,22 % |

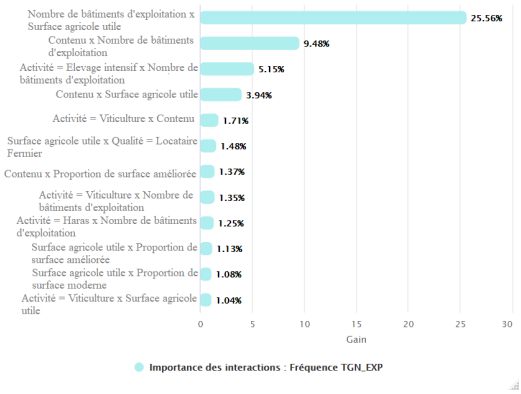
TABLE 5.7 – Résultats de la modélisation : XGBoost fréquence TGN exploitation

Le modèle XGBoost sur la garantie TGN présente de meilleures performances en termes de Gini et de RMSE que le GLM. Les indicateurs sont stables sur la base d'apprentissage et la base de validation.

Les résultats de l'analyse de l'importance des variables et des interactions dans la prédiction sont les suivants :




(a) Importance des variables




(b) Importance des interactions

FIGURE 5.11 – XGBoost fréquence TGN exploitation

La tendance observée sur l'importance des variables est la même que celle observée sur le GLM. Les combinaisons obtenues à partir du nombre de bâtiments d'exploitation, de la surface agricole utile, du contenu de l'exploitation et de l'activité d'élevage intensif apportent le plus d'amélioration dans la prédiction.

En conclusion, les résultats des modèles XGBoost sont cohérents avec ceux des GLM, au regard de l'importance des variables. Cependant, ils n'apportent pas un gain significatif en termes de performance par rapport aux modèles GLM obtenus. Par conséquent, les modèles GLM exploitation seront retenus pour étudier l'apport des données en *Open Data*.

## 5.3 Modèles de coût moyen

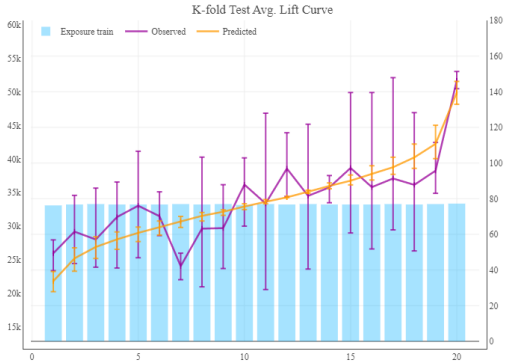
Comme évoqué dans la section relative à l'ajustement de lois 4.2.1, la modélisation du coût moyen s'effectuera avec une loi gamma.

### Incendie

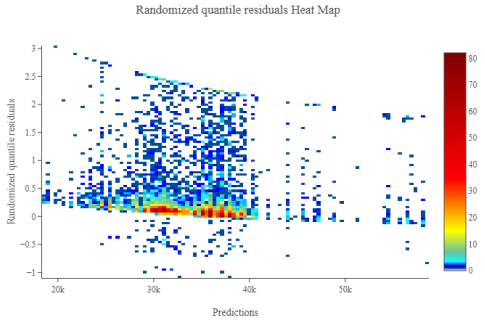
Les résultats obtenus sur la modélisation du coût moyen incendie sont les suivants :

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Gamma | 64150 | 13.22% | 64490 | 7.74% |

TABLE 5.8 – Résultats de la modélisation coût moyen incendie



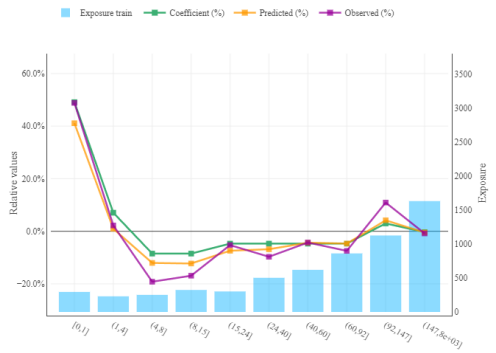

(a) *lift curve* : coût moyen incendie



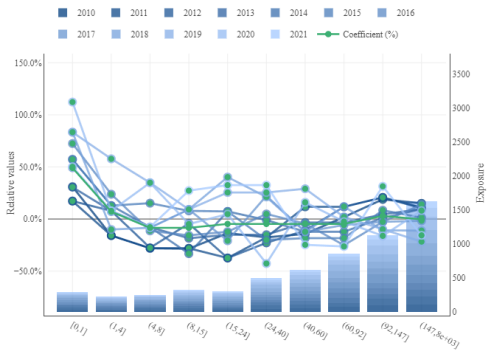

(b) Résidus quantiles : coût moyen incendie

FIGURE 5.12 – Graphiques des résultats : modélisation coût moyen incendie

Les métriques d'évaluation du modèle, présentées dans le tableau 5.8, ne sont pas stables entre la base d'apprentissage et la base de validation. Cet écart serait probablement lié à une faiblesse du volume de données pour calibrer un modèle stable.
Néanmoins, comme observé plus haut (figure 4.1-b), la loi gamma sous-estime le coût moyen des petits sinistres mais s'ajuste bien à la tendance observée (figure 5.12-a). Les résidus quantiles ne présentent pas une tendance mais sont compris entre -2 et 2 (graphique 5.12-b). La loi ne serait peut-être pas la plus adaptée sur ces données.




(a) Prédiction et observé : surface agricole utile




(b) Stabilité temporelle : surface agricole utile

FIGURE 5.13 – Stabilité temporelle et prédiction : coût moyen incendie

Toutefois, en analysant la stabilité temporelle du modèle sur la figure 5.13, le modèle est assez stable dans le temps et prédit bien la tendance observée sur les variables les plus importantes.

**Résumé de la modélisation :**
De manière générale, le coût moyen diminue quand la surface agricole utile augmente.

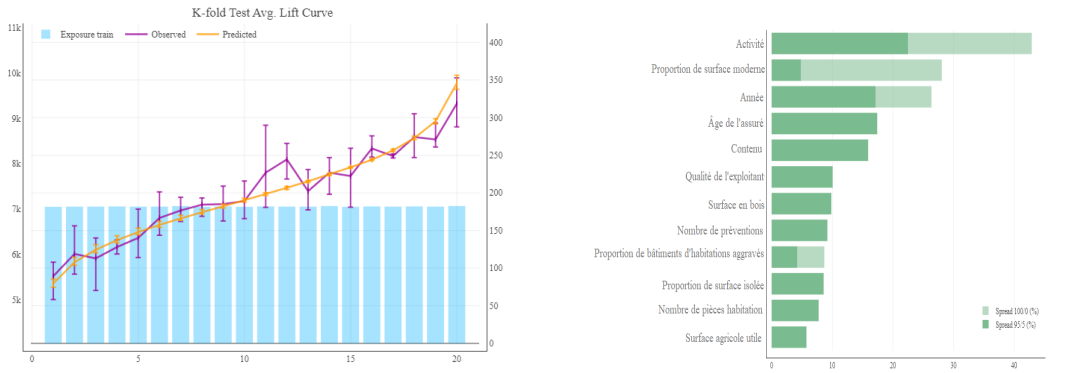
Le coût moyen est plus élevé sur les petites surfaces, ce qui pourrait s'expliquer par le fait que l'ampleur de l'incendie soit plus grand sur les petites surfaces. Par ailleurs, le coût moyen incendie augmente avec le nombre de préventions. L'hypothèse émise est que les exploitations où le risque est élevé sont celles avec le plus de préventions. De plus, cette croissance du coût moyen avec le nombre de préventions pourrait être le signal d'un incendie de forte intensité qui n'a pu être contrôlé par le nombre de préventions.

**TGN**

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Gamma | 7185 | 9,27% | 7219 | 7,62% |

TABLE 5.9 – Résultats de la modélisation coût moyen TGN

L'écart entre le Gini sur la base d'apprentissage et la base de validation est moins important pour cette garantie. Cela serait probablement lié au nombre de sinistres plus conséquent que sur l'incendie.




(a) *lift curve* : coût moyen TGN (b) Variables significatives : coût moyen TGN

FIGURE 5.14 – Graphiques des résultats : modélisation coût moyen TGN

Au regard de la *lift curve* 5.14-a, la tendance du coût moyen observé est bien ajustée par le modèle, à part une sur-estimation des gros sinistres.

**Résumé de la modélisation :**
L'information clé de la modélisation est que le coût moyen est significativement plus élevé pour les assurés exerçant l'activité d'élevage intensif. Le coût moyen est plus faible chez les assurés moins âgés, dû à une expérience acquise sur la gestion des intempéries. En outre, cela pourrait justifier la tendance observée sur la variable qualité où le coût moyen est plus faible chez les retraités et propriétaires non exploitants. Enfin, plus le coût moyen augmente, plus il y a de surface en bois, ce qui serait probablement lié à la fragilité des matériaux.

**A retenir** :

- **Modélisation de la fréquence de sinistres** :
  - Habitation : fréquence empirique en fonction du nombre de pièces et de la valeur du contenu de l'habitation pour la garantie incendie et GLM Poisson robuste et segmentant pour la garantie TGN ;
  - Exploitation : GLM Poisson robuste et segmentant pour la garantie incendie et la garantie TGN ;
- **Modélisation du coût moyen** : GLM gamma moins robuste mais qui décrit bien la tendance observée avec une stabilité temporelle assez correcte ;
- **Modélisation XGBoost de la fréquence exploitation** :
  - Les résultats des XGBoost sont cohérents avec ceux des GLM ;
  - Par rapport au GLM, le gain de précision et de segmentation avec le modèle XGBoost n'est pas significatif.

