Institut des ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : ~~Madame~~ / Monsieur CHELBON Alexandre

Titre du mémoire : Refonte de la prime commerciale du périmètre DOM en assurance MRH

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

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

Membres présents du jury de la filière :

Signature :

Entreprise :

Nom : AXA France IARD

Signature :

Directeur de mémoire en entreprise

Nom : DECAUX Brice

Signature :

Membres présents du jury de l'Institut des Actuaires :

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

Résumé

Dans le contexte d'un marché de l'assurance habitation fortement concurrentiel, les actuaires de l'équipe *Actuariat Pricing MRH* d'AXA France doivent sans cesse établir une multitude de modèles de prime pures pour différents segments et garanties et les maintenir dans le temps afin d'accoler à la sinistralité réelle et proposer ainsi des produits rentables. Se pose alors la question de la simplification du processus tarifaire dans le but de réduire le nombre de modèle à réaliser.

L'objet de ce mémoire est ainsi d'étudier la pertinence de l'intégration des contrats situés dans les DOM à la modélisation de la prime pure actuelle pour les contrats de France métropolitaine. On sélectionnera donc le modèle optimal de prime pure pour la garantie dégâts des eaux dans les DOM. Le modèle devra être optimal d'un point de vue opérationnel (faut-il intégrer la tarification *DOM* dans la tarification *Métropole* ?) mais également d'un point de vue technique (la prime pure prédite correspond-elle bien aux risques encourus ?)

On commencera par analyser les contrats DOM d'AXA France au niveau commercial puis au niveau sinistralité afin d'avoir une vue d'ensemble du produit. On détaillera ensuite les étapes de la construction des bases de modélisations pertinentes pour répondre à notre problématique au mieux.

Le coeur du mémoire sera la modélisation de la prime pure dégâts des eaux grâce à un modèle coût-fréquence. On utilisera les modèles linéaires généralisés avec les distributions classiques dans un premier temps. Dans un second temps on utilisera une modélisation novatrice : le *Generalised Additive Models for Location Scale and Shape*, dit GAMLSS qui va nous permettre de modéliser de nouvelles distributions mieux adaptées à nos données. On complètera l'étude par une analyse du risque géographique avec une méthode de zoniers innovante intitulée *GeoRev*.

Enfin, on étendra les conclusions à toutes les garanties et on réalisera une étude commerciale des contrats situés dans les DOM en comparant ancien et nouveau tarif. Pour cela on utilisera les modèles selectionnés précedemment sur une nouvelle base de données et on pourra ainsi apprécier les impacts de notre nouvelle tarification.

Mots-clés : *Assurance MultiRisques Habitation, Dégât des eaux, Prime pure, Sélection de modèle, Simplification tarifaire, Modèles linéaires généralisés, Pénalisation, Generalised Additive Models for Location Scale and Shape, Adéquation de lois de probabilités, Zoniers, Lissage Spatial, Prime commerciale*

Abstract

In the context of a highly competitive home insurance market, the actuaries of AXA France's *Home Insurance Pricing* team must constantly establish a multitude of pure premium models for different segments and guarantees and maintain them over time in order to attach to the actual loss and thus offer profitable products. This raises the question of simplifying the pricing process in order to reduce the number of models to be carried out.

The purpose of this thesis is thus to study the relevance of the integration of contracts located in the French overseas departments (DOM) to the modelling of the current pure premium for metropolitan France contracts. We will therefore select the optimal model of pure premium for the water damage guarantee in the DOM. The model will have to be optimal from a business point of view (should DOM pricing be integrated into Metropolitan pricing ?) but also from a technical point of view (does the predicted pure premium correspond well to the risks involved ?)

We will start by analyzing AXA France's DOM contracts at the commercial level and then at the claims level in order to have an overview of the product. We will then detail the stages of the construction of the relevant modeling bases to best meet our problem.

The heart of the thesis will be the modeling of the pure water damage premium thanks to a cost-frequency. Conventional generalized linear models will be used with classical distributions in a first step. In a second step, an innovative method will be used : the *Generalised Additive Models for Location Scale and Shape*, known as GAMLSS, which will allow us to model new distributions better adapted to our data. The study will be supplemented by a geographical risk analysis with a new zoniers method entitled GeoRev.

Finally, the conclusions will be extended to all guarantees and a commercial study of contracts located in the French overseas departments will be carried out by comparing old and new pricing. For this we will use the models selected previously on a new database and we will be able to assess the impacts of our new pricing.

Keywords : *Home Insurance, Water Damage, Pure Premium, Model Selection, Price Simplification, Generalized Linear Models, Penalization,*
Generalised Additive Models for Location Scale and Shape, Adequacy of probability distributions, Zoning, Spatial Smoothing, Profitability

Notes de synthèse

Contexte de l'étude

Cette étude se place dans le cadre d'une simplification des processus tarifaires au sein de l'équipe MRH d'AXA France. L'équipe MRH d'AXA France doit en effet développer et maintenir une multitude de modèles dans le but d'expliquer au mieux la sinistralité de son portefeuille. Et, théoriquement, en prenant en compte une segmentation appartement / maison et en réalisant des modèles pour chaque segment(au nombre de 4 dont font partie les DOM) et garanties (au nombre de 10), c'est 80 modèles qui seraient à réaliser et à maintenir dans le temps.

Ainsi, ce processus de simplification tarifaire passe par l'étude de la pertinence de l'intégration de la sinistralité des DOM à la sinistralité de France métropolitaine afin de valider ou non la fusion des ces deux pérmimètres qui peut permettre de réduire le nombre de modèle à maintenir / réaliser . L'étude a été faite sur toutes les garanties mais seules les analyses sur la garantie dégâts des eaux ont été explicitées.

De plus, cette étude peut également permettre d'introduire le nouveau produit MRH d'AXA France : "Ma Maison" dans les territoires ultramarins où le marché de l'assurance IARD et particulièrement celui de la MRH est réputé complexe et concurrentiel pour les assureurs (difficulté à appliquer l'anti-sélection des risques, innassurabilité de certaines zones géographiques, phénomène de non-assurance). Ainsi, AXA représente environ 9% de ce marché et est notamment en concurrence avec les bancassureurs qui y sont en forte croissance : l'introduction du nouveau produit pourrait permettre de gagner des parts de marché.

Préparation des données

La préparation des données a été une étape clé de notre étude. La première étape de la construction des bases de données a été de déterminer un seuil de sinistre grave. En effet, dans notre étude, nous avons fait la distinction entre sinistralité grave (charge élevée et faible fréquence d'occurence) et sinistralité attritionnelle (charge plus classique et fréquence d'occurence plus importante). Ainsi la charge des sinistres a été écrêtée grâce à ce seuil de sinistre grave que nous avons déterminé à l'aide d'un "Mean Excess Plot"

d'une part mais aussi en analysant la sinistralité réelle d'autre part. Cette analyse a montré des seuils de graves différents pour le périmètre "DOM" et le périmètre "métropole", or, pour comparer les sinistralités attritionnelles, ce seuil doit être identique. Ainsi, le seuil de grave de la métropole (bien plus élevé) a été retenu et de ce fait, il n'y a quasiment pas de sinistres graves dans les DOM au vu des données.

Après avoir vieilli nos charges de sinistres en utilisant la méthode de Chain Ladder et en distinguant les sinistres graves et les sinistres attritonnelles, nous avons pu effectivement construire les bases de données de modélisation en utilisant d'une part les "bases images" contenant les informations liées au contrat des assurés (type de logement, nombre de pièces...), les "bases sinistres" contenant les informations liées aux sinistres subis par les assurés (garanties concernée, montant...) et enfin les "bases clients" contenant les informations propres à l'assuré (âge, ancienneté du contrat...). Chacune de ces bases a été extraite sur une fenêtre d'observation allant de 2009 à 2019 (afin d'avoir un maximum de données pour les DOM). Un retraitement des codes INSEE des contrats situés dans les DOM grâce à une base de données externe a également été effectué afin d'être cohérent avec le découpage administratif actuellement en vigueur.

Grâce à ces bases, ont été construites les bases fréquences (avec une ligne par numéro de contrat x année de survenance) et les bases sinistres (avec une ligne par numéro de sinistre x année de survenance) sur deux périmètres : l'un avec uniquement des contrats "DOM" et l'autre avec des contrats "DOM" et des contrats "métropole". Les variables de ces bases de modélisation ont été retraitées avec notamment une gestion des valeurs manquantes / aberrantes, une discrétitation et un regroupement de modalités afin de n'avoir que des variables catégorielles. Une analyse de corrélation a également été effectuée afin de déterminer s'il fallait écarter des variables trop corrélées. Enfin, une analyse descriptive a été réalisée en traçant notamment des graphiques des variations de nos variables à expliquer en fonction de certaines variables explicatives dans le but d'avoir une première idée des variables significatives pour notre étude.

Modélisation de la sinistralité DDE hors zonier

Comme mentionné précedemment, ce mémoire se concentre en particulier sur la modélisation de la garantie dégât des eaux.

La première étape de cette modélisation a été de construire un modèle "hors zonier" où aucune variable géographique n'a été prise en compte. En effet, les bases de données (fréquence et coût moyen) ont été divisée en trois : une "base HZ" (pour hors zonier), une "base Z" (pour zonier) et une "base V" (pour validation) pour chaque périmètre ("DOM" et "métropole + DOM"). Le but a été ainsi de construire d'une part un modèle hors zonier basé sur les DOM (base D HZ) et un autre basé sur la métropole et les DOM (base M+D HZ), et de comparer les performances de ces deux modèles sur une base DOM (base D Z) selon le schéma suivant :

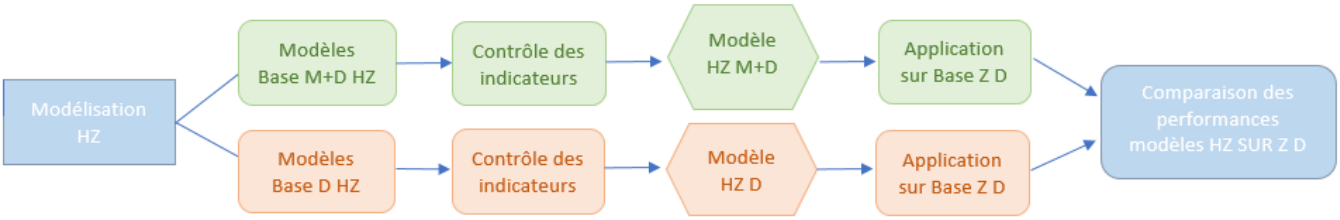


Schéma de la modélisation hors zoniers

Une approche "coût-fréquence" a été utilisée afin d'aboutir à cette modélisation et les modèles ont été évalués à l'aide d'une validation croisée par k-folds et de métriques de performances telles que le coefficient de Gini. Des modèles linéaires généralisés (GLM) avec les lois classiques (Poisson pour la fréquence et Gamma pour le coût moyen) ont ainsi été utilisés, mais très vite, nous nous sommes aperçus à l'aide QQ-plots que nos variables à expliquer ne suivaient globalement pas ces lois usuelles.

Ainsi, nous avons utilisé une extension du GLM : le GAMLSS (Generalised Additive Models for Location Scale and Shape) qui proposent de nombreuses lois possédant notamment des paramètres de formes (en plus des paramètres de position et d'échelle) ce qui étend les possibilités d'ajustement. Ainsi, grâce à ce nouveau paradigme de modélisation, nous avons déterminé que la loi de Sichel (introduite par les GAMLSS) était plus adaptée à la distribution du nombre de sinistre que les lois de "Poisson" et "Binomiale négative". De même pour la distribution du montant de charge, les lois "Box-Cox" et "Béta Généralisée II" sont plus adaptées que les lois "Gamma" ou encore "Inverse Gaussienne". De ce fait, tout au long de la construction de nos modèles (sélection des variables, sélection de fonction de lissage introduite par le GAMLSS), nous avons comparé les "lois classiques" aux "lois GAMLSS" afin d'établir si ces dernières donneraient effectivement de meilleures performances.

Pour la modélisation de la fréquence, la loi de Sichel a offert de meilleures performances et c'est elle qui a été retenue à la fois pour le modèle "DOM" et le modèle "métropole + DOM", la comparaison des performances entre ces deux modèles sur la base "Z DOM" sont dans le tableau suivant :

Indicateurs	Modèle fréquence DOM	Modèle fréquence Mét+DOM
Nb variables	12	12
Gini	36,33%	25,18%
RMSE	0,1474	0,1475
MAE	0,0230	0,0229
Avg Deviance	0,1054	0,1078
Erreur globale	2,110%	3,675%

Comparaison des performances des modèles fréquence hors zonier sur la base "Z DOM"

Nous avons ainsi observé que le modèle basé uniquement sur les DOM segmente bien mieux le risque que celui basé sur la métropole et les DOM.