# Chapitre 6

# Apport de l'*Open Data* et zonier MRA

Ce chapitre aura pour objectif d'évaluer le gain de segmentation engendré par l'intégration des zoniers et des données en *Open Data* dans les modèles.

## 6.1 *Open Data*

Cette section vise à étudier l'apport des données en *Open Data* dans les modèles mis en place précédemment.

### 6.1.1 Modèles de fréquence

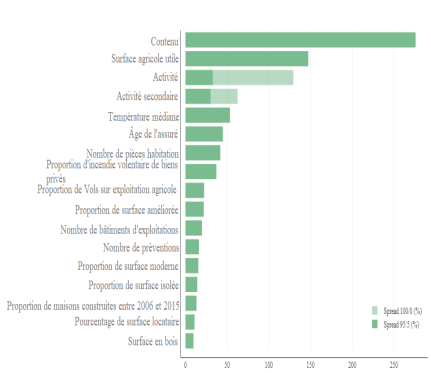
**Incendie : Fréquence exploitation**

L'ajout des données externes dans le modèle GLM réalisé précédemment a permis d'obtenir les performances suivantes :

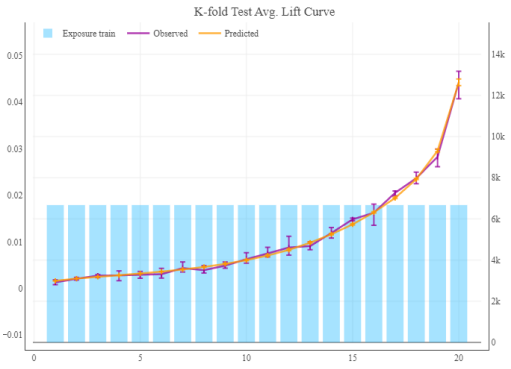
| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* | 0,118 | 49,08% | 0,118 | 48,44% |
| GLM poisson avec *Open Data* | 0,118 | 49,98% | 0,118 | 49,08% |

TABLE 6.1 – Résultats de la modélisation fréquence incendie exploitation avec *Open Data*

Les données en *Open Data* ont permis d'améliorer la qualité de segmentation du modèle obtenu précédemment, comme observé dans le tableau 6.1 et sur la figure 6.1-b. Le nouvel ordre d'importance des variables est le suivant, après l'ajout des données en *Open Data* (6.1-a) :




(a) Variables significatives : fréquence incendie exploitation *Open Data*




(b) *lift curve* : fréquence incendie exploitation *Open Data*

FIGURE 6.1 – *Open Data* : fréquence incendie exploitation

**Résumé de la modélisation :**
Il ressort de cette modélisation que le risque incendie exploitation augmente significativement dans les zones Insee où plus de la moitié des températures observées sur l'année sont supérieures à 13°C. La fréquence incendie augmente avec la proportion d'incendies volontaires de biens privés et la proportion de vols sur les exploitations agricoles. Pour rappel, il s'agit d'informations sur la criminalité de l'année $n-1$. Enfin, le risque incendie diminue lorsque la proportion de maisons construites entre 2006 et 2015 dans la zone Insee augmente. Le prix au mètre carré n'a pas été retenu comme variable significative car l'information apportée par la variable était obtenue à partir des autres variables.

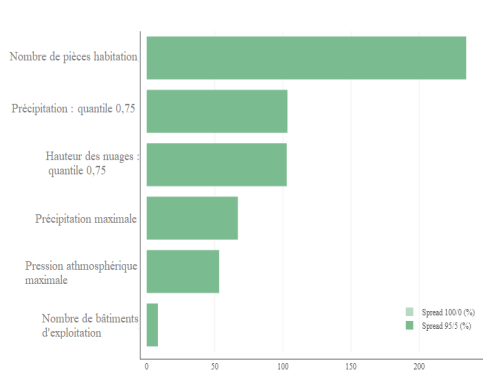
### TGN : Fréquence habitation

Après intégration des données *Open Data* dans le GLM les performances suivantes sont obtenues :

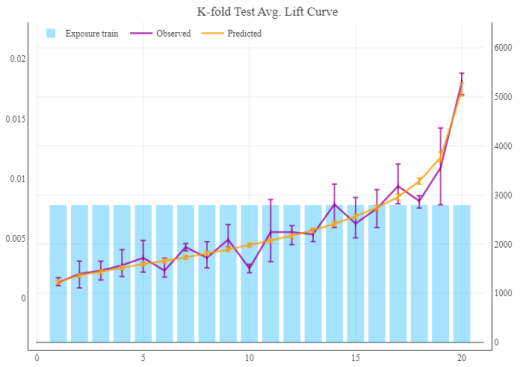
| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* | 0,089 | 21,71 % | 0,089 | 21,23 % |
| GLM Poisson avec *Open Data* | 0,089 | 36,71 % | 0,089 | 33,64 % |

TABLE 6.2 – Résultats de la modélisation fréquence TGN habitation avec *Open Data*

La qualité de segmentation du modèle, observable dans le tableau 6.2, a été nettement améliorée avec la présence de variables en *Open Data* (6.2-a). Bien que l'indice de Gini soit moins stable entre la base d'apprentissage et la base de validation, il est possible d'observer une amélioration significative de la *lift curve* (6.2-b). La fréquence prédite s'est rapprochée de la fréquence observée sur la base de validation avec l'ajout de la hauteur des nuages, de la pression atmosphérique et des précipitations.




(a) Variables significatives : fréquence TGN habitation *Open Data*




(b) *lift curve* : fréquence TGN habitation *Open Data*

FIGURE 6.2 – Open Data : fréquence TGN habitation

**Résumé de la modélisation :**
Il ressort de l'analyse de ces variables que la fréquence d'évènements climatiques diminue avec la hauteur des nuages. La métrique qui a été utilisée est le quantile 0,75 de la hauteur observée des nuages sur l'année. Cette valeur permet de connaître le niveau globalement atteint par la hauteur des nuages dans chaque zone. La hauteur des nuages joue un rôle particulier dans le climat. Les nuages bas sont des nuages qui ne contiennent pas beaucoup d'eau et qui ont un niveau de réflectivité des rayons du soleil très élevé. Ce niveau de réflectivité élevé participe au réchauffement de la surface et joue un rôle dans la formation de la grêle. En outre, les nuages hauts, remplis de précipitations en phase liquide, annoncent des perturbations et des orages. L'interprétation qui pourrait en résulter est que dans les zones avec de fortes précipitations, les habitations ont une structure adéquate pour résister aux phénomènes orageux. Cette interprétation se justifie avec le niveau de la fréquence d'évènements climatiques qui est plus faible dans les zones avec des précipitations élevées au regard de la précipitation maximale sur l'année et du quantile 0,75. De même, la fréquence augmente significativement dans les zones où la pression atmosphérique atteint 1040hPa. Or, les zones à haute pression sont des endroits avec de fortes températures.

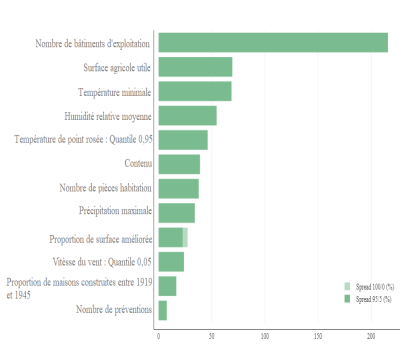
### TGN : Fréquence exploitation

Les performances suivantes sont obtenues en intégrant les données *Open Data* dans la modélisation :

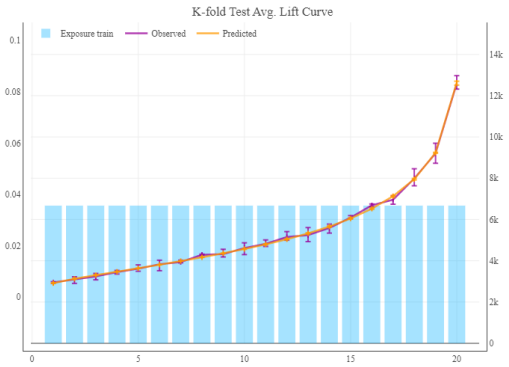
| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* | 0,193 | 36,59 % | 0,193 | 36,22 % |
| GLM poisson avec *Open Data* | 0,193 | 38,7% | 0,193 | 38,14% |

TABLE 6.3 – Résultats de la modélisation fréquence TGN exploitation avec *Open Data*

L'ajout des données en *Open Data* a permis d'améliorer la qualité de segmentation du modèle avec un gain en Gini de deux points, comme observé dans le tableau 6.3. Il est possible d'observer une amélioration de l'ajustement de la fréquence prédite relativement à la fréquence observée sur la *lift curve* 6.3-b.




(a) Variables significatives : fréquence TGN exploitation *Open Data*




(b) *lift curve* : fréquence TGN exploitation *Open Data*

FIGURE 6.3 – *Open Data* : fréquence TGN exploitation

Cette amélioration du modèle est liée à l'ajout de variables telles que la température minimale observée, l'humidité relative moyenne, la température de point de rosée, la précipitation maximale, la vitesse du vent et la proportion de maisons construites de 1919 à 1945 (6.3-a).

**Résumé de la modélisation :**
Il en ressort que la fréquence TGN est significativement plus élevée dans les zones où des températures minimales extrêmement basses sont observées (inférieures à -8°C). De plus, les zones avec une plus grande humidité relative et une température de point de rosée faible ont une plus grande fréquence, ce qui soutient l'hypothèse d'une fréquence plus élevée sur le risque exploitation dans les régions à basses températures. Par ailleurs, plus le quantile 5% de la vitesse du vent sur l'année augmente, plus la fréquence observée augmente. Il en est de même avec la proportion de maisons construites entre 1919 et 1945. Enfin, la fréquence diminue avec le niveau maximal de précipitations observé sur l'année, ce qui pourrait être lié à la résilience des assurés face au niveau élevé de précipitations.

### 6.1.2 Coût moyen

De manière générale, les données en *Open Data* ne présentent pas de véritable tendance discriminante apparente pour le coût moyen des deux garanties. Néanmoins, l'ajout de ces données permet d'améliorer la qualité de segmentation globale du modèle.
Le modèle relatif à la garantie incendie avec la variable distance au centre d'incendie est présenté brièvement.

#### Coût incendie

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Gamma sans *Open Data* | 64150 | 13,22% | 64490 | 7,74% |
| GLM Gamma avec *Open Data* | 64130 | 14,44% | 64370 | 9,41% |

TABLE 6.4 – Résultats GLM gamma coût moyen *Open Data*

L'ajout de la distance à la caserne de pompier la plus proche a permis d'améliorer la qualité de segmentation du modèle de coût moyen incendie. Bien qu'il y ait une phase de croissance du coût moyen avec les exploitations situées entre 0 et 4 kilomètres d'une caserne, cette variable n'est pas assez discriminante. En effet, il n'y a pas de tendance globale observée sur cette variable. Cette absence de tendance pourrait être liée au fait qu'il s'agit de la distance observée en 2022 et non celle au moment du sinistre.

## 6.2 Construction des zoniers

Dans cette étude, l'intérêt de la construction d'un zonier est qu'il pourrait améliorer la capacité de prédiction et de segmentation du modèle (6.2.2).
Il existe plusieurs méthodes de réalisation de zoniers. L'approche développée dans ce mémoire a été retenue en raison de contraintes opérationnelles. La maille géographique considérée pour la construction du zonier est l'Insee.
Les zoniers sont construits à partir des résidus de chaque modèle. Ces zoniers ainsi construits sont réinjectés dans leurs modèles respectifs comme de nouvelles variables.
Du fait de la différence de la nature des risques modélisés au sein de chaque garantie, il est nécessaire de construire un zonier pour chaque garantie.

### 6.2.1 Contexte : Intérêt des nouveaux zoniers

Des études en interne ont permis d'observer que le zonier actuel ne réussit plus à segmenter le risque correctement. Cette difficulté de segmentation serait liée à sa structure et à son ancienneté [1]. Le zonier actuel est un zonier à la maille département qui dépend de l'activité. De plus, ce zonier est un zonier global sans distinction de garantie.

1. Le zonier actuel date du début des années 2000.

Néanmoins, les coefficients du zonier habitation et exploitation sont différents.

La solution à court terme adoptée est basée sur l'utilisation des ratios S/P [2] 10 ans et 3 ans par département et par activité. Ces ratios sont utilisés pour actualiser les coefficients des zoniers habitation et exploitation. Cependant, cette méthode conduit à classer les risques en deux zones : rouge et verte.

Cette méthode ne peut être intégrée dans la nouvelle structure de modélisation de la prime pure par garantie. Ainsi, il est nécessaire de construire un nouveau zonier adapté à chaque garantie.

### 6.2.2 Approche bayesienne

#### Théorie

Définition : Modèle Bayésien
Pour une variable aléatoire (ou une suite de variables aléatoires), un modèle Bayésien est la donnée d'une loi conditionnelle et d'une loi a priori :

$$X \sim f(X|\theta)$$

$$\theta \sim \pi$$

Dans la modélisation Bayésienne, il est possible de calculer une loi a posteriori sur $\theta$. Cette loi est la loi de $\theta$ conditionnellement aux valeurs $X$.

Définition : Loi a posteriori, cas continu
La loi a posteriori est la loi dont la densité est donnée par :

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int_{u\in\Theta} f(X|u)\pi(u)du}$$

Définition : Loi marginale
La loi marginale est la loi définie par :

$$m_\pi(X) = \int_{u\in\Theta} f(X|u)\pi(u)du$$

La loi marginale ne dépend pas du paramètre $\theta$ mais uniquement de la loi de $X$ et de la loi a priori. Elle joue le rôle de constante de normalisation de la loi a posteriori.
Par conséquent, la loi marginale est inutile dans la maximisation de la loi a posteriori, le calcul peut s'effectuer à une constante multiplicative près. Ainsi, la notation suivante est adoptée :

$$\pi(\theta|X) \propto f(X|\theta)\pi(\theta)$$

2. Ratio S/P = Somme de sinistres/Somme des primes

**Application à la création d'un zonier**

L'idée dans la création d'un zonier est de supposer qu'il existerait des facteurs géographiques qui ne sauraient être expliqués par le modèle. En d'autres termes, l'erreur de prédiction se décomposerait en une partie liée à localisation de l'assuré et un aléa. La formulation mathématique est la suivante :

$$\begin{aligned} g(Y_i) - \hat{\beta} X_i &= R_i \\ &= \textit{facteur géographique} + \textit{aléa} \end{aligned}$$

Avec : $Y_i$ la valeur observée sur le risque $i$, $\hat{\beta}$ les coefficients du modèle GLM, $X_i$ les variables explicatives du risque $i$, $R_i$ le résidu associé au risque $i$, et $g$ la fontion de lien du modèle GLM.
Dans cette approche bayésienne, les hypothèses suivantes sont émises :

- Le facteur géographique est considéré comme une variable aléatoire ;
- La loi a priori du vecteur aléatoire des facteurs géographiques $H$ est supposée la loi d'un vecteur gaussien centré ;
- L'aléa $\epsilon$ est un bruit gaussien centré.

En utilisant le modèle GLM obtenu sur la fréquence, le nombre de sinistres prédits pour un Insee est : $E_{Insee}\ exp(\beta\, X_{Insee})$, avec $E_{Insee}$ l'exposition de l'Insee. L'écriture de l'équation du modèle GLM à la maille Insee est la suivante :

$$Y_i = E_i\ exp(\beta X_i + h_i + \epsilon_i)$$

Avec $h$ une réalisation de $H$.

Notations importantes :

- $\sigma_H^2$ : variance du risque global au niveau de la carte ;
- $D_{i,j}$ : distance entre la zone Insee $i$ et $j$ ;
- $H_i$ : variable aléatoire gaussienne du risque géographique pour l'Insee $i$ avec $H_i \sim \mathcal{N}(0,\, \sigma_h^2)$
- $K$ : matrice de corrélations des facteurs géographiques $H_i$, avec :

$$\begin{aligned} K_{i,j} &= \frac{Cov(H_i, H_j)}{\sigma_H^2} \\ &= \begin{cases} 1 & \text{si } i{=}j \\ \dfrac{exp(-\frac{D_{i,j}^2}{2\sigma_H^2})}{\sigma_H^2} & \text{sinon} \end{cases} \end{aligned}$$

- Le bruit $\epsilon_i \sim \mathcal{N}(0,\, \frac{\sigma^2}{E_i})$ avec $E_i$ l'exposition de l'Insee $i$ ;
- $W$ : matrice diagonale des expositions de chaque code Insee ;
- $f_{R|H}(r|h)$ : une densité gaussienne de moyenne h et $f_H(h)$ une densité gaussienne centrée.

L'objectif est d'obtenir une estimation optimale $h$. Il faudra déterminer la valeur $\hat{h}^* = (\hat{h}_i^*)_{i\in[|1,p|]}$ qui maximise la loi posteriori $\pi_{H|R}(h|r)$ .

$$\pi_{H|R}(h|r) \propto f_{R|H}(r|h) f_H(h)$$

Il s'agit de maximiser la quantité de gauche par la méthode du maximum de vraisemblance. En utilisant la log-vraisemblance, cela revient à :

$$Log(L(H|R)) \approx Log(L(h)) + Log(L(R|H))$$

La log-vraisemblance obtenue à partir de $f_H(h)$ est la suivante :

$$Log(L(h)) \approx -{}^tH \frac{1}{\sigma_H^2} K^{-1} H$$

Ensuite, la log-vraisemblance obtenue à partir de $f_{R|H}(r|h)$ s'exprime de la manière suivante :

$$Log(L(R|H)) \approx -\frac{1}{\sigma^2} {}^t(H-R) W (H-R)$$

Finalement, la log-vraisemblance totale à maximiser est la suivante :

$$Log(L(H|R)) \approx -{}^tH \frac{1}{\sigma_H^2} K^{-1} H - \frac{1}{\sigma^2} {}^t(H-R) W (H-R)$$

En pratique, ce problème de maximisation est résolu à partir des équations normales comme dans l'estimation des coefficients d'une régression pénalisée.
La solution théorique est la suivante :

$$\hat{h}^* = \left( KW + \frac{\sigma^2}{\sigma_H^2} Id \right)^{-1} KW * R$$

Le rapport $\frac{\sigma_2}{\sigma_H^2}$ définit le paramètre de lissage des coefficients des facteurs géographiques. Plus ce rapport est petit, moins les coefficients sont lissés. Plus ce rapport est grand, plus le lissage est important.

— Le choix d'un $\sigma_H^2$ petit conduit $\hat{h}^* \to 0$ ;
— Le choix d'un $\sigma^2$ petit conduit $\hat{h}^* \to R$.