Pour le coût moyen, nous avons noté que, bien que la loi de Box-Cox (sur le périmètre "DOM") et la loi Béta généralisée II (sur le périmètre "métropole+DOM") ajustaient mieux le montant de charges, c'était la loi Gamma qui offrait de meilleures performances sur les deux périmètres, et voici le tableau permettant de comparer les performances du modèle "DOM" et du modèle "métropole + DOM" :

Indicateurs	Modèle coût moyen DOM	Modèle coût moyen métropole + DOM
Nb variables	8	11
Gini	18,06%	15,96%
RMSE	1211,69	1219,88
MAE	661,29	667,67
Avg Deviance	0,56270	0,58540
Erreur globale	2,24%	1,72%

Comparaison des modèles coût moyen hors zonier sur la base "Z DOM"

Là aussi, le modèle basé sur les DOM offre une meilleure segmentation du risque.

Modélisation de la sinistralité finale

Après avoir déterminer quels étaient les meilleurs modèles hors zonier, nous avons introduit une variable géographique (le code INSEE) afin de compléter nos modèles (nous avons ensuite ajouter la modélisation des antécédents). Pour cela, nous avons utilisé l'algorithme GeoRev (interne à AXA) qui se base sur une approche par crédibilité mais aussi par similarité du risque, cette méthode nous a permis de classer les risques selon leur localisation.

Nous avons utiliser la même approche que précédemment : en partant du modèle hors zonier "DOM", nous avons construit un zonier sur les DOM nous avons ensuite intégré cette nouvelle variable à notre modéle afin d'analyser ses performances sur notre base de validation (base Z DOM). La même opération a été effectuée sur le périmètre "métropole + DOM" avec la construction d'un zonier sur les DOM mais aussi sur la métropole.

Cependant, nous avons eu l'idée de créer un modèle "hybride " , partant du modèle hors zonier basé sur le périmètre "métropole+DOM" mais en faisant notre zonier uniquement sur les DOM (et non plus sur les DOM et la métropole). La méthodologie de modélisation est résumée dans le schéma suivant :

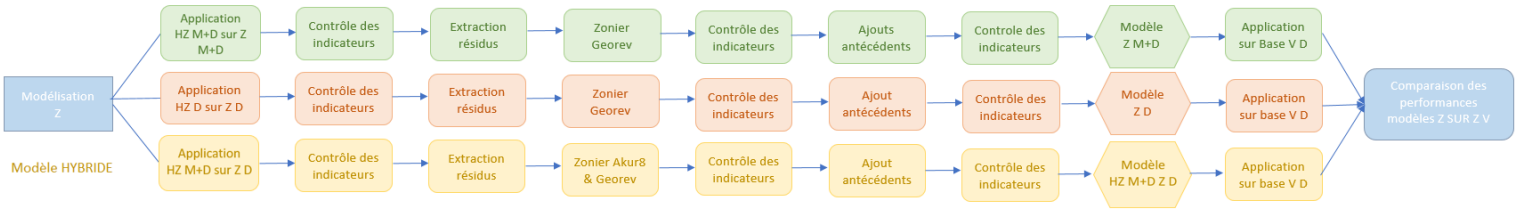


Schéma de la modélisation avec zonier avec le modèle hybride

Ainsi, pour la modélisation de la garantie dégâts des eaux, nos conclusions sur la sinistralité attritionnelle sont les suivantes :

- Pour la modélisation du coût moyen, le modèle hybride est meilleur que le modèle basé uniquement sur les DOM, et concernant les distributions à utiliser, c'est la loi "Gamma" qu'il faut choisir.

- Pour la modélisation de la fréquence, notre modélisation hybride permet de nous rapprocher des performances du modèle basé uniquement sur les DOM, et celle-ci est retenue afin d'avoir un équilibre entre les aspects techniques et opérationnels (maintenance des modèles). Concernant les distributions à retenir, la loi de Sichel offre les meilleures performances, et à défaut, il faut choisir la loi binomiale négative.

Enfin, nous avons terminé l'étude de la garantie dégâts des eaux en modélisant la sinistralité grave à l'aide d'un arbre de régression permettant de mutualiser les sinistres graves avec une légère forme de segmentation. De plus, un coefficient a été appliqué aux contrats ultramarins afin d'être cohérent avec la sinistralité grave réelle.

Analyse de la prime commerciale DOM

L'étude de l'intégration de la modélisation des contrats "DOM" aux contrats "métropole" a également été réalisée sur les principales garanties couverte par AXA en utilisant la même méthodologie que pour le dégat des eaux. Cette intégration a ainsi été validée pour chacune des garanties modélisées notamment grâce aux modèles hybrides.

Grâces à ces différents modèles, nous avons construit étape par étape (prime pure avec "base level adjustment", frais et réassurance) la nouvelle prime commerciale du périmètre "DOM" afin de l'analyser sur les affaires nouvelles de 2020 et 2021. La nouvelle prime commerciale se voit augmenter de 30% en moyenne, c'est bien plus élevée qu'auparavant et un changement de prime aussi important peut engendrer une hausse des résiliations. Nous appliquons un coefficient afin de s'aligner à l'ancien tarif, de cette façon, l'écart relatif moyen de la prime commerciale est contenu comme le montre le tableau suivant :

DOM	Ecart moyen par contrat avant correction	Ecart moyen par contrat après correction
Guadeloupe	24,70%	4,88%
Guyane	42,00%	1,14%
Martinique	21,48%	3,60%
Réunion	39,63%	2,55%
Tout DOM	34,10%	2,99%

Ecarts relatifs moyen de la prime commerciale par contrat

Nous avons ensuite étudié en détails quels étaient les types d'assurés les plus concernés par ce changement de tarif grâce à un arbre de régression.

Enfin, la même étude a été réalisée sur la rentabilité technique grâce à l'indicateur ELR (Expected Loss Ratio) comme le montre le tableau suivant :

DOM	ELR
Guadeloupe	78,10%
Guyane	66,70%
Martinique	75,66%
Réunion	77,66%
Tout DOM	76,41%

ELR Globaux par DOM

Nous voyons ainsi que l'ELR est de 76,41% sur les DOM ce qui est une valeur classique pour les affaires nouvelles.

Executive summary

Context of the study

This study is part of a simplification of the tariff processes within the team of AXA France. AXA France's HRM team has to develop and maintain a multitude of models in order to models in order to best explain the claims experience of its portfolio. Theoretically, by taking into account an apartment/home segmentation and by creating models for each segment (four of which include the French overseas departments) and coverages (10), 80 models would have to be created and maintained over time.

Thus, this process of tariff simplification involves a study of the relevance of integrating the claims experience of the DOM (French overseas) into the claims experience of metropolitan France in order to validate or not the merger of these two segments, which can reduce the number of models to be maintained / produced. The study was carried out on all guarantees but only the analyses on the water damage guarantee were explained.

In addition, this study can also be used to introduce AXA France's new property and casualty product, "Ma Maison", in the French overseas territories where the property and casualty insurance market, and particularly the property and casualty market, is considered to be complex and competitive for insurers (difficulty in applying anti-selection of risks, uninsurability of certain geographical areas, phenomenon of non-insurance). AXA accounts for around 9% of this market and is in competition with the fast-growing bancassurers : the introduction of the new product could enable it to gain market share.

Data preparation

Data preparation was a key step in our study. The first step in the construction of the databases was to determine a threshold for severe claims. In our study, we distinguished between severe claims (high burden and low frequency of occurrence) and attritional claims (more traditional burden and higher frequency of occurrence), and the burden of claims was capped using this severe claims threshold, which we determined using a "Mean Excess Plot" on the one hand, but also by analyzing the actual claims experience. This

analysis showed different severe loss thresholds for the "DOM" perimeter and the "Metropolitan France" perimeter, but in order to compare attritional losses, this threshold must be identical. Thus, the threshold of seriousness of the "metropolis" (much higher) was retained and as a result, there are almost no serious claims in the DOM in view of the data.

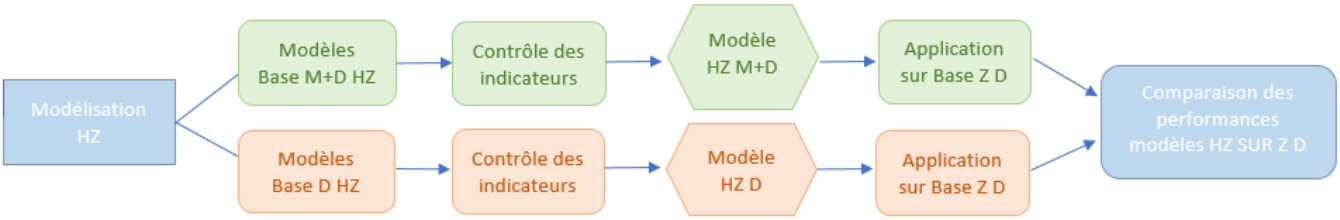
After having aged our claims loads using the Chain Ladder method and distinguishing between serious and attritonnal claims, we were able to effectively build the modeling databases using on the one hand the "image databases" containing information related to the policyholders' contract (type of housing, number of rooms. ...), the "claims databases" containing the information related to the claims suffered by the insured (guarantees concerned, amount...) and finally the "customer databases" containing the information specific to the insured (age, contract seniority...). Each of these databases has been extracted over an observation window from 2009 to 2019 (in order to have a maximum of data for the DOM). The INSEE codes of contracts located in the DOM were also reprocessed using an external database in order to be consistent with the administrative division currently in force.

Thanks to these databases, frequency databases (with one line per contract number x year of occurrence) and claims databases (with one line per claim number x year of occurrence) were built on two perimeters : one with only "DOM" contracts and the other with "DOM" and "metropolitan" contracts. The variables of these modeling bases have been reprocessed with, in particular, a management of missing/outlying values, a discretization and a grouping of modalities in order to have only "categorical" variables. A correlation analysis was also carried out to determine whether variables that were too correlated should be discarded. Finally, a descriptive analysis was carried out by drawing graphs of the variations of our variables to be explained according to certain explanatory variables in order to have a first idea of the significant variables for our study.

Modeling of water damage claims without geographic data

As mentioned above, this paper focuses in particular on the modelling of water damage cover.

The first step of this modelling was to build a "non-zonal" model where no geographical variables were taken into account. In fact, the databases (frequency and average cost) were divided into three : an "HZ base" (for non-zonal), a "Z base" (for zonal) and a "V base" (for validation) for each perimeter ("DOM" and "metropolitan France + DOM"). The aim was thus to build a non-zonal model based on the DOM (D HZ base) and another based on metropolitan France and the DOM (M+D HZ base), and to compare the performances of these two models on a DOM base (D Z base) according to the following scheme :



Non-zonal modelling scheme

A "cost-frequency" approach was used to arrive at this modelling and the models were evaluated using k-fold cross vaidation and performance metrics such as the Gini coefficient. Generalized linear models (GLM) with classical distributions (Poisson for the frequency and Gamma for the average cost) were used, but very quickly, we realized with the help of QQ-plots that our variables to be explained did not globally follow these usual distributions.

Thus, we used an extension of the GLM : the GAMLSS (Generalised Additive Models for Location Scale and Shape) which proposes many distributions with shape parameters (in addition to the position and scale parameters) which extends the possibilities of adjustment. Thus, thanks to this new modelling paradigm, we have determined that the Sichel distribution (introduced by GAMLSS) was more suitable for the distribution of the number of claims than the "Poisson" and "negative Binomial" distributions. Similarly, for the distribution of the amount of charge, the "Box-Cox" and "Generalized Beta II" distributions are better suited than the "Gamma" or "Inverse Gaussian" distributions. Therefore, throughout the construction of our models (selection of variables, selection of smoothing function introduced by GAMLSS), we compared the "classical distributions" to the "GAMLSS distributions" in order to establish whether the latter would indeed give better performances.

For frequency modelling, Sichel's distribution offered better performances and it is the one that was retained for both the "DOM" model and the "metropolis + DOM" model, the comparison of performances between these two models on the "Z DOM" basis are in the following table :

Indicators	DOM frequency model	Met+DOM frequency model
Nb of features	12	12
Gini	36,33%	25,18%
RMSE	0,1474	0,1475
MAE	0,0230	0,0229
Avg Deviance	0,1054	0,1078
Global error	2,110%	3,675%

Comparison of the frequency non-zonal models on the basis of "Z DOM"

We thus observed that the model based solely on the DOM segments the risk much better than the one based on metropolitan France and the DOM.

For the average cost modelling, we noted that, although the Box-Cox distribution (on the "DOM" perimeter) and the generalized Beta II distribution (on the "metropolitan France + DOM" perimeter) adjusted the amount of charges better, it was the Gamma distribution that offered better performances on the two perimeters :

Indicators	DOM average cost model	metropole + DOM average cost model
Nb of features	8	11
Gini	18,06%	15,96%
RMSE	1211,69	1219,88
MAE	661,29	667,67
Avg Deviance	0,56270	0,58540
Global error	2,24%	1,72%

Comparison of the average cost non-zonal models on the basis of "Z DOM"

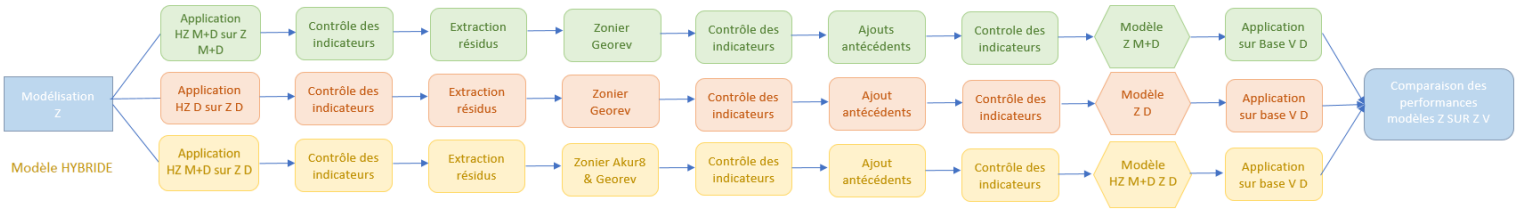
Here again, the DOM-based model offers a better segmentation of risk.

Final loss modelling

After having determined the best models outside the zone, we introduced a geographical variable (the INSEE code) in order to complete our models (we then added the modelling of the antecedents). To do this, we used the GeoRev algorithm (internal to AXA) which is based on a credibility approach but also on risk similarity, this method allowed us to classify the risks according to their location.

We used the same approach as before : starting from the "DOM" non-zonal model , we built a zoning system on the DOM and then integrated this new variable into our model in order to analyse its performance on our validation base (DOM Z base). The same operation was carried out on the "metropolitan France + DOM" perimeter with the construction of a zoning system on the DOM but also on metropolitan France.

However, we had the idea of creating a 'hybrid' model, starting from the non-zonal model based on the 'mainland + DOM' perimeter, but making our zonal model solely on the DOM (and no longer on the DOM and mainland). The modelling methodology is summarised in the following scheme :



Scheme of the zoning model with the hybrid model

Thus, for the modelling of water damage cover, our conclusions on attritional claims are as follows :

- For the modelling of the average cost, the hybrid model is better than the model based only on the DOM, and concerning the distributions to be used, it is the "Gamma" distribution that should be chosen.

- For the frequency modelling, our hybrid model allows us to get closer to the performance of the DOM-only model, and it is retained in order to have a balance between technical and operational aspects (model maintenance). Concerning the distributions to be retained, the Sichel distribution offers the best performances, and failing that, the negative binomial distribution should be chosen.

Finally, we completed the study of water damage coverage by modelling the severe claims experience using a regression tree to pool severe claims with a slight form of segmentation. In addition, a coefficient was applied to the overseas contracts in order to be consistent with the actual serious claims experience.

Analysis of the DOM commercial premium

The study of the integration of the modelling of "DOM" contracts into "metropolitan" contracts was also carried out on the main coverages covered by AXA using the same methodology as for water damage. This integration was validated for each of the modelled coverages, notably thanks to the hybrid models.

Thanks to these different models, we have constructed step by step (pure premium with "base level adjustment", costs and reinsurance) the new commercial premium of the "DOM" perimeter in order to analyse it on the new business of 2020 and 2021. The new commercial premium is increased by 30% on average, which is much higher than before and such a significant change in premium can lead to an increase in cancellations. We apply a coefficient to align with the old tariff, so that the average relative difference in the commercial premium is contained as shown in the following table :

DOM	Avg deviation per contract before correction	Avg deviation per contract after correction
Guadeloupe	24,70%	4,88%
Guyane	42,00%	1,14%
Martinique	21,48%	3,60%
Réunion	39,63%	2,55%
All DOM	34,10%	2,99%

Average relative deviation of the commercial premium by contract

We then studied in detail which types of policyholders were most affected by this tariff change using a regression tree.

Finally, the same study was carried out on the technical profitability using the ELR (Expected Loss Ratio) indicator as shown in the following table :

DOM	ELR
Guadeloupe	78,10%
Guyane	66,70%
Martinique	75,66%
Réunion	77,66%
All DOM	76,41%

Global ELR per DOM

We can see that the ELR is 76.41% in the DOM, which is a classic value for new business.

Remerciements

Je tiens tout d'abord à remercier mon tuteur , Brice DECAUX. Je le remercie particulièrement pour le temps qu'il a passé à m'encadrer ainsi que pour pour son écoute et sa grande pédagogie.

Je souhaite également remercier mon ancien tuteur, Romain TOESCA, qui m'a grandement aidé dans la structuration de mon mémoire à ses débuts. Je le remercie en particulier de m'avoir offert l'opportunité de travailler au sein de son équipe.

Je tiens aussi à remercier mon ancien manager, François LUU pour sa disponibilité et sa bienveillance, mais surtout pour tous les conseils qu'il a pu me donner lors de mes différents travaux.

Je remercie enfin les autres membres de l'équipe Actuariat Pricing pour toute l'aide et les conseils qu'ils ont pu m'apporter.

Enfin, je remercie tous les membres du service Multirisque habitation pour l'accueil qu'ils m'ont réservé, qui a contribué à mon intégration et au bon déroulement de mon alternance tant sur le plan professionel qu'humain.

Table des matières

Introduction

L'assurance habitation est une assurance primordiale : elles protègent en effet les assurés dans leurs vies quotidiennes et couvre ainsi de nombreuses garanties telles que le bris de glaces, le dégât des eaux, l'incendie, le vol mais également les catastrophes naturelles et évènements climatiques. Il s'agit d'un marché en croissance en France. Ainsi, en 2020, le marché a progressé d'1,5 % pour atteindre les 43 millions de contrats Multirisque Habitation. Dans le même temps, le nombre de logements en 2020 est estimé à 37 millions, en hausse de 1,1 %. Les cotisations s'établissent à plus de 11 milliards d'euros soit une hausse de 3,5 %. On a donc toujours des perspectives de croissance pour ce marché.

Cependant, malgré sa croissance, le marché de l'assurance habitation est profondément concurrentiel ce qui peut affecter son essor. Cette situation s'explique notamment par l'entrée en vigueur de la loi Hamon en 2015 qui permet aux assurés de résilier leurs contrats au bout d'un an, mais également par une concurrence accrue avec les bancassureurs. Cela pousse ainsi les assurés à aller chez la concurrence afin de diminuer leurs primes.

Dans cet environnement très concurrentiel qui impacte le dynamisme du produit d'assurance habitation, les assureurs doivent sans cesse adapter leurs tarifs aux risques réellement encourus dans le but de rester rentable, masi surtout de proposer le tarif le plus juste aux clients pour conserver une bonne dynamique commerciale. En effet, il ne serait pas déontologique d'augmenter les primes sans aucune forme de segmentation dans l'unique but d'améliorer sa rentabilité. Ainsi, chez AXA France IARD, l'équipe Actuariat Pricing MRH développe une multitude de modèles de prime pure pour chaque garantie et pour différents segments. L'équipe doit ainsi maintenir ses modèles et les revoir annuellement afin de correspondre à la sinistralité à laquelle AXA est exposée.

Maintenir ces modèles dans le temps demande beaucoup de ressources, tant humaines que technologiques. S'est ainsi posé la question de la simplification du processus de tarification afin de réduire le nombre de modèle à réaliser. Mon mémoire traite ainsi de la pertinence de l'intégration des contrats situés dans les DOM dans la modélisation actuelle de la prime pure des contrats de France métropolitaine. Cela permettrait en effet de n'avoir qu'un unique modèle à réaliser. D'autant plus que cela accélèrerait l'introduction des contrats DOM dans le nouveau produit habitation d'AXA. Ainsi, cette étude va permettre de réaliser une refonte de la prime pure dans les DOM, le modèle actuel étant ancien.

Dans cette étude, on cherche donc à savoir si les contrats situés dans les DOM ont besoin d'une modélisation spécifique ou bien si les modèles actuels basés sur la France métropolitaine suffisent. On va

restreindre l'étude à la **garantie dégâts des eaux** dont les sinistres sont les plus représentés dans les DOM.

Dans une première partie, on va réaliser une revue de la situation des contrats DOM d'AXA France et analyser leur sinistralité. Dans une seconde partie, on détaillera la contruction des bases de données servant à la modélisation. On obtiendra une base de données contenant uniquement des contrats DOM et une base de données contenant en plus les contrats situés en France métropolitaine.

Viendra ensuite la troisième partie, le coeur du mémoire, à savoir la sélection d'un modèle optimal pour la garantie dégâts des eaux dans les DOM. On procèdera à la modélisation de la prime pure par une approche classique de **modélisation coût-fréquence** . Pour cela, on évaluera les performances d'une succession de modèles construits d'une part sur la base de données *DOM* et d'autre part sur la base de données *Métropole+DOM*.

Les modèles utilisés seront les modèles linéaires généralisés avec les distributions classiques et on ira plus loin en utilisant un type de modèle innovant basé sur les modèles linéaires généralisés : le *Generalised Additive Models for Location Scale and Shape*, dit GAMLSS, qui permet d'étendre le GLM à des distributions non usuelles et plus proches des données. On s'attardera ensuite sur la problématique du zonier en utilisant une méthode innonvante intitulée *GeoRev*. Cela permettra entre autres de voir si l'utilisation d'un zonier peut compenser le fait de ne pas réaliser de modèles spécifiques aux DOM.

Enfin, on étendra les conclusions à toutes les garanties et on réalisera une étude commerciale des contrats situés dans les DOM en comparant ancien et nouveau tarif. Pour cela on utilisera les modèles selectionnés précedemment sur une nouvelle base de données et on pourra ainsi apprécier les impacts de notre nouvelle tarification.

Chapitre 1

Contexte de l'étude

1.1. Le marché de l'assurance Multirisque Habitation

Note : les chiffres et graphiques présentés ci-dessous en section 1.1.1 et 1.1.2 proviennent du "Rapport d'activité FFA 2020".

1.1.1 Le marché de l'assurance IARD en France

L'assurance IARD pour « Incendie, Accidents et Risques Divers » est une famille d'assurances visant à couvrir les dommages et la protection des biens par opposition aux assurances VIE qui protègent les personnes. Il s'agit d'une assurance aussi bien destinée aux particuliers (pour assurer leurs véhicule ou logements) qu'aux professionels (pour assurer leurs locaux et autres machines).

En 2020, on comptait 174 compagnies d'assurance non-vie et l'assurance de biens et de responsabilités totalisait 60,1 milliards d'€ de cotisations (en augmentation de 2,4 % par rapport à 2019) sur un total de 201,9 milliards d'€ de cotisations sur l'ensemble de l'assurance française. Sur ces 60,1 milliards d'€ de cotisations, 38,1 milliards d'€ concernent les particuliers et 22,1 milliards d '€ les professionnels.

Concernant la charge de prestations, elle s'établit à 43,1 milliards d'€ (également en augmentation de 2,4 % par rapport à 2019) sur 182,0 milliards d'€ pour l'ensemble de l'assurance française.

Avec les cotisations et les charges, et en ajoutant les frais, on peut estimer le ratio combiné de l'ensemble du marché de l'assurance IARD en France :

L'assurance IARD comporte plusieurs types d'assurances, les plus notables étant l'assurance automobile et l'assurance habitation.