L'ajustement du paramètre de lissage s'effectue en tenant compte de la distance (ou indice) Moran. Cette distance peut être interprétée comme la distance moyenne à parcourir pour observer un changement important des coefficients des facteurs géographiques. Enfin, le niveau de lissage optimal est celui qui permet d'améliorer au mieux le Gini du modèle sans sur-apprentissage [3].

L'étape finale est la création de zones géographiques de sorte que toutes les zones Insee appartenant à une zone aient le même coefficient. En général, des méthodes de classification non supervisées sont utilisées pour réaliser cette opération de regroupement des

3. Un Gini assez stable sur la base d'apprentissage et la base de validation

coefficients. Toutefois, il existe une approche qui consiste à utiliser un arbre à décision pour faire ce regroupement. Cette méthode consiste à prédire les coefficients des facteurs géographiques à partir d'eux-mêmes avec un arbre de décision. Bien évidemment, il s'agit d'un arbre de régression avec une profondeur qui est ajustée en fonction du nombre de zones désirées. Finalement, les noeuds terminaux obtenus contiennent des groupes de coefficients de facteurs géographiques. Chaque groupe correspond à une zone et le coefficient associé à la zone est la moyenne des coefficients des facteurs géographiques pondérée par l'exposition de chaque zone Insee.

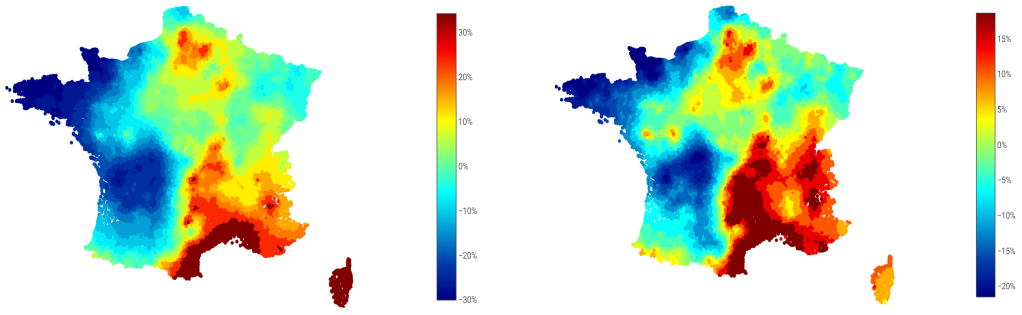
### 6.2.3 Application

L'objectif dans cette section est de pouvoir expliquer le risque géographique et de mesurer l'apport de l'*Open Data* dans notre modélisation.

A partir des résultats du chapitre précédent, les modèles avec des données *Open Data* présentaient de meilleurs résultats que les modèles sans *Open Data*. Ce gain en termes de qualité dans les modèles est en partie dû au fait que les données *Open Data* contiennent une partie de l'information contenue dans les facteurs géographiques. Dans ce contexte, il est intéressant de comparer un modèle utilisant un zonier construit avec un modèle sans *Open Data* contre un modèle utilisant un zonier construit avec un modèle avec *Open Data*. Cette confrontation permettra de mesurer l'intérêt des données *Open Data* dans la modélisation du risque.
Les zoniers seront construits à partir des modèles de fréquence exploitation. Le territoire est segmenté en 20 zones géographiques avec l'ajustement du lissage qui permet d'avoir un modèle avec de meilleures performances.

#### Incendie

En appliquant la théorie bayésienne, les zoniers suivants sont obtenus avec les modèles avec et sans *Open Data*.




(a) Zonier issu du modèle sans *Open Data* (b) Zonier issu du modèle avec *Open Data*

FIGURE 6.4 – Zonier exploitation incendie

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* | 0,118 | 49,08% | 0,118 | 48,44% |
| GLM Poisson avec *Open Data* | 0,118 | 49,98% | 0,118 | 49,08% |
| GLM Poisson sans *Open Data* + zonier | 0,117 | 51,63% | 0,117 | 49,84% |
| GLM Poisson avec *Open Data* + zonier | 0,117 | 51,85% | 0,117 | 49,85% |

TABLE 6.5 – Résultats à la suite de l'ajout du zonier - exploitation incendie

**Analyse des zoniers** :
Les zones à très fortes températures présentent un risque incendie exploitation plus grand que les zones à basses températures. En effet, le risque géographique incendie est très élevé dans le sud-ouest de la France (Corse, Languedoc-Roussillon, Provence-Alpes-Côte d'Azur). Ce risque peut être aussi considéré comme élevé en Auvergne-Rhône-Alpes et moyen au centre. En outre, le risque incendie lié au risque des facteurs géographiques diminue considérablement en Bretagne et en Aquitaine. Cette tendance s'observe sur les deux zoniers.
Il est important de faire une distinction entre les facteurs géographiques représentés par le zonier issu du modèle sans *Open Data* et celui avec des données *Open data*. Dans la modélisation sans données *Open Data*, le zonier obtenu représente la totalité du risque (ou effet) géographique ; tandis que, dans la modélisation avec des données *Open Data*, le zonier ne représente que le risque géographique résiduel. En effet, les variables *Open Data* contiennent déjà une partie de l'information géographique. C'est pour cette raison que les coefficients des facteurs géographiques sont plus élevés dans le modèle sans données *Open Data* que dans le modèle avec données *Open Data*.

Concernant les performances du modèle, l'apport du zonier a permis une augmentation du Gini d'environ deux points sur le modèle sans données *Open Data* et un point sur le modèle avec les données *Open Data* (car effet géographique résiduel). En comparant les deux modèles, le modèle sans données *Open Data* combiné au zonier donne des performances quasi-similaires à celui avec des données *Open Data*. L'apport final des données *Open Data* est assez minime et ces données ne serviraient qu'à mieux interpréter de les effets géographiques.


(a) Bêta zonier issu du modèle sans *Open Data*

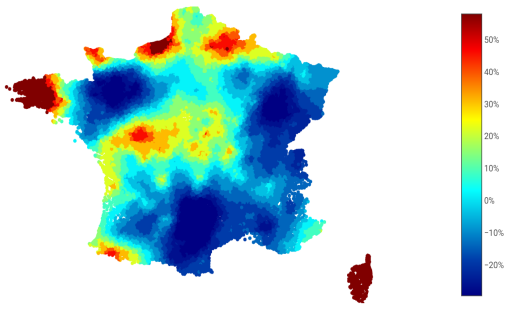

(b) Bêta zonier issu du modèle avec *Open Data*

FIGURE 6.5 – Bêta zonier exploitation incendie

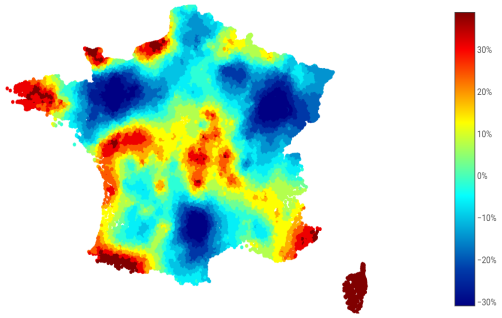
La figure 6.5 permet d'observer une croissance des coefficients du modèle en fonction des zones.

### TGN

Un zonier habitation TGN n'a pas pu être construit à cause du faible taux de couverture [4] par commune. Ce faible taux de couverture conduirait à l'obtention d'un zonier trop lissé. Par conséquent, le zonier TGN n'a pu être construit que sur l'exploitation. Les zoniers obtenus sont les suivants :



(a) Zonier issu du modèle sans *Open Data*



(b) Zonier issu du modèle avec *Open Data*

FIGURE 6.6 – Zonier exploitation TGN

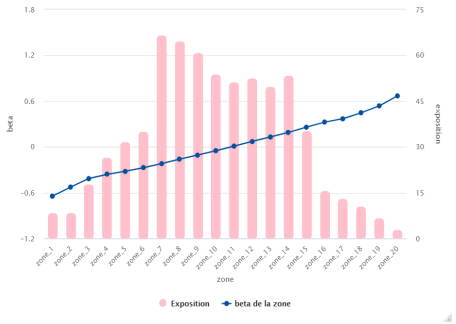
4. Nombre de communes Insee dans la base de modélisation rapporté au nombre de zones Insee en France.

| | RMSE train | Gini train | RMSE test | Gini test |
|---|---|---|---|---|
| GLM Poisson sans *Open Data* | 0,193 | 36,59 % | 0,193 | 36,22 % |
| GLM poisson avec *Open Data* | 0,193 | 38,7% | 0,193 | 38,14% |
| GLM Poisson sans *Open Data* + zonier | 0,193 | 42,24% | 0,193 | 40,49% |
| GLM Poisson avec *Open Data* + zonier | 0,193 | 43,16% | 0,193 | 41,25% |

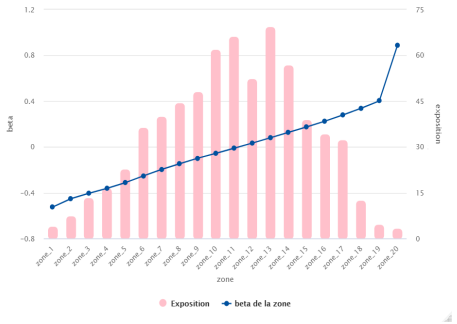
TABLE 6.6 – Résultats à la suite de l'ajout du zonier - exploitation TGN

**Analyse des zoniers**
Le risque géographique d'évènements climatiques est très faible dans les Pays de la Loire, la Lorraine et la Languedoc-Roussilon sur les deux zoniers. Par contre, le risque géographique est très élevé en Corse, Bretagne, dans le nord de la Basse-Normandie et Haute-Normandie.
A l'échelle du risque géographique résiduel, il est possible d'observer un risque géographique plus élevé dans le sud de l'Aquitaine, du Midi-Pyrénnées et Provence-Alpes-Cote d'Azur. De même, le risque géographique résiduel est plus élevé dans le centre et le centre-est.