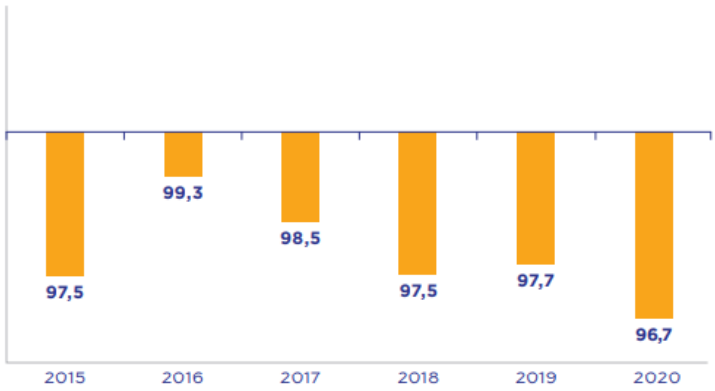


FIGURE 1.1.1 – Evolution du ratio combiné net de réassurance du marché IARD

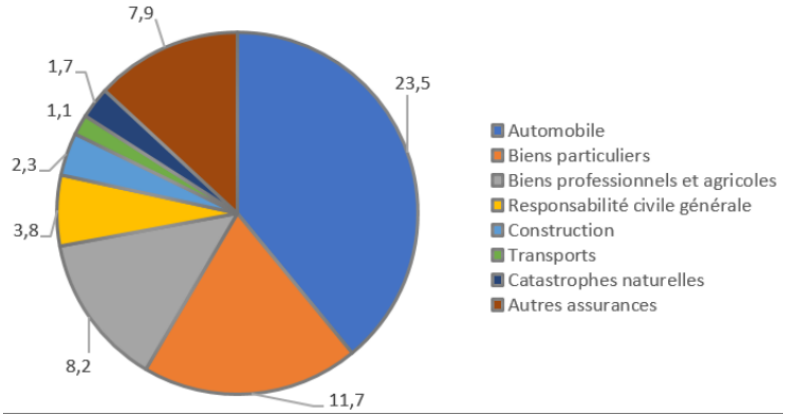


FIGURE 1.1.2 – Repartion des cotisations IARD en 2020 en Mds d'€

1.1.2 L'assurance Multirisque Habitation

Le contrat d'assurance Multirisque Habitation est un contrat multi garanties qui permet de protéger son habitation et son mobilier que l'on soit responsable ou bien victime d'un sinistre.
Cette assurance est obligatoire pour les locataires et les copropriétaires occupants ou non occupants, et elle est facultative pour les propriétaires dont le logement n'est pas en copropriété.

Les biens assurables en MRH :

Dans le cadre d'un contrat d'assurance Multirisque Habitation, les biens assurables sont les suivant :

- Les **bâtiments** appartenant à l'assuré ainsi que leurs aménagements et installations qui ne peuvent être détachés sans être détériorés ou sans détériorer la construction (maison, appartement, greniers, cave, garages, abris de jardins, etc.).

- Le **mobilier personnel** : l'assureur garantit les meubles et objets personnels appartenant à l'assuré, aux membres de sa famille, et à toute autre personne résidant où se trouvant momentanément dans les lieux assurés.

Les garanties en MRH :

Avec une assurance Multirisque Habitation, les dommages assurables sont les suivant :

- La garantie **incendie-explosion** : tous les contrats multirisques offrent en garantie de base la couverture des dommages matériels résultant d'un incendie, d'une explosion, d'une implosion, de la chute de la foudre, ainsi que les dégâts provoqués en éteignant un feu.

- La garantie **dégâts des eaux** : couvre les conséquences d'un dégât des eaux mais n'a pas pour objet l'indemnisation des réparations de la partie de la construction ou de l'appareil à l'origine du dommage.

- Les garanties **vol et vandalisme** : couvrent la disparition, la destruction ou la détérioration des biens mobiliers résultant de vols, tentatives de vol et/ou d'actes de vandalisme commis dans les circonstances prévues au contrat et dont l'assuré doit apporter la preuve.

- La garantie **bris de glace** : couvre les dommages matériels (bris, fissures, etc.) subis par les vitres, les fenêtres, les baies vitrées, les vélux, les garde-corps, les parois séparatives de balcons, ainsi que les verres et glaces du mobilier.

- La garantie **catastrophes naturelles** : est une garantie légale obligatoire. L'assuré est automatiquement couvert contre les dégâts dus aux catastrophes naturelles. La mise en jeu de la garantie par l'assureur

est subordonnée à la constatation de l'état de catastrophe naturelle par un arrêté interministériel publié au Journal officiel.

- La garantie **tempête et autres événements climatiques** : les contrats d'assurance habitation garantissant les dommages d'incendie ou tous autres dommages à des biens situés en France couvrent obligatoirement les effets du vent dû aux tempêtes, ouragans et cyclones.

- La garantie contre les **actes de terrorisme** ou **d'attentats** : elle couvre les dommages matériels liés au terrorisme ou à un attentat.

- La garantie **responsabilité civile vie privée** : elle couvre les conséquences pécuniaires de la responsabilité civile encourue par l'assuré à la suite de dommages corporels, matériels ou immatériels consécutifs causés à des tiers au cours de la vie privée.

- La garantie **protection juridique** fait l'objet d'un contrat distinct ou d'un chapitre distinct à l'intérieur du contrat. Elle consiste à prendre en charge les frais de procédure, en cas de différend ou de litige opposant l'assuré à un tiers dans le cadre d'une procédure civile, pénale, administrative

Le marché de la MRH en détails :

L'assurance Multirisque habitation est le second marché de l'assurance IARD avec un chiffre d'affaire de 11,28 milliards d'€ en 2020 (en augmentation de 3,2 % par rapport à 2019). Concernant la prime moyenne hors taxes, elle est de 260 € en 2020 (en augmentation de 1,7 % par rapport à 2019).

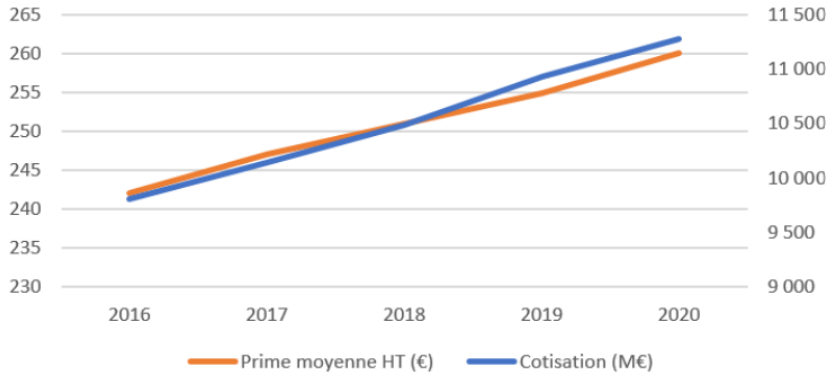


FIGURE 1.1.3 – Evolution du chiffre d'affaire et la prime moyenne sur le marché MRH

Il s'agit donc d'un marché en croissance : en 2020, le marché a progressé d'1,5 % pour atteindre les 43 millions de contrats Multirisque Habitation. Dans le même temps, le nombre de logements en 2020 est

estimé à 37 millions, en hausse de 1,1 %.

Cependant, malgré sa croissance, le marché de l'assurance habitation est profondément concurrentiel ce qui peut affecter son essor. Cette situation s'explique notamment par l'entrée en vigueur de la loi Hamon en 2015 qui permet aux assurés de résilier leurs contrats au bout d'un an, mais également par une concurrence accrue avec les bancassureurs.

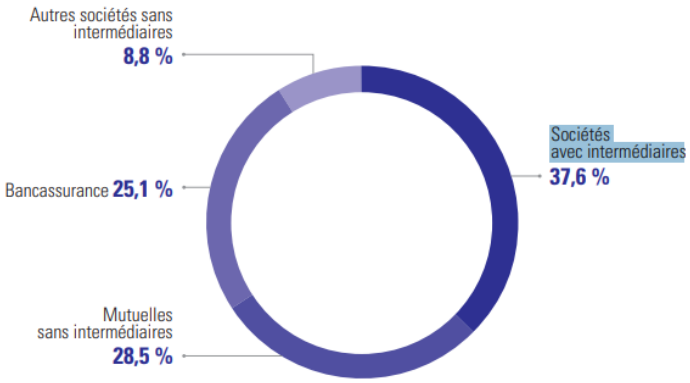


FIGURE 1.1.4 – Répartition des cotisations dommages aux biens des particuliers

1.1.3 AXA France sur le marché de l'assurance Multirisque Habitation

AXA se positionne comme le deuxième assureur sur le marché français de l'assurance Multirisque Habitation. En effet, le chiffre d'affaires de l'entreprise est d'environ 1,2 milliards d'€ et est resté stable entre 2016 et 2020.

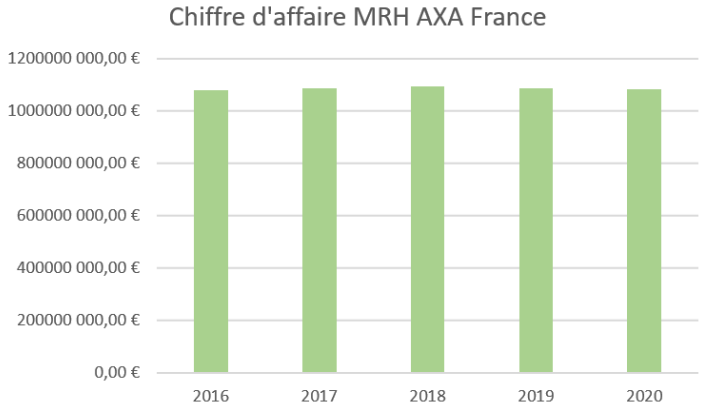


FIGURE 1.1.5 – Evolution du CA MRH AXA

AXA propose actuellement deux produits d'assurance Multirisques habitation :

- Le produit « *Confort* » qui est l'ancien produit MRH d'AXA. Il est destiné à disparaître et ne peut donc être souscrit uniquement dans les DOM et à Monaco. En revanche, les assurés possédant ce produit n'ont aucune obligation à changer. Le produit propose entre autres les garanties suivantes : l'incendie (couvrant également les effets du courant électrique), le dégât des eaux, le bris de glaces, les évènements climatiques, les catastrophes naturelles et technologiques et les attentats et actes de terrorisme, la responsabilité civile ainsi que le vol et le vandalisme.

- Le produit « *Ma Maison* », introduit en 2017, qui est le nouveau produit MRH d'AXA. Ce produit se veut plus flexible par rapport au produit « *Confort* ». En effet, en plus de bénéficier d'un socle commun de garanties, l'assuré peut moduler son contrat à sa convenance avec un large panel d'options à souscrire afin de mieux correspondre à ses besoins. Ainsi le dommage électrique, le bris de glace ainsi que le vol et le vandalisme deviennent des options sous « *Ma Maison* » tandis que le gel devient inclus aux dégâts des eaux par exemple. Les primes des contrats « *Ma Maison* » se veulent alors mieux ajustées aux risques des assurés.

Ainsi aujourd'hui, le portefeuille d'AXA est composé à plus de 30% de contrats « *Ma Maison* ».

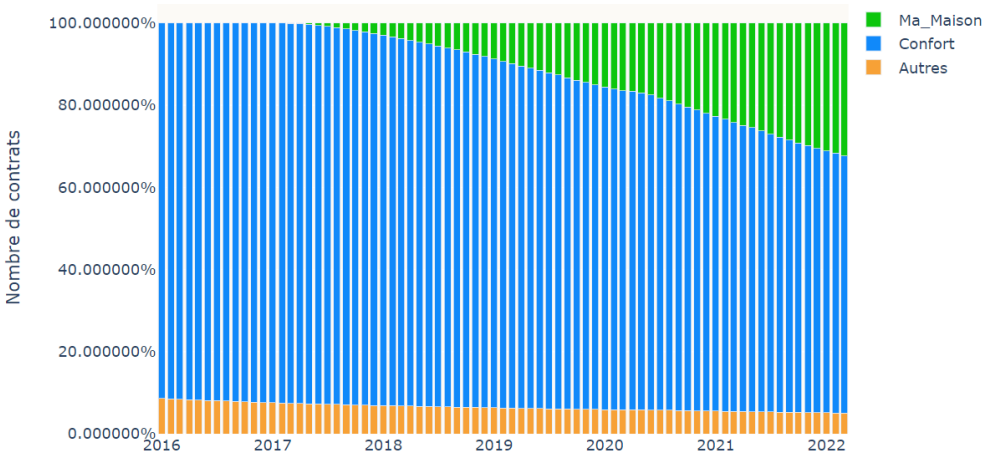


FIGURE 1.1.6 – Répartition des contrats Confort - Ma Maison

1.2. La situation des contrats DOM au sein d'AXA France

Les Outre-Mer françaises se composent des DROM qui sont à la fois départements et régions que sont La Réunion, la Guyane, la Guadeloupe, la Martinique et Mayotte, des COM qui sont des collectivités territoriales, que sont Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Barthélemy, Wallis-et-Futuna et la Polynésie française. S'ajoute également la Nouvelle-Calédonie qui possède un statut spécial. On parle ainsi des DROM-COM autrefois appellés DOM-TOM. Ce mémoire se focalisera ainsi sur les DROM hors Mayotte sous l'intitulé « *DOM* ».