Au regard des performances des deux modèles avec zonier, les *Open Data* permettent non seulement d'améliorer la qualité de segmentation mais apportent également une interprétabilité du risque géographique.




(a) Bêta zonier issu du modèle sans *Open Data*




(b) Bêta zonier issu du modèle avec *Open Data*

FIGURE 6.7 – Bêta zonier exploitation TGN

**A retenir** :
- **Contrainte** : zonier construit uniquement sur l'exploitation car faible taux de couverture sur l'habitation ;
- **Description des zoniers** :
  - Approche Bayésienne ;
  - A partir des résidus de chaque modèle, les zoniers sont construits puis ils sont réinjectés dans les modèles ;

  - Deux zoniers sont construits, un premier à partir du modèle sans données *Open Data* et avec des données *Open Data* ;
- **Intérêt des nouveaux zoniers** :
  - Segmentation : amélioration de la segmentation des modèles ;
  - Distinction par garantie : apporte une segmentation plus fine du risque en fonction de la garantie, par exemple la Bretagne est très peu risquée (zone bleue) pour l'incendie, mais très risquée (zone rouge) pour la garantie TNG ;
  - Par rapport à l'ancien zonier : corrige le biais sur l'ancien zonier dû au fait qu'il soit calibré sur toutes les garanties ;
- **Comparaison des modèles** : après avoir intégré les zoniers construits dans leurs modèles respectifs, l'écart de segmentation entre les modèles est de l'ordre de 1% à 2% ;
- **Apports principaux des données *Open Data*** :
  - Modélisation fréquence avec *Open Data* : amélioration de la segmentation des modèles avec intégration des données *Open Data* ;
  - Modélisation du coût moyen avec *Open Data* : absence de tendances discriminantes sur les données *Open Data* ;
  - Après intégration des zoniers : interprétation du risque géographique.

# Chapitre 7

# Synthèse de la modélisation

L'objectif final de la création des différents modèles calibrés est la mise en place d'un modèle de prime pure. Cette prime pure peut être considérée comme le juste prix que devrait payer l'assuré par rapport à son risque. La prime pure modélisée dans ce mémoire est le produit de la sortie du modèle de coût et de fréquence. La prime pure prédite à partir des modèles de fréquence et de coût doit être en cohérence avec la prime pure observée.
En pratique, la *lift curve* est utilisée pour vérifier l'adéquation entre prime pure observée et prime pure prédite. Bien évidement il faudra s'assurer que les métriques sur l'erreur de prédiction n'explosent pas. Enfin, la prime pure prédite permettra de mesurer le réel apport des données *Open Data*. Les modèles de coût moyen sans *Open Data* seront utilisés dans ce chapitre.

## 7.1 Prime pure incendie

### 7.1.1 Habitation

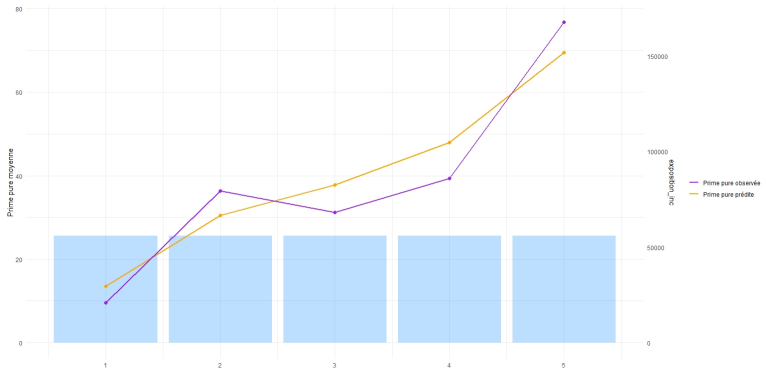



FIGURE 7.1 – *Lift curve* - Prime pure habitation incendie

La *lift curve* 7.1 obtenue en utilisant le modèle de coût moyen incendie et la table de fréquence habitation est plutôt satisfaisante. La prime pure prédite décrit assez bien la tendance de la prime pure observée.

| | Erreur globale | GINI |
|---|---|---|
| Prime pure prédite | 5% | 27,48% |

TABLE 7.1 – Indicateurs de performances - Prime pure incendie habitation

L'évaluation du Gini et de la *lift curve* du modèle permettent de voir que le modèle de prime pure obtenu est segmentant et qu'il possède une assez bonne qualité de prédiction globale. Par ailleurs, la prime pure prédite sur-estime la prime pure observée globale de 5%, ce qui est assez rassurant. Le pouvoir de segmentation du modèle obtenu peut s'observer avec la courbe de Lorenz.

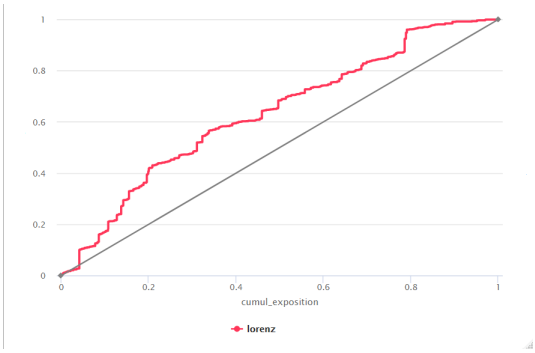
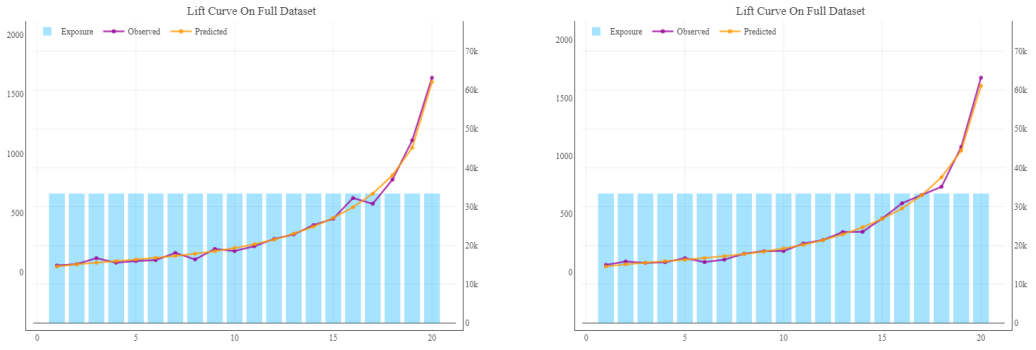



FIGURE 7.2 – Courbe de Lorenz - Prime pure habitation incendie

### 7.1.2 Exploitation

Les *lift curve* des primes pures prédites sont les suivantes :




(a) Prime pure sans *Open Data* (b) Prime pure avec *Open Data*

FIGURE 7.3 – *Lift curve* - Prime pure incendie exploitation

Les deux graphiques 7.3 permettent d'observer que les modèles obtenus réussissent à prédire une prime pure cohérente avec la prime pure observée. De plus, les erreurs globales de prédiction, présentées dans le tableau 7.2, sont assez faibles. Les primes pures prédites sous-estiment la prime pure observée globale de 7%. La capacité de segmentation des modèles est assez significative.

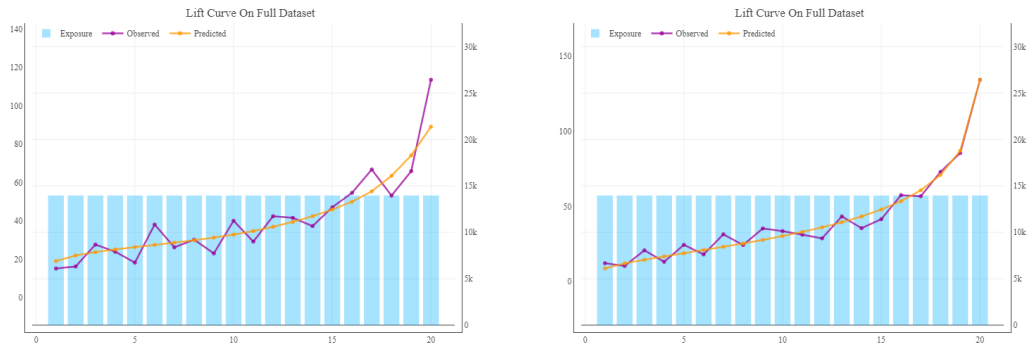
| | Erreur globale | Gini |
|---|---|---|
| Prime pure avec *Open Data* | -7% | 50,7% |
| Prime pure sans*Open Data* | -7% | 50,7% |

TABLE 7.2 – Indicateurs de performances - Prime pure incendie exploitation

Au regard de ces résultats, les modèles avec et sans *Open Data* ont des performances de prédiction et de segmentation équivalentes. Par conséquent, il est possible de se passer des coûts liés à la mise à jour et l'alimentation des données en *Open Data* ; sans que le modèle ne perde en qualité. Toutefois, l'apport des données en *Open Data* ne saurait être négligeable en termes d'interprétation et de quantification des effets géographiques.

## 7.2 Prime pure TGN

### 7.2.1 Habitation




(a) Prime pure sans *Open Data* (b) Prime pure avec *Open Data*

FIGURE 7.4 – *Lift curve* - Prime pure TGN habitation

La prime pure prédite avec le modèle de fréquence obtenu avec les données *Open Data* s'ajuste mieux à la prime pure observée. Cette différence est due à l'écart de Gini des deux modèles de fréquence. La prime pure prédite avec les données en *Open Data* permet d'obtenir une meilleur segmentation du risque. Néanmoins, les métriques d'évaluation de l'erreur de prédiction des deux modèles ne sont pas significativement éloignées, comme observé dans le tableau 7.3.

| | Erreur globale | Gini |
|---|---|---|
| Prime pure avec *Open Data* | -15% | 34,79% |
| Prime pure sans *Open Data* | -16% | 26,74% |

TABLE 7.3 – Indicateurs de performances - Prime pure TGN habitation

Une méthode pour améliorer la prime pure modélisée par le modèle sans *Open Data* est l'ajustement de l'intercept (ou niveau de base, $e^{\beta_0}$) du modèle. Cette méthode consiste à calculer un nouveau $e^{\beta_0}{}'$.

$$e^{\beta_0'} = e^{\beta_0} \frac{Somme\ prime\ pure\ observée}{Somme\ prime\ pure\ prédite}$$

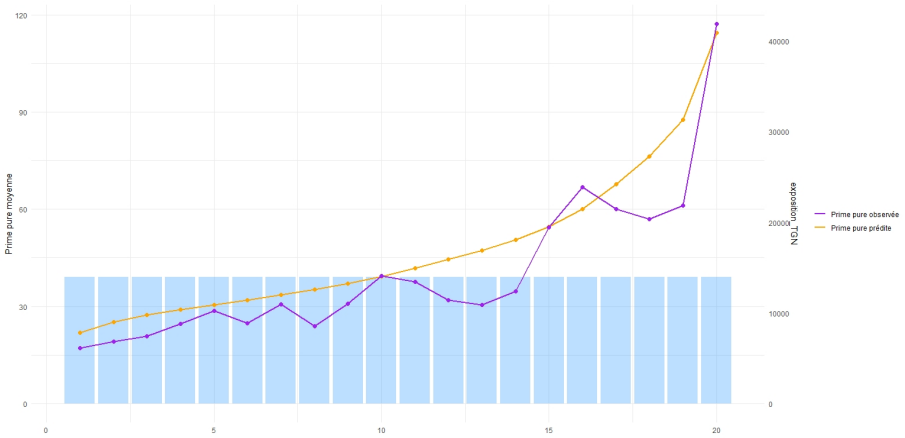
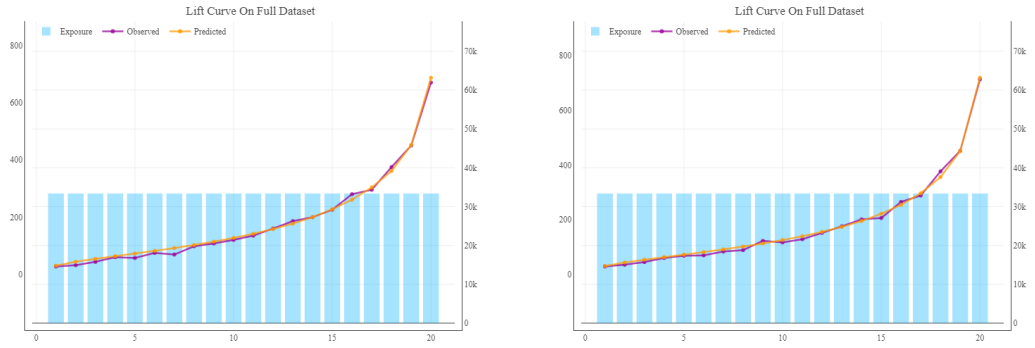



FIGURE 7.5 – *Lift curve* - Prime pure habitation TGN habitation avec ajustement

La nouvelle prime pure prédite surestime la valeur des petites primes pures observées. Cependant, elle permet d'avoir une meilleure prédiction des grosses primes.

En somme, il apparaît que, pour cette garantie, les variables *Open Data* présentent un intérêt majeur dans la qualité de prédiction du modèle.

### 7.2.2 Exploitation




(a) Prime pure sans *Open Data* (b) Prime pure avec *Open Data*

FIGURE 7.6 – *Lift curve* - Prime pure TGN exploitation

Concernant la modélisation du risque exploitation, les graphiques et les métriques permettent d'observer que les modèles avec et sans données *Open Data* sont équivalents.

| | Erreur globale | Gini |
|---|---|---|
| Prime pure avec *Open Data* | -14% | 44,7% |
| Prime pure sans *Open Data* | -14% | 44,5% |

TABLE 7.4 – Indicateurs de performances - Prime pure TGN exploitation

L'intérêt des données *Open Data* concerne l'interprétabilité d'une partie de l'effet géographique. De plus, au regard des enjeux climatiques, la présence de variables pour mesurer l'intensité du risque climatique est un atout non négligeable. En effet, l'analyse des tendance du modèle a permis d'observer la capacité des variables *Open Data* à segmenter le risque climatique.

L'erreur globale négative observée sur la plupart des modèles est liée à l'ajustement de la loi gamma, qui s'ajuste bien à la tendance du coût moyen mais la sous-estime (graphiques 4.3 et 4.1). Par conséquent, avant toute mise en production des modèles obtenus, un ajustement du niveau de base est effectué pour améliorer l'erreur globale ; comme ce qui a été réalisé pour la prime pure habitation TGN.

**A retenir** : Les modèles de prime pure obtenus ont une bonne capacité de prédiction et de segmentation.

# Conclusion

L'étude réalisée au travers de ce mémoire a permis de mieux comprendre le produit multirisque agricole d'AXA France. L'analyse des résultats obtenus permet d'identifier les facteurs influençant la sinistralité liée aux garanties incendie et TGN sur les risques habitation et exploitation agricole. La démarche utilisée dans la modélisation met en exergue la pertinence d'une modélisation distincte du risque habitation et exploitation. D'une part, cette démarche a abouti sur le fait qu'une modélisation distincte de la fréquence exploitation et habitation est indispensable. D'autre part, celle-ci démontre qu'une modélisation unique du coût moyen serait plus opérationnelle.

Ensuite, l'étude sur l'apport des données *Open Data* a montré que le gain de segmentation lié à l'utilisation de ces variables est différent en fonction de la garantie. En effet, bien que les variables *Open Data* permettent de discriminer les risques, l'amélioration de la segmentation globale des modèles de fréquence est plus importante sur la garantie TGN. Néanmoins, l'intégration de ces variables contribue inéluctablement à l'amélioration de l'ajustement de la fréquence de sinistres prédite à la fréquence de sinistres observée.

En outre, la création des différents zoniers de fréquence de sinistres facilite l'insertion de l'information géographique dans la modélisation. Ces zoniers ont apporté une plus-value en termes de qualité de segmentation. Cependant, cette bonification est plus marquée dans la modélisation sans données *Open Data*. Cette moindre amélioration est due au fait qu'une partie de l'effet géographique est déjà contenue dans les données *Open Data*. Toutefois, il apparaît indéniablement que ces données *Open Data* permettent d'interpréter l'information géographique captée par les zoniers.

Enfin, les modèles de prime pure obtenus sur les deux garanties ont des capacités de segmentation satisfaisantes. De plus, sauf pour la garantie habitation TGN, le gain de segmentation obtenu avec les données *Open Data* dans les modèles de prime pure serait négligeable. Ainsi, il semble alors possible de se passer du coût d'alimentation des données *Open Data* sans perdre en qualité de segmentation.

Toutefois, une étude annexe pourrait être menée sur la modélisation des facteurs géographiques à partir des données *Open Data*. D'un autre côté, une recherche plus approfondie de données en *Open Data* discriminant significativement le coût moyen des sinistres per-

mettrait une évaluation plus pertinente de l'intérêt de ces données.

De plus, la segmentation la plus juste serait une modélisation distincte du coût moyen exploitation et habitation, qui n'a pu être effectuée à cause d'un faible volume de données. Enfin, bien qu'une tendance soit globalement observée en termes de stabilité temporelle des modèles, celle-ci n'est pas parfaite. Il conviendrait d'actualiser les modèles obtenus quelques années plus tard pour challenger la segmentation des modèles.

En somme, les primes pures incendie et TGN (les deux garanties les plus importantes) modélisées permettent de mesurer le niveau de sinistralité réel des clients. Ces modèles de primes pures constituent un nouvel atout majeur dans l'ajustement tarifaire de la prime globale multirisque agricole.

# Annexe A

# Test de Kolmogorov-Smirnov

Soit $X = (x_1, .., x_n)$ une série de données indépendante et de même loi.
Le test de Kolmogorov-Smirnov est un test d'ajustement de loi, à partir de fonctions de répartition.

Avec $H_0$ : la loi de la distribution de $X$ a la même fonction de répartition $F$ qu'une loi continue donnée.
Si $H_0$ est vraie, la fonction de répartition empirique $\hat{F}$ de $X$ doit être "proche" de $F$.

# Annexe B

# Indice de Moran

L'indice de Moran est un indice d'autocorrélation spatiale.

$$I = \frac{N}{\Sigma_{i,j=1}^{N} w_{ij}} \frac{\Sigma_{i,j=1}^{N} w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\Sigma_{i=1}^{N}(X_i - \overline{X})^2}$$

— $X$ un champ de réel ;
— $N$ le nombre de sites ;
— $W$ la matrice carrée des poids poids de dimension N ;
— $w_{ij}$ élément de W, quantifiant quantifiant les influences de $j$ sur $i$ ;
— $\overline{X}$ la moyenne de $X$.