L'analyse du marché ci-dessous concerne cependant tout le marché ultramarin, les données uniquement DOM n'étant pas disponibles.

Note : les chiffres et informations présentés ci-dessous en section 1.2.1 proviennent de l'étude "Le phénomène de non-assurance dans les départements et collectivités d'Outre-mer" (2020) de l'IGF (Inspection Générale des Finances).

1.2.1 L'assurance IARD et les spécificités de la MRH dans les DOM

Les difficultés du marché IARD dans les DOM :

Le marché de l'assurance IARD ultramarin est réputé complexe et difficile pour les assureurs : sont en cause les spécificités de ce marché en matière d'économie de l'assurance ce qui représente un véritable défi technique. Ainsi, du fait du caractère industriel de la MRH (grâce à des primes relativement faibles), il est difficile de segmenter convenablement les risques ce qui empêche en l'état de pratiquer l'anti-sélection des risques.

De même, certains risques ultramarins en assurance dommage aux biens semblent innassurables en raison de l'absence d'aléa, avec par exemple des constructions en zones d'aléa fort ou des bâtis ayant des soucis de constructions ou de résistance. On note également la présence d'habitats sans droit ni titre qu'il est difficile d'assurer.

Enfin, l'étroitesse du portefeuille d'assurés dans les DOM restreint d'une part la possibilité de réaliser une modélisation fine des risques en raison d'un manque d'historique et d'autre part la mutualisation des risques au sein des territoire ultramarins.

Les acteurs du marché IARD dans les DOM :

Le nombre d'acteurs dans les Outre-Mer est limité comparé à la France métropolitaine. En effet,

4 compagnies d'assurances concentrent près de 70 % du marché alors que cette part de marché est partagée par 11 compagnies en métropole. De plus, le marché ultramarin représente une part peu significative du chiffre d'affaire de ces assureurs. Les difficultés liées à ce marché expliquent sans doutes cette situation.

On retrouve ainsi les acteurs principaux suivant :

- **Generali** et ses deux filiales : GFA Caraïbes aux Antilles et Prudence créole dans l'océan Indien. Avec 271 millions d'€ de chiffres d'affaires (0,3 % du CA Generali France) en 2018, Generali concentre près de 25 % du marché IARD dans les territoires ultramarins.

- **Groupama** avec ses caisses régionales GAN Pacifique, GAN Antilles-Guyane et Groupama océan Indien, possède environ 20 % du marché IARD ultramarin.

- **Allianz** est présent dans tous les territoires ultramarins et sur tous les segments de marché IARD (15% de part de marché IARD outre-mer, 1,5 % du chiffre d'affaires d'Allianz France).

- **AXA**, avec 93 M€ (0,9 % du chiffre d'affaires IARD d'Axa France), le groupe concentre une part de marché IARD outre-mer d'environ 9 % pour les particuliers.

- On note également la présence des bancassureurs qui sont en forte croissance sur ce marché.

Le phénomène de non-assurance en MRH dans les DOM :

Malgré une amélioration depuis 25 ans, le taux de souscription d'une assurance Multirisque Habitation dans les DOM est très inférieur à celui de la France métropolitaine :

En % des ménages	1995	2001	2006	2011	2017
Guadeloupe	29	32	44	53	**59**
Guyane	47	38	52	42	**49**
Martinique	39	41	50	52	**62**
La Réunion	29	45	59	62	**68**
Mayotte	N.D.	N.D.	N.D.	6	**6**
Métropole	*96*	*96*	*96*	*96*	*97*

Source : Pôle Science des données de l'IGF, d'après Budget de famille, Insee. Sur la résidence principale. L'échantillon pour la France métropolitaine est constitué de 10 342 ménages et l'échantillon Drom est constitué de 5 455 ménages.

FIGURE 1.2.1 – Taux de souscription à l'assurance multirisque habitation dans les DOM

Cette situation s'explique d'abord par un niveau de vie moins élevé dans ces territoires : l'assurance des biens ne fait pas partie des dépenses prioritaires. Cela n'est cependant pas la seule explication, on peut en effet évoquer la faiblesse de l'offre d'assurance du fait de la difficulté du marché.

1.2.2 Analyse commerciale des contrats MRH DOM d'AXA France

Etude du portefeuille DOM :

Les contrats DOM en MRH représentent 1,2 % du portefeuille MRH d'AXA France, soit environ 40 000 contrats début 2022. Cette situation fait écho à la situation du marché IARD dans les DOM évoquée plus haut.

Parmis ces 40 000 contrats, plus de la moitié se trouvent à La Réunion :

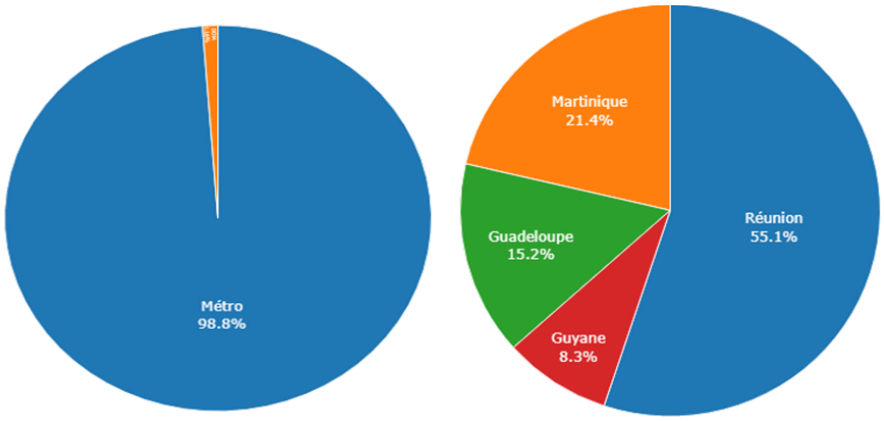


FIGURE 1.2.2 – Répartion du portefeuille MRH d'AXA France

Le portefeuille MRH DOM a vu une croissance de près de 6 % entre 2016 et 2020, là où le portefeuille MRH Métropole est en baisse :

Les pentes tarifaires suivent la même trajectoire entre la métropole et les DOM, mais les prime y sont en moyenne environ 50 € plus élevées. Dans le détails, les primes en Guadeloupe et en Martinique se rapprochent de celles de la métropole alors qu'elles sont plus basses à La Réunion et davantages en Guyane.

Etude des affaires et résiliations dans les DOM :

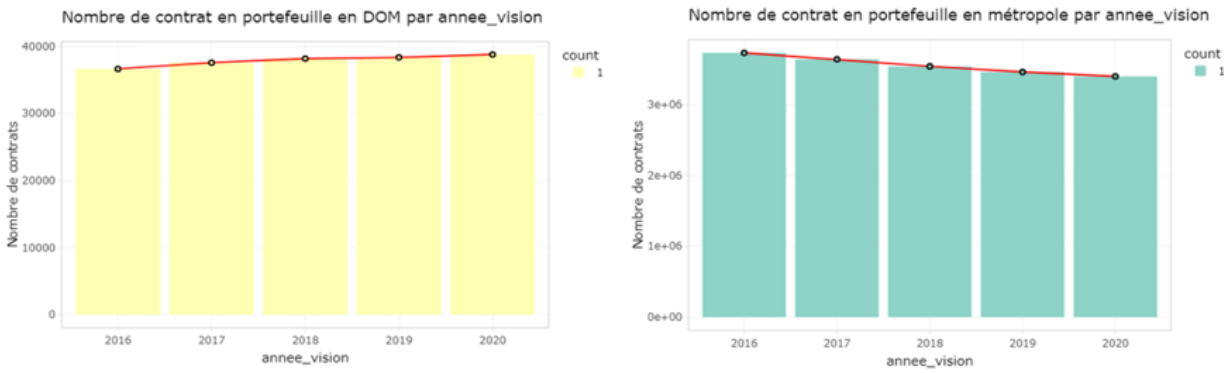


FIGURE 1.2.3 – Evolution du nombre de contrats dans les DOM et en Métropole

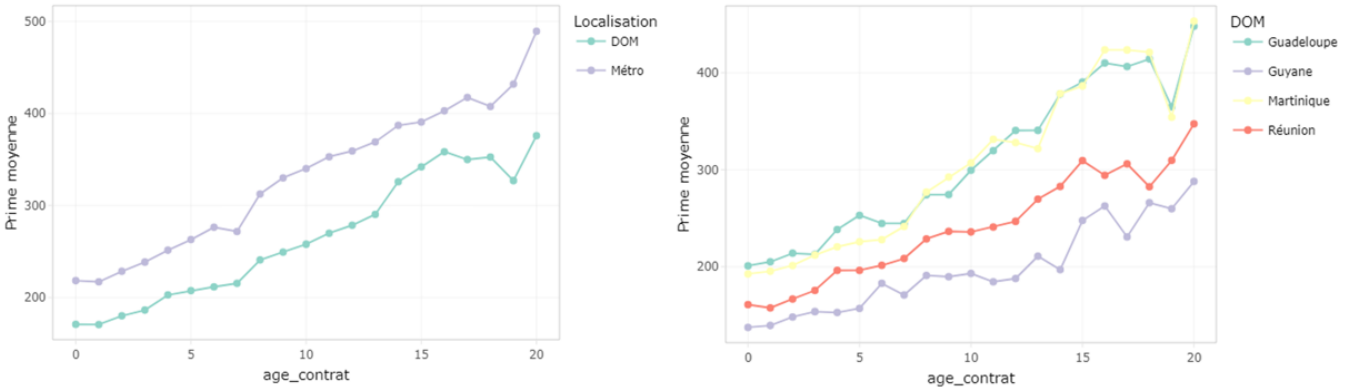


FIGURE 1.2.4 – Evolution de la prime moyenne par ancienneté

1.2.3 Analyse de la sinistralité des contrats MRH DOM d'AXA France

La répartition du nombre sinistres dans les DOM :

Les 4 unités de prestations (ou GUP) les plus répandues dans les DOM sont le dégât des eaux (44 %), le vol (15 %), le dommage électrique (10 %) et les évènements climatiques (9 %) qui représentent plus de 80 % des sinistres entre 2009 et 2020. Par rapport à la métropole, il y a une sur-représentation du dégât des eaux (44 % dans les DOM contre 39 % en métropole) et du vol (15 % dans les DOM contre 10 % en métropole) et on observe en revanche bien moins de bris de glace (3 % dans les DOM contre 11 % en métropole). La part du dégât des eaux représente plus de 60 % des sinistres dans les DOM depuis 2018.

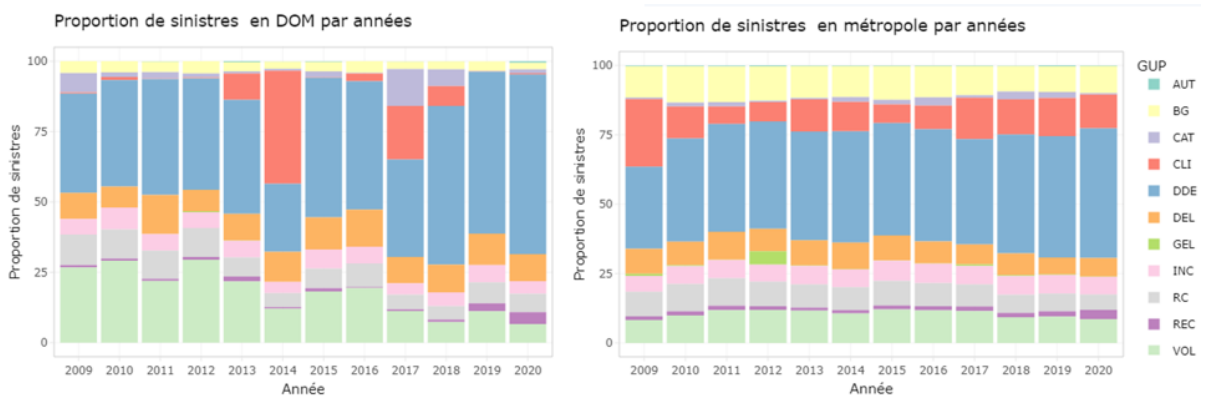


FIGURE 1.2.5 – Evolution de la proportion de GUP par années en Métropole et dans les DOM

La répartition de la charge de sinistres dans les DOM :

Le dégât des eaux est l'unité de prestation la plus représentée en charge dans les DOM du fait de la proportion plus élevée de ce type de sinistre avec près de 24 % de la charge annuelle totale. En métropole, la part du dégât des eaux en charge annuelle est également de 24 % . Ainsi, un sinistre dégât des eaux est en moyenne moins cher de 300 € dans les DOM par rapport à la métropole (1 500 € en métropole contre 1 200 € dans les DOM sur notre historique de 12 ans). Le vol en revanche représente près de 20 % de la charge annuelle totale dans les DOM contre environ 12 % en métropole, il est en moyenne 100 € plus cher en métropole (2 900 € en métropole contre 2 800 € dans les DOM sur notre historique de 12 ans).

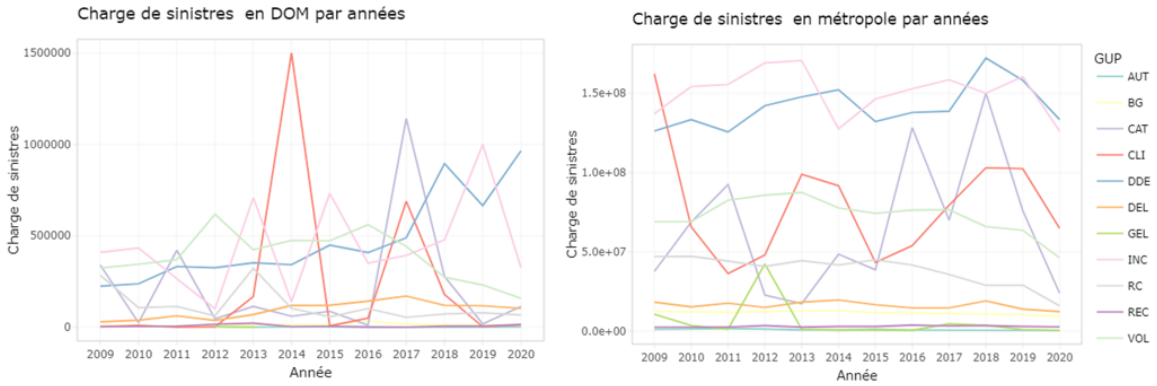


FIGURE 1.2.6 – Evolution de charge totale de sinistres par années en Métropole et dans les DOM

Focus sur l'évolution du sinistre dégât des eaux :

La fréquence des sinistres dégâts des eaux a augmenté d'année en année, elle était d'environ 0,6 % en 2009 pour finir à environ 1,5 % en 2020 avec un pic à 1,6 % en 2018. De même le coût moyen d'un sinistre dégât des eaux est globalement en hausse et notamment depuis 2017 où il était d'environ 1 100 € pour arriver à plus de 1 500 € en 2020. Ainsi, la sinistralité de ce type de sinistre est clairement en hausse et ce mémoire se focalisera donc sur cette garantie.

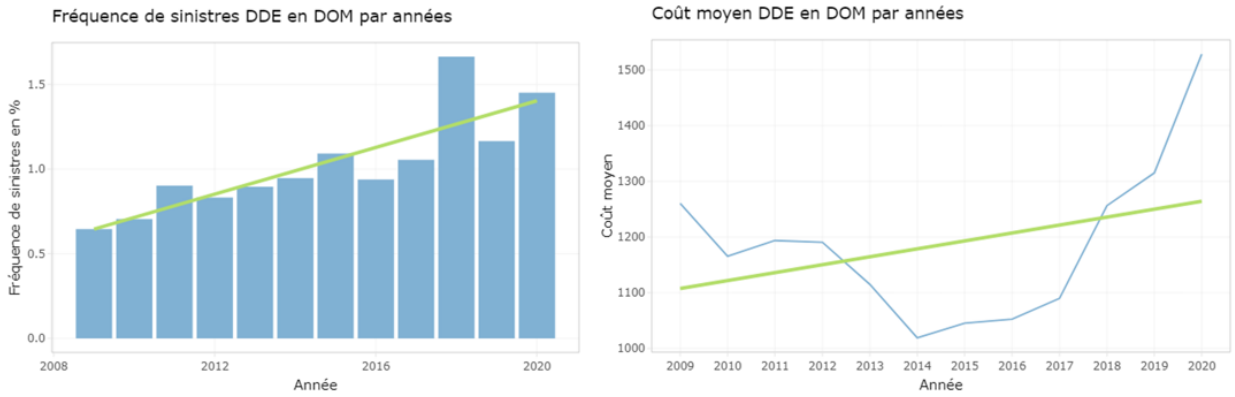


FIGURE 1.2.7 – Evolution fréquence de sinistre et du coût moyen du DDE dans les DOM

Evolution du ratio S/C dans les DOM :

Le ratio "sinistres sur cotisations" noté S/C est un indicateur de rentabilité d'un portefeuille utilisé notamment en assurance non-vie. Il est défini selon la formule suivante :

$$\mathbf{S/C} = \frac{\textbf{Somme des charges}}{\textbf{Somme des cotisations}}$$

Ainsi, si le ratio S/C est égal à 100, cela signifie que l'assureur reçoit juste assez de cotisations pour rembourser les sinistres de ces assurés. Pour être rentable, le S/C doit donc être absolument inférieur à 100. Notons qu'en plus de charges sinistres, l'assureur a d'autres coûts à couvrir (frais de réassurance, commissions, frais de gestion...). Ces frais sont eux pris en compte dans l'indicateur du ratio combiné qui ajoute les frais au montant des charges sinistres.

Comparons à présent les ratios S/C de la métropole et ceux des DOM :

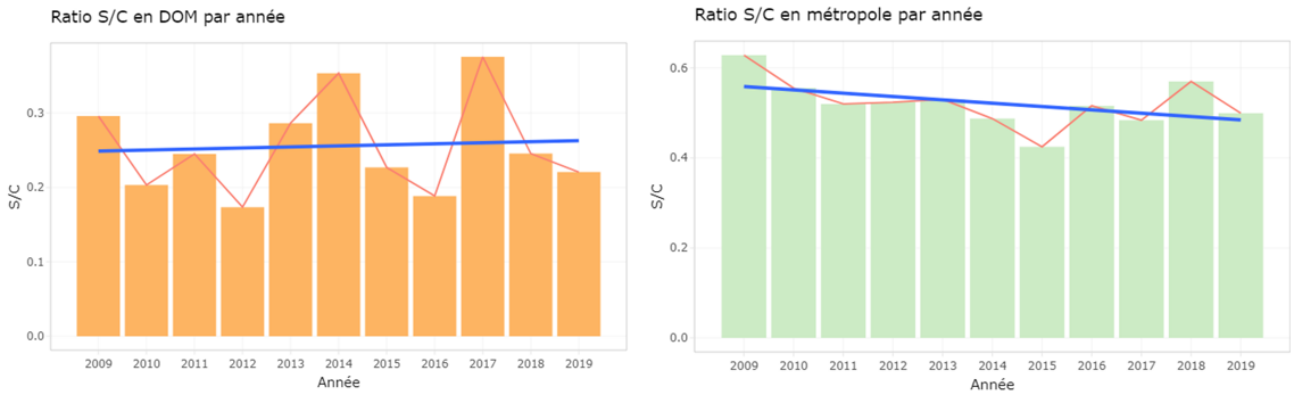


FIGURE 1.2.8 – Evolution des ratios S/C entre les DOM et la métropole

Les ratios S/C les plus élevés dans les DOM (environ 35%) correspondent aux années 2014 et 2017 où la charge de sinistres était exceptionnellement élevée du fait de sinistres concernant la garantie liée aux évènements climatiques. A l'inverse le ratio S/C le plus faible (environ 17%) correspond à l'année 2012. L'évolution des ratios S/C est plus instable dans les DOM par rapport à la métropole ce qui peut s'expliquer par une plus faible volumétrie de contrats dans les territoires ultramarins.

Enfin, le S/C dans les DOM est en moyenne deux fois moins élevé (à 25,6%) que celui de la métropole (à 52,2%). Cette situation fait des DOM un segment particulièrement rentable par rapport à la métropole (en considérant des frais équivalents).

1.3. Objectifs de l'étude

1.3.1 Le besoin de simplification des processus tarifaires

Cette étude se place dans le cadre d'une simplification des processus tarifaires au sein de l'équipe MRH d'AXA France. L'équipe MRH d'AXA France doit en effet développer et maintenir une multitude de modèles dans le but d'expliquer au mieux la sinistralité de son portefeuille. De fait, le nombre de modèle existant est de 20 uniquement pour le segment "MRH Classique" (les contrats situés en France métropolitaine hors DOM, mobilhome, et propriétaire non occupant, identifiée en tant que segment "métropole" dans la suite du mémoire). En théorie le nombre idéal de modèle est :

$$\textbf{N.modeles} = \textbf{N. garanties} \times \textbf{Segmentation appartement/maison} \times \textbf{N. segments}$$

Ainsi, avec un nombre de garanties égal à 10, la segmentation appartement/maison égale à 2 et le nombre de segment égal à 4 (MRH classique, DOM, Propriétaire non occupant et mobilhomes), cela donnerait un total de 80 modèles à développer et maintenir dans la durée. Or, d'un point de vue opérationnel, réaliser autant de modèles peut s'avérer complexe en terme de temps et de ressources humaines. C'est pourquoi l'idée d'une simplification du processus tarifaire est essentielle afin de n'avoir qu'un nombre raisonnable de modèles à actualiser.

Le processus de simplification tarifaire au sein de l'équipe MRH d'AXA France consiste à intégrer des segments à la modélisation de la sinistralité "MRH Classique". En effet, le segment métropole représentant environ 98% du portefeuille MRH d'AXA France, il peut paraître cohérent d'intégrer la modélisation des segments à bien plus faible volumétrie au segment regroupant la quasi-totalité de la masse des contrats. Cependant, rien n'indique que les risques de ces segments puissent se confondrent avec ceux du segment métropole, la nature des risques entre segments peut être intrinsèquement différente, notamment en prenant en compte la géographie, on peut par exemple penser aux risques des mobilhomes qui se situent majoritairement sur les littoraux.

Par conséquent, pour valider la simplification tarifaire d'un segment hors "MRH Classique" en l'intégrant à la modélisation du segment "MRH Classique", il faut que le modèle regroupant les deux segments ait un pouvoir prédictif équivalent (on peut également accepter un modèle détérioré dans une certaine mesure) à un modèle basé uniquement sur le segment hors "MRH Classique" sur une base de données hors "MRH Classique". Il n'est en effet pas nécéssaire de regarder les performances sur une base de données "MRH Classique", représentant 98% de la volumétrie, le modèle ne devrait pas impacté les performances par rapport au modèle sans l'ajout du nouveau segment.

1.3.2 Simplification de la tarification et revue de la prime DOM

Dans cette partie, nous allons voir en détails la simplification de la tarification des contrats DOM. Comme nous l'avons vu précedemment, cette démarche va consister en l'intégration des contrats situés dans les DOM dans la modélisation des contrats situés en métropole. Notre étude va se concentrer sur la garantie dégâts des eaux, mais une extension aux garanties incendie, vol, bris de glace, responsabilité civile et climatique sera réalisée en vue d'une analyse des résultats d'un point de vue commercial à la fin de ce mémoire.

L'objectif va être de réaliser un modèle basé sur les contrats "DOM" et un modèle basé sur les contrats "Métropole + DOM" afin de comparer leur pouvoir prédictif sur une base de données contenant des contrats DOM. Si les performances de ce nouveau modèle sont semblables au modèle basé uniquement sur les contrats DOM, alors l'intégration sera justifiée (nous verrons dans la suite sur quelles métriques seront jugés les performances des modèles).

Nos modélisations de la sinistralité dans les DOM nous permettrons de revoir le tarif sur ce segment. En effet la prime commerciale dans les territoires ultramarins est basée sur un modèle ancien comportant un faible nombre de variables ce qui implique donc une faible segmentation. De plus il n'y a pas de zonier à la maille INSEE (nous détaillerons les différentes mailles géographique dans la suite), mais seulement des coefficients différents par DOM. Nous pourrons donc jugé si l'intégration d'une segmentation géographique plus fine et nécéssaire ou non.