# Annexe C

# *Grid search* XGBoost

| max_depth | eta | nround | Gini train | Gini test |
|---|---|---|---|---|
| 2 | 0,025 | 100 | 32,93% | 31,01% |
| 3 | 0,025 | 100 | 33,60% | 33,07% |
| 4 | 0,025 | 100 | 34,67% | 34,15% |
| 5 | 0,025 | 100 | 35,53% | 34,19% |
| 2 | 0,025 | 250 | 40,36% | 39,84% |
| 3 | 0,025 | 250 | 42,83% | 41,73% |
| 4 | 0,025 | 250 | 44,28% | 41,87% |
| 5 | 0,025 | 250 | 46,00% | 43,35% |
| 6 | 0,025 | 250 | 48,04% | 42,05% |
| 16 | 0,025 | 250 | 68,22% | 44,36% |
| 2 | 0,05 | 100 | 37,43% | 35,95% |
| 3 | 0,05 | 100 | 39,76% | 38,59% |
| 4 | 0,05 | 100 | 40,42% | 39,17% |
| 5 | 0,05 | 100 | 42,08% | 39,26% |
| 6 | 0,05 | 100 | 43,10% | 39,58% |
| 16 | 0,05 | 100 | 59,87% | 41,30% |
| 2 | 0,05 | 150 | 43,41% | 42,76% |
| 3 | 0,05 | 150 | 45,42% | 44,24% |
| 4 | 0,05 | 150 | 47,45% | 44,81% |
| 5 | 0,05 | 150 | 48,97% | 45,66% |
| 6 | 0,05 | 150 | 51,05% | 46,28% |
| 16 | 0,05 | 150 | 73,13% | 46,58% |
| 2 | 0,05 | 200 | 47,82% | 45,48% |

(a) Incendie exploitation

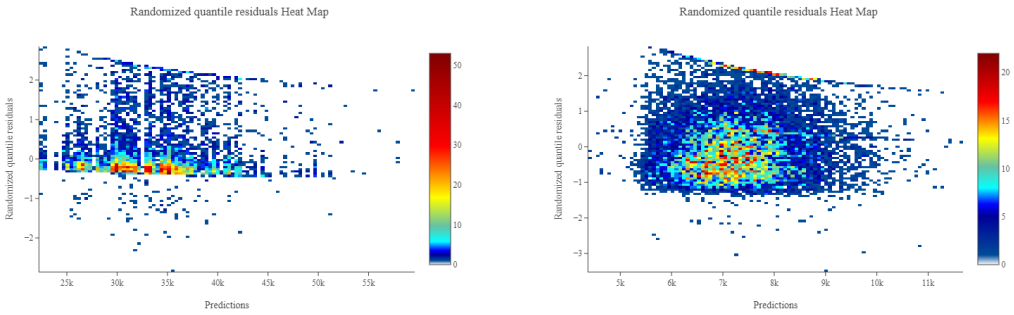
| max_depth | eta | nround | Gini train | Gini test |
|---|---|---|---|---|
| 2 | 0,025 | 150 | 32,43% | 30,75% |
| 3 | 0,025 | 150 | 34,35% | 31,70% |
| 16 | 0,025 | 150 | 66,82% | 34,88% |
| 2 | 0,025 | 200 | 32,77% | 31,40% |
| 6 | 0,025 | 200 | 39,14% | 34,55% |
| 16 | 0,025 | 200 | 69,68% | 35,43% |
| 2 | 0,05 | 100 | 33,22% | 32,26% |
| 3 | 0,05 | 100 | 34,43% | 33,19% |
| 4 | 0,05 | 100 | 35,69% | 33,46% |
| 2 | 0,05 | 120 | 33,30% | 32,31% |
| 3 | 0,05 | 120 | 34,81% | 33,85% |
| 3 | 0,05 | 150 | 35,28% | 35,06% |
| 4 | 0,05 | 150 | 36,64% | 35,51% |
| 5 | 0,05 | 150 | 37,91% | 36,36% |
| 6 | 0,05 | 150 | 40,40% | 36,52% |
| 16 | 0,05 | 150 | 72,71% | 37,76% |
| 2 | 0,05 | 200 | 34,33% | 35,80% |
| 3 | 0,05 | 200 | 35,74% | 35,89% |
| 4 | 0,05 | 200 | 37,00% | 36,57% |
| 5 | 0,05 | 200 | 39,27% | 37,04% |
| 6 | 0,05 | 200 | 41,69% | 37,19% |

(b) TGN exploitation

FIGURE C.1 – Quelques résultats du Grid search XGBoost

# Annexe D

# Résidus quantiles GLM log-normale




(a) Incendie (b) TGN

FIGURE D.1 – Résidus quantiles GLM log-normale coût moyen

# Table des figures


1 Fréquence empirique des sinistres incendie habitation . . . . . v
2 Empirical frequency of home fire claims . . . . . ix

1.1 Source : [AGRESTE, 2022] . . . . . 3
1.2 Évolutions des différents types d'exploitations source : AGRESTE . . . . . 4
1.3 Cycle de production . . . . . 5
1.4 Évolution de l'apport net . . . . . 8

2.1 Étapes de construction de la base des contrats . . . . . 14
2.2 Étapes de construction de la base des sinistres . . . . . 15
2.3 Évolution de facteurs d'actualisation . . . . . 16
2.4 Création de la base de modélisation . . . . . 16
2.5 Poids des primes pures observées par garantie . . . . . 17
2.6 Fréquence de sinistres exploitation : contenu et surface . . . . . 18
2.7 Répartition des images de risque selon le contenu et la surface de l'exploitation . 19
2.8 Fréquence de sinistres exploitation : activité et qualité sur l'exploitation . 19
2.9 Fréquence de sinistres habitation : contenu, nombre de pièces, qualité . . . 20
2.10 Corrélations : $V$ de Cramer . . . . . 21
2.11 Application de la fonction *getbb* sur la Brest . . . . . 22
2.12 Répartition des centres d'incendie et de secours . . . . . 24
2.13 Répartition des stations météorologiques . . . . . 26
2.14 Répartition des crimes par département . . . . . 27
2.15 Prix au mètre carré des logements . . . . . 28
2.16 Proportion de maisons construites entre 2006 et 2015 . . . . . 29

3.1 *QQ-plot* charge incendie exploitation - habitation . . . . . 35
3.2 Détermination du seuil incendie . . . . . 36
3.3 *QQ-plot* charge TGN exploitation - habitation . . . . . 38
3.4 Fonction d'excès moyen TGN . . . . . 38
3.5 Détermination du seuil TGN . . . . . 39
3.6 Facteurs de Chain-Ladder . . . . . 41

4.1 Ajustement de lois - Garantie incendie . . . . . 46
4.2 Ajustement de lois - Garantie TGN . . . . . 47

4.3 Ajustement de lois sur le coût moyen - Garantie TGN . . . . . 47
4.4 Interaction des variables « type de risque » et « contenu de l'exploitation» incendie (exploitation à gauche, habitation à droite) . . . . . 49
4.5 Déviance du GLM gamma en fonction de $\lambda$ . . . . . 50
4.6 Sélection des modèles en fonction du Gini . . . . . 51
4.7 Exemple d'arbre de décision : source [R-bloggers, 2021] . . . . . 52
4.8 Courbe de Lorenz . . . . . 55
4.9 Lift curve . . . . . 56
4.10 Résidus quantiles . . . . . 57
4.11 Structure de la validation croisée . . . . . 58

5.1 Proportion de contrats habitation . . . . . 59
5.2 Fréquence de sinistres habitation . . . . . 61
5.3 Courbes de Lorenz sur base de validation : *k-folds* . . . . . 62
5.4 Résidus quantiles : fréquence incendie exploitation . . . . . 62
5.5 *lift curve* : fréquence incendie exploitation . . . . . 63
5.6 Variables significatives : fréquence incendie exploitation . . . . . 63
5.7 Stabilité temporelle : fréquence incendie . . . . . 64
5.8 Graphiques des résultats : modélisation fréquence TGN habitation . . . . 65
5.9 Graphiques des résultats : modélisation fréquence TGN exploitation . . . 66
5.10 XGBoost fréquence incendie exploitation . . . . . 68
5.11 XGBoost fréquence TGN exploitation . . . . . 69
5.12 Graphiques des résultats : modélisation coût moyen incendie . . . . . 70
5.13 Stabilité temporelle et prédiction : coût moyen incendie . . . . . 70
5.14 Graphiques des résultats : modélisation coût moyen TGN . . . . . 71

6.1 *Open Data* : fréquence incendie exploitation . . . . . 74
6.2 Open Data : fréquence TGN habitation . . . . . 75
6.3 *Open Data* : fréquence TGN exploitation . . . . . 76
6.4 Zonier exploitation incendie . . . . . 81
6.5 Bêta zonier exploitation incendie . . . . . 83
6.6 Zonier exploitation TGN . . . . . 83
6.7 Bêta zonier exploitation TGN . . . . . 84

7.1 *Lift curve* - Prime pure habitation incendie . . . . . 86
7.2 Courbe de Lorenz - Prime pure habitation incendie . . . . . 87
7.3 *Lift curve* - Prime pure incendie exploitation . . . . . 87
7.4 *Lift curve* - Prime pure TGN habitation . . . . . 88
7.5 *Lift curve* - Prime pure habitation TGN habitation avec ajustement . . . . . 89
7.6 *Lift curve* - Prime pure TGN exploitation . . . . . 90

C.1 Quelques résultats du Grid search XGBoost . . . . . 95

D.1 Résidus quantiles GLM log-normale coût moyen . . . . . 96

# Liste des tableaux


1 Résultats de la modélisation fréquence de sinistres TGN habitation . . . . v
2 Results of the SHN housing loss frequency modelling . . . . . x