Cependant, bien que la prime commerciale doit restée technique (proche de la sinistralité), il faut garder à l'esprit la rentatibilité du produit, d'autant plus que ce segment est déjà relativement rentable. Ainsi, notre nouvelle segmentation ne doit pas se faire au détriment des enjeux commerciaux. Il faut trouver le bon équilibre entre mutualisation et segmentation, et on voudra, par exemple, peu d'écarts de primes commerciales entre des risques proches (notamment d'un point de vue géographique).

Enfin, toute cette démarche, va au final, permettre d'intégrer les contrats situés dans les DOM au nouveau produit MRH d'AXA France : "Ma Maison", ce qui va permettre d'analyser tous les contrats sur la même base dans le futur.

Chapitre 2

Préparation des données

2.1. Les bases de données utilisées

2.1.1 La base image

La base image est la base de données centrale dans toute étude en MRH au sein d'AXA France. Elle contient les informations concernant les contrats des assurés avec les spécificités de leurs habitations telles que : le type de logement, le nombre de pièces, le capital mobilier, la localisation du bien. A cela s'ajoute les informations liées à la souscription du contrat telles que : la date d'affaire nouvelle, la date de résiliation, la cotisation, les options souscrites. Dans la suite, on utilisera certaines des caractéristiques des contrats en tant que variable lors de la modélisation afin de segmenter le risque au mieux.

La base image est une base par image de risque. En effet, tout au long de la vie du contrat, le risque peut en effet évoluer du fait d'un remplacement à cause de modifications d'informations dans le cas d'un déménagement ou encore d'un changement de garantie par exemple . Ces actions ont pour effet d'ajouter une nouvelle ligne dans la base image, car on va considérer qu'il s'agit alors de risque différents, mais les numéros de contrats concernés resteront inchangés. De même, à chaque passage à un nouvel exercice, on considère qu'il s'agit d'un nouveau risque et une nouvelle ligne est ainsi ajoutée.

La base image est mise à jour chaque mois. On a donc une image mensuelle de tous nos contrats. Dans notre cas, on souhaite une base annualisée, on récupère donc la dernière image du risque de l'exercice considéré pour chaque contrat.

Afin d'avoir un maximum de données pour la consistance de notre modélisation de la sinistralité dans les DOM, le périmètre d'étude est composée des contrats en portefeuille sur la période du 01/01/2009 au 31/12/2019 soit 11 années d'historique. En effets, les DOM représentant environ 1% du portefeuille, on a

besoin de plus d'historique, là où 3 années auraient suffies si l'on modélisait uniquement les contrats situés en France métropolitaine.

La base image MRH est construite grâce aux données des contrats d'AXA France IARD. Il s'agit d'un regroupement de données provenant des "régions AXA" qui découpent la France métropolitaine avec les régions Ile-de-France, Nord-Est, Ouest, Sud-Est et Sud-Ouest, à cela s'ajoute la région DOM et une région "AXA partenaires" qui n'a pas de réalité géographique mais qui représente les collaborateurs d'AXA. Ainsi, avec la base image, on dispose directement des infomations concernant les contrats situés dans les DOM, et aucun traitement supplémentaire n'est à effectuer.

2.1.2 Les bases sinistres

Les bases sinistres contiennent les caractéristiques des sinistres en portefeuille avec les libellés des garanties impactées (les unites de prestations), la date de survenance, le numéro du sinistre, l'état du sinistre (en cours ou clos), les réglements, les recours et les réserves. On obtient la charge totale par sinistre avec la formule suivante :

$$\textbf{Charge totale} = \textbf{reglement principal} + \textbf{reserves} + \textbf{recours}$$

Les bases sinistres sont créées par année de survenance à différentes visions des sinistres. On reprend ici le même périmètre d'étude que pour la base image, à savoir les sinistres dégâts des eaux survenus de 2009 à 2019, toujours dans l'optique d'avoir un maximum de données. Les sinistres sont à vision 31/12/2020 afin d'avoir une année de développement des sinistres supplémentaires pour les sinistres survenus en 2019, ce qui est généralement suffisant pour avoir une bonne vision de la charge totale des sinistres dégâts des eaux. De plus l'étude se faisant de 2009 à 2019, la majorité des sinistres sont clos et on a ainsi la vision finale de leurs charges.

La récupération des données concernant les sinistres survenus dans les DOM a été plus délicate que dans le cas de la base d'image. En effets, nos bases sinistres ne contenaient pas l'intégralité du périmètre DOM. De ce fait, nous avons dû retirer les filtres nous empêchant d'obtenir l'entièreté des données DOM dans notre code. Il a fallu ensuite s'assurer de la validité du nombre de sinistre obtenus par années de survenance et les réconcilier avec les données de l'équipe actuariat en charge des Outre-Mer basée à Marseille.

2.1.3 Les autres bases

Deux autres bases de données vont être utilisées lors de notre étude, l'un interne et l'autre externe.

Les dernières bases de données interne que nous avons utilisé sont les bases Client. Celle-ci contiennent les caractéristiques propres aux assurés telles que leur âge, l'ancienneté de leur contrat MRH, leur ancienneté dans le portefeuille IARD AXA (un assuré peut en effet être multi-détenteur de contrats, MRH et Auto notamment), le statut marital, le nombre d'enfant et la catégorie socio-professionnelle entre autres. Seuls l'âge du client, l'ancienneté de son contrat et MRH et son ancienneté dans le portefeuille nous intéresseront dans notre étude. Ces bases de données ayant un historique allant jusqu'en 2013, il suffira de retirer le nombre d'années nécéssaires pour les contrats de 2009 à 2012 lorqu'on intégrera ces informations à la base de données servant à la modélisation. Par exemple pour l'année 2010, on retire trois années à l'âge des clients et de leurs contrats afin d'avoir la vision au 31/12/2010.

Enfin nous allons utiliser une base de données externe contenant la liste actuelle des codes INSEE, codes postaux et noms de communes des territoires ultramarins étudiés (Guadeloupe, Martinique, Guyane et Réunion) et une base contenant ces mêmes caractéristiques mais prenant en compte les anciens découpages administratif. En effet, notre problématique concernant les codes INSEE dans les DOM est la suivante : notre étude allant de 2009 à 2019, nos bases de données contiennent des codes INSEE ne correspondant pas à la réalité administrative actuelle. Certains quartiers d'une même commune dans les DOM possédaient leur propre code INSEE alors qu'aujourd'hui ces codes ont été regoupés en un unique code INSEE. On va ainsi utiliser ces deux bases de données afin de construire une table de correspondance pour obtenir les codes INSEE actuellement en vigueur. Regardons un exemple avec la commune de Saint-Denis de La Réunion :

Commune	Code INSEE	Code postal
Saint-Denis	97411	97400
-	-	97417
-	-	97490

TABLE 2.1.1 – Code INSEE et postaux actuels de la commune de Saint-Denis de La Réunion

Le code INSEE de cette commune est le 97411, mais on remarque qu'elle possède trois codes postaux. On va se servir de ces codes postaux pour réaliser la correspondance avec les anciens codes INSEE :

Commune	Code INSEE	Code postal
Saint Denis	974411	97400
Saint Denis Tadar	974425	97400
Saint Denis Camelias	974445	97400
Le Brule	974464	97400
Saint Francois	974470	97400
Belle Pierre	974475	97400
Saint Denis	974411	97417
Saint Bernard	974442	97417
La Montagne	974447	97417
Saint Denis	974411	97490
Sainte Clotilde	974454	97490
Saint Denis Chaudron	974459	97490
Bois de Nefles	974473	97490
Moufia	974489	97490
La Bretagne	974499	97490

TABLE 2.1.2 – Anciens codes INSEE et postaux de la commune de Saint-Denis de La Réunion

On remarque ainsi que l'on a 15 codes INSEE différents pour la commune de Saint-Denis qui deviendront donc "97411" après correspondance. On construit de cette manière la table de correspondance pour chaque commune située dans les DOM. De même, des erreurs probablement d'origine humaine ont indiqué les codes postaux à la place des codes INSEE dans nos bases de données, il a donc fallu corriger ces valeurs. De fait, grâce à ces correspondances et corrections, on passe de plus de 200 codes INSEE initialement présents dans notre base de données à seulement une centaine au final. Nous verrons par la suite que cela nous sera grandement utile lors de la modélisation géographique du risque.

2.2. Gestion des sinistres graves

2.2.1 Problématique

Lors de la modélisation de la sinistralité, on prend pour hypothèse l'homogéneité des risques de notre portefeuille à des fins de mutualisation de la charge totale. C'est dans ce contexte que nous allons faire la distinction entre les sinistres graves qui ont des charges élevées mais de faibles fréquences d'occurence, des sinistres attritionnels avec des montants de charge plus classique mais des fréquences d'occurence plus importantes. Ainsi, dans notre modélisation, nous allons séparer ces deux types de sinistres en écrêtant les charges selon un seuil que nous allons fixer. De ce fait on aura :

$$\textbf{Charge grave} = (\textbf{Charge reelle} - \textbf{Seuil}) \ . \ \mathbb{1}_{\textbf{Charge reelle} > \textbf{Seuil}}$$

et :

$$\textbf{Charge attritionnelle} = \textbf{max}(\textbf{Charge reelle} \ , \ \textbf{Seuil})$$

La charge, grave qui est la partie résiduelle, sera alors modélisée de manière plus uniforme (avec moins de segmentation pour une meilleure mutualisation de ces montants élevés) contrairement à la charge attritionnelle qui sera bien plus segmentée et proche du risque réel.

Notre problématique réside donc dans le choix de notre seuil de sinistre grave qui va nous permettre de distringuer la sinistralité attritionnelle de la sinistralité grave. Cependant, dans le cadre de notre étude, nous allons avoir d'une part la sinistralité des DOM et d'autre part celle de la France métropolitaine. On a ainsi deux seuils de graves à déterminer. Or la finalité étant de comparer des modélisations (DOM et Métropole + DOM) de la sinistralité attritionnelle, celle-ci doit être écrêtée avec le même seuil pour rendre les comparaisons de modèles pertinentes. La question est donc de savoir si les deux sinistralités ont le même seuil de grave (ou bien un seuil suffisamment proche).

Regardons alors les distributions de la charge de la garantie dégât des eaux dans les DOM et en métropole :

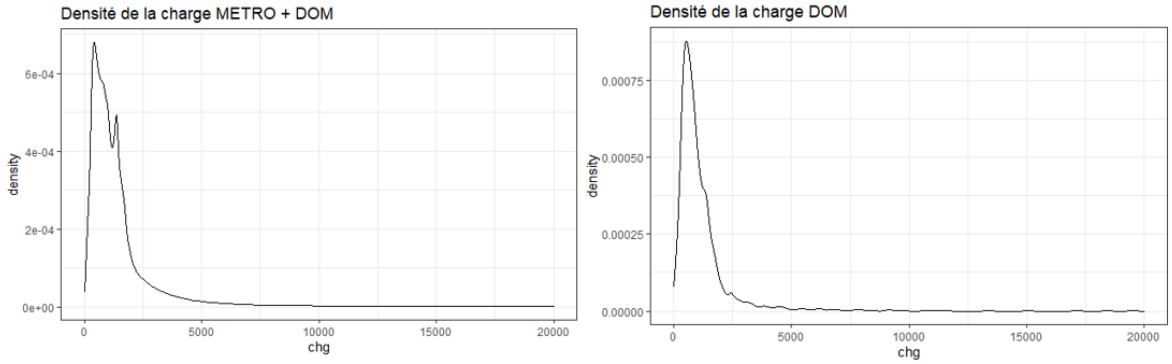


FIGURE 2.2.1 – Distribution de la charge de la garantie DDE dans les DOM et métropole+DOM

Les deux distributions semblent avoir la même forme, bien que la distribution dans les DOM montre une probabilité plus faible d'avoir des sinistres avec un montant important.

On peut regarder en détail si le fait d'être dans les DOM ou non a un impact sur la distribution des charges. Soit X la variable aléaoire égale à la charge d'un sinistre et Y la variable aléatoire de Bernouilli indiquant si le sinistre a eu lieu dans les DOM ou non. On va donc réaliser le test U de Mann-Whitley afin de savoir si les deux groupes (Y=1 ou Y=0) ont la même distribution. Le test de Mann-Whitley est un test non paramétrique et réputé simple qui permet de tester si la distribution de deux groupes est proches. On utile l'hypothèse nulle H_0 : " Les deux distributions sont égales ".