1.1 Ratio S/C par année : FRANCE ASSUREURS . . . . . 4
1.2 Garanties risque habitation . . . . . 6
1.3 Garanties risque exploitation . . . . . 7

2.1 Crimes sélectionnés . . . . . 26
2.2 Taux de croissance annuels . . . . . 27

3.1 Synthèse des résultats - Fonction d'excès moyen - Garantie incendie . . . . 37
3.2 Triangle de liquidation cumulé . . . . . 40

4.1 Exemples de lois appartenant à la famille exponentielle . . . . . 45

5.1 Résultats de la modélisation fréquence incendie habitation . . . . . 60
5.2 Résultats de la modélisation fréquence incendie exploitation . . . . . 61
5.3 Résultats de la modélisation fréquence TGN habitation . . . . . 65
5.4 Résultats de la modélisation fréquence TGN exploitation . . . . . 66
5.5 Paramètres XGBoost . . . . . 67
5.6 Résultats de la modélisation : XGBoost fréquence incendie exploitation . . 67
5.7 Résultats de la modélisation : XGBoost fréquence TGN exploitation . . . 68
5.8 Résultats de la modélisation coût moyen incendie . . . . . 69
5.9 Résultats de la modélisation coût moyen TGN . . . . . 71

6.1 Résultats de la modélisation fréquence incendie exploitation avec *Open Data* 73
6.2 Résultats de la modélisation fréquence TGN habitation avec *Open Data* . 74
6.3 Résultats de la modélisation fréquence TGN exploitation avec *Open Data* 75
6.4 Résultats GLM gamma coût moyen *Open Data* . . . . . 77
6.5 Résultats à la suite de l'ajout du zonier - exploitation incendie . . . . . 82
6.6 Résultats à la suite de l'ajout du zonier - exploitation TGN . . . . . 84

7.1 Indicateurs de performances - Prime pure incendie habitation . . . . . 87
7.2 Indicateurs de performances - Prime pure incendie exploitation . . . . . 88
7.3 Indicateurs de performances - Prime pure TGN habitation . . . . . 89

7.4 Indicateurs de performances - Prime pure TGN exploitation . . . . . 90

# Bibliographie


[AGRESTE, 2010a] AGRESTE (2010a). L'agriculture française en 2010 premiers résultats du recensement agricole. https://agriculture.gouv.fr/sites/default/files/documents/pdf/DP_recensement_agricole.pdf.

[AGRESTE, 2010b] AGRESTE (2010b). Nombre d'exploitations. https://www.observatoire-des-territoires.gouv.fr/nombre-dexploitations.

[AGRESTE, 2021] AGRESTE (2021). Primeur recensement agricole 2020. https://agreste.agriculture.gouv.fr/agreste-web/download/publication/publie/Pri2105/Primeur%202021-5_Recensement-Agricole-2020.pdf.

[AGRESTE, 2022] AGRESTE (2022). L'agriculture française en 5 chiffres fous. https://www.lesechos.fr/industrie-services/conso-distribution/lagriculture-francaise-en-5-chiffres-fous-1371637.

[AKAFFOU, 2020] AKAFFOU, D. H. (2020). *Méthode alternative de tarification santé : GLM/XGBOOST*. Mémoire d'actuariat.

[Akur8, 2019] Akur8 (2019). *Gaussian Processes and geographic smoothing*. Akur8.

[ALLAIRE, 2020] ALLAIRE, O. (2020). *Comparaison de différentes méthodes pour la modélisation de la prime pure d'un produit risque aggravé en assurance automobile*. Mémoire d'actuariat.

[AXA, 2022] AXA, F. (2022). *Étude Fonctionnelle Tarification du Produit Multirisque Agricole*. AXA France.

[BERNANOSE, 2020] BERNANOSE, A. (2020). *Modélisation du risque Incendie en assurance MultiRisques Habitation*. Mémoire d'actuariat.

[BERRADA, 2021] BERRADA, N. (2021). *Elaboration de zoniers en assurance MRH à partir de l'open data*. Mémoire d'actuariat.

[Boursedescredits, 2022] Boursedescredits (2022). Définition d'un sinistre. https://www.boursedescredits.com/lexique-definition-sinistre-4001.php.

[Cahiers Agricultures, 2021] Cahiers Agricultures (2021). Agriculture et systèmes alimentaires face à la covid-19. https://www.cirad.fr/les-actualites-du-cirad/actualites/2021/cahiers-agricultures-face-a-la-covid-19#:~:text=Au%20d%C3%A9but%20de%20cette%20crise,la%20s%C3%A9curit%C3%A9%20alimentaire%20des%20populations.

[Camille RISI - Sandrine BONY, 2019] Camille RISI - Sandrine BONY (2019). Les nuages, enfants terribles du climat. https://theconversation.com/les-nuages-enfants-terribles-du-climat-113102#:~:text=Les%20nuages%20jouent%20un%20r%C3%B4le,%C3%A0%20l'effet%20de%20serre.

[Code des assurances, 2022] Code des assurances (2022). Code des assurances : Assurance de dommages | juin 2022. https://www.dalloz.fr/documentation/Document?id=DZ%2FOASIS%2F001103.

[COULIBALY, 2021] COULIBALY, A. (2021). *Modélisation des sinistres de la garantie incendie en Multirisque Professionnelle.* Mémoire d'actuariat.

[DENNIEL, 2021] DENNIEL, C. (2021). *Lissage des résidus par Krigeage dans la création d'un zonier : Application sur un portefeuille MRH.* Mémoire d'actuariat.

[France Assureurs, 2020] France Assureurs (2020). L'assurance franÇaise données clés 2020. https://www.franceassureurs.fr/wp-content/uploads/VF-Donnees-cles-2020.pdf.

[Goodassur, 2022] Goodassur (2022). Indice ffb. https://goodassur.com/assurance-habitation/indice-ffb.

[IFOP, 2020] IFOP (2020). L'impact de la crise du covid-19 sur l'agriculture. https://www.lavoixdunord.fr/910009/article/2020-12-18/l-impact-de-la-crise-du-covid-19-sur-l-agriculture.

[INSEE, 2020] INSEE (2020). Les agriculteurs : de moins en moins nombreux et de plus en plus d'hommes. https://www.insee.fr/fr/statistiques/4806717#tableau-figure1.

[Insee, 2020] Insee (2020). Chef d'exploitation agricole et coexploitants. https://www.insee.fr/fr/metadonnees/definition/c1326.

[INSEE, 2021a] INSEE (2021a). Données logements. https://www.insee.fr/fr/statistiques/5395859?sommaire=5395912&q=LOG1+%E2%80%93%20Logements+construits+avant+2016+par+type%2C%20cat%C3%A9gorie+et+%C3%A9poque+d%27ach%C3%A8vement+de+la+construction.

[INSEE, 2021b] INSEE (2021b). Exploitation agricole. https://www.insee.fr/fr/metadonnees/definition/c1186.

[INSEE, 2021c] INSEE (2021c). Logements en france : les chiffres de l'insee au 1er janvier 2021. https://monimmeuble.com/actualite/logements-en-france-les-chiffres-de-linsee-au-1er-janvier-2021.

[LEFEBVRE, 2018] LEFEBVRE, A. (2018). *Modeling of the Policy Net Present Value in Agricultural Multi-Risk insurance.* Mémoire d'actuariat.

[LEGIFRANCE, 2016] LEGIFRANCE (2016). Loi n° 2016-1321 du 7 octobre 2016 pour une république numérique (1). https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000033205212/2022-08-03.

[NICOLLE, 2017] NICOLLE, C. (2017). *Tarification au trajet à l'aide de l'Open Data.* Mémoire d'actuariat.

[NOAA, 2022] NOAA (2022). Données météorologiques. https://www1.ncdc.noaa.gov/pub/data/noaa/.

[ONDRP, 2022] ONDRP (2022). Chiffres départementaux mensuels relatifs aux crimes et délits enregistrés par les services de police et de gendarmerie depuis janvier 1996. https://www.data.gouv.fr/fr/datasets/chiffres-departementaux-mensuels-relatifs-aux-crimes-et-delits-enregistres-par-les-serv

[PARIENTE, 2017] PARIENTE, J. (2017). *Modélisation du risque géographique en assurance habitation.* Mémoire d'actuariat.

[R-bloggers, 2021] R-BLOGGERS (2021). Exemple arbre de décision. https://www.r-bloggers.com/2021/04/decision-trees-in-r/.

[Richou, 2018] RICHOU, Y. T. A. (2018). *Introduction aux Statistiques Bayésiennes.* Université de Bordeaux.

[SEPULVEDA, 2016] SEPULVEDA, C. (2016). *Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial.* Mémoire d'actuariat.

[THUILLIER, 2021] THUILLIER, M. (2021). *Calcul de la valeur contrat sur la branche Multirisque Immeuble comme aide opérationnelle à la relation client.* Mémoire d'actuariat.

[TOESCA, 2019] TOESCA, R. (2019). *Tarification de la garantie incendie en Dommages Aux Biens - Entreprises.* Mémoire d'actuariat.

[VERMET, 2021] VERMET, F. (2021). *Arbres de décision et méthodes ensemblistes.* Note de cours.

[Wikipedia, 2022] WIKIPEDIA (2022). Indice de moran. https://fr.wikipedia.org/wiki/Indice_de_Moran.

[Yves FOUQUART, 2015] YVES FOUQUART (2015). Les nuages et leur rétroaction. https://www.futura-sciences.com/planete/dossiers/climatologie-tant-incertitudes-previsions-climatiques-638/page/4/.