Si on a deux populations N et M de tailles respectives n et m, on calcule la statistique :

$$\mathbf{U} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{m}} \begin{cases} 1 \ si \ N_i \ < M_j \\ \frac{1}{2} \ si \ N_i \ = M_j \\ 0 \ si \ N_i \ > M_j \end{cases}$$

On applique ce test statistique sur les variables aléatoires X|Y=1 et X|Y=0 que l'on a groupé en 20 quantiles. On retrouve alors une p-value inférieure à $2,2.10^{-16}$ (bien plus faible que notre seuil de 5%), on rejette donc l'hypothèse et on ne peut pas considérer que la distribution des charges de sinistres soit la même en métropole et dans les DOM.

2.2.2 Modélisation du seuil de sinistres graves

Nous avons déterminé que la distribution des charges de sinistres n'était pas la même selon la localisation du sinistre (DOM ou métropole). On va ici devoir déterminer les deux seuils de sinistres graves. Pour cela, on va utiliser la théorie des valeurs extrêmes qui proposent de multiples méthodes pour déterminer ce seuil permettant de séparer les sinistres attritionnels des graves. En effet, avec le théorème de Pickands (1975), on déduit que sous les bonnes conditions et pour un seuil u suffisemment élevé, la loi de Pareto généralisée (GPD) est une très bonne approximation de la loi des excès F_u d'une variable aléatoire X définit par :

$$\mathbf{F_u(x) = \mathbb{P}[X - u < x | X > u],\ x \geq 0}$$

et la distribution de Pareto généralisée est définie ainsi pour $\sigma(u) \geq 0$:

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

Pour estimer le seuil de graves on définit la fonction moyenne des excès e :

$$\mathbf{e(u) = \mathbb{E}[X - u | X > u]}$$

On utilise alors des méthodes graphiques afin d'estimer le seuil le plus approprié : en supposant que la distribution GPD est valide pour modéliser les excès d'un seuil u, on peut montrer que pour v > u, la fonction e(v)=E[X-v | X > v] est une fonction linéaire de v de pente $\frac{\xi}{1-\xi}$. De ce fait, le graphique de la fonction moyenne des excès (le mean excess plot) devrait rester raisonnablement proche d'une fonction linéaire (*Kratz, 2020,* **[1]**). On doit donc déterminer visuellement le seuil à partir duquel le mean excess plot est linéaire. Regardons donc les mean excess plot pour les DOM et la métropole + DOM :

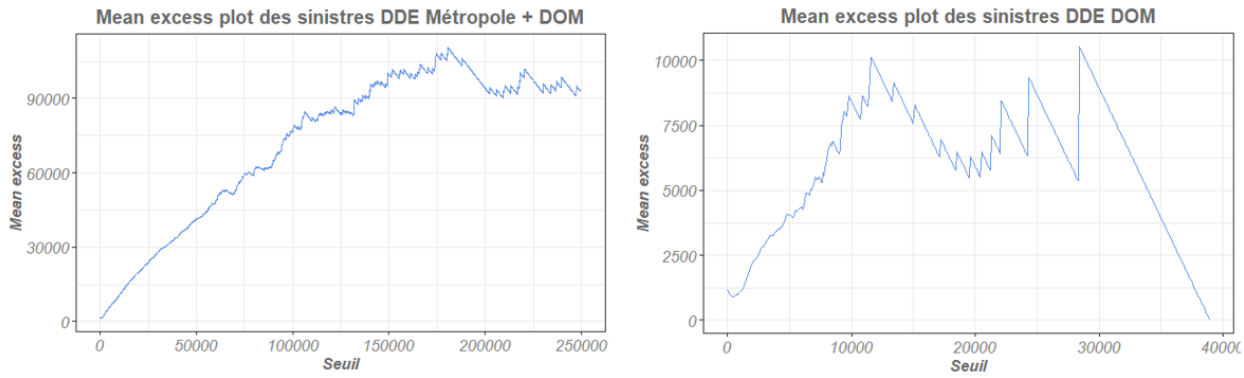


FIGURE 2.2.2 – Mean excess plot de la charge de la garantie DDE dans les DOM et métropole+DOM

On remarque premièrement que le mean excess plot est moins stable dans les DOM du fait du nombre de données plus faible comparé à la métropole + DOM. Dans l'un cas ou l'autre il paraît difficile d'estimer visuellement le point à partir duquel la courbe devient linéaire. Pour répondre à cet problématique on peut réaliser une régression linéaire "glissante" à partir d'un point et sur une fenêtre d'un nombre défini de point. Cela sera effectué dans le but d'obtenir la statistique du R^2 aussi appelé coefficient de détermination. En notant y_i les valeurs réelles, $\hat{y}_i$ les valeurs prédites et $\bar{y}$ la moyenne des valeurs prédites, cette statistique est définie dans le cadre d'une regression linéaire simple par :

$$\mathbf{R^2} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})^\mathbf{2}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \bar{\mathbf{y}})^\mathbf{2}}$$

Le coefficient de détermination va nous permettre de juger de la qualité de nos régression linéaires. On va donc implémenter ces regressions sur une fenêtre de 1000 points dans le cas de la métropole + DOM et de 100 points dans le cas des DOM. On intègre ensuite les valeurs du R^2 dans un tableau afin de les comparer. Voici un exemple sur les DOM de la forme de ces tableaux :

Seuil	R^2
0	0,9840
100	0,9837
200	0,9827
300	0,9811
400	0,9804
....	...

TABLE 2.2.1 – Coefficient R^2 en fonction du seuil de grave dans les DOM :

Ici, il ne suffit pas de relever le R^2 le plus élevé pour définir notre seuil de grave comme nous allons le voir en traçant le coefficient de détermination en fonction du seuil :

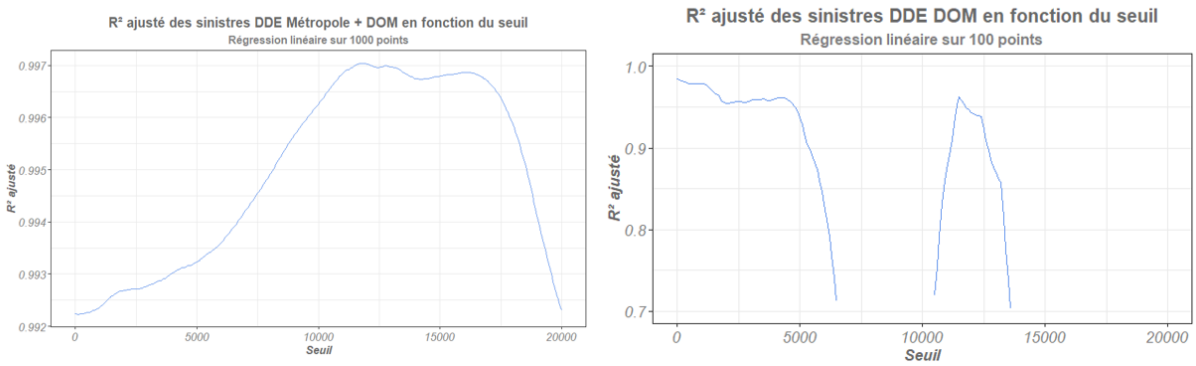


FIGURE 2.2.3 – Seuil de grave en fonction du R^2 pour la garantie DDE dans les DOM et métropole+DOM

Grâce à ce graphique on a une meilleure idée du seuil à choisir : dans le cadre du segment métropole + DOM, le R^2 est maximisé pour un seuil de grave d'environ 11 500 €. Pour le cas du segment DOM, on remarque que le graphique est en deux parties, la seconde moitié correspondant en réalité à la pente descendante du mean excess plot, ce qui ne nous intérèsse pas : dans la première partie du graphique, on a dès le début les valeurs de R^2 les plus élevées mais qui correspondent à des seuils relativement faibles, on a ensuite une décroissance de la courbe avant de remonter pour atteindre un extremum local, c'est là qu'on va choisir un seuil qui se situe aux alentours des 4 500 €.

Une des conditions de validité du théorème de Picklands énoncé plus tôt est que la fonction des excès suivent une loi de Pareto Généralisée, on doit donc vérifier cette condition. Pour cela, on ajuste les excès au delà du seuil à une distribution GPD afin d'estimer ses paramètres de forme et de position grâce à l'estimateur du maximum de vraisemblance. Cela va nous permettre de tracer un graphique quantile-quantile (ou QQ-plot) qui représente les quantiles théoriques d'une GPD ayant les paramètres que nous avons estimé en fonction des quantiles empiriques. Il est alors possible de déterminuer visuellement si notre loi empirique est bien en adéquation avec la loi théorique : si le QQ-plot forme une droite linéaire, alors les deux lois sont effectivement en adéquation et l'utilisation de la méthode du mean excess plot est justifiée.

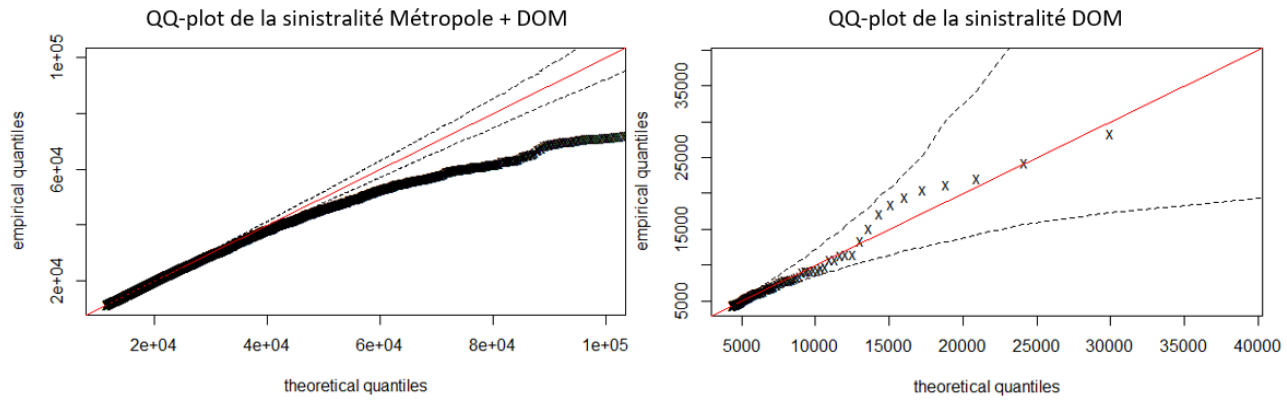


FIGURE 2.2.4 – QQ-plot de la fonction des excès pour la garantie DDE dans les DOM et métropole+DOM

En ce qui concerne le QQ-plot pour le segment métropole + DOM, nos données suivent une loi de Pareto généralisée jusqu'aux sinistres ayant un montant de 45 000 € environ (qui représentent 0,05 % des sinistres). Pour le segment DOM, les données suivent une loi GPD jusqu'aux sinistres ayant un montant de 13 000 € environ (qui représentent 0,2% des sinistres). On s'aperçoit donc qu'à partir d'un certain seuil, nos sinistres peuvent suivre une loi différente. Dans ces cas là, il peut être utile de déterminer un seuil supérieur à celui du seuil de graves qu'on peut le seuil des sinistres atypiques. Dans le cadre de notre étude, on ne prendra en compte que les sinistres graves. De ce fait, au vu de nos données, on considèrera que l'hypothèse d'une loi de Pareto généralisée pour la fonction des excès est valide.

On a ainsi déterminer deux seuils de graves différents pour le segment métropole + DOM (11 500€) et le segment DOM (4 500€) grâce à la méthode du mean excess plot. On peut cependant challenger cette méthode par d'autres méthodes provenant également de la théorie des valeurs extrêmes. Dans le cadre de notre, nous allons utiliser l'estimateur de Hill qui permet d'estimer l'indice de queue $\xi = \alpha^{-1}$. En notant n le nombre d'observations, k le nombre d'excès (avec $k \leq n$) , et $X_{(i),n}$ les statistiques d'ordre d'un échantillon $X_1, ..., X_n$, on le définit ainsi :

$$\mathbf{H_{k,n}} = \frac{\mathbf{1}}{\mathbf{k}} \sum_{\mathbf{i=0}}^{\mathbf{k-1}} \mathbf{ln}(\frac{\mathbf{X_{(i),n}}}{\mathbf{X_{(k),n}}})$$

Cet estimateur converge en probabilité vers α^{-1} (il s'agit d'un estimateur faiblement consistant). Ainsi en traçant l'estimateur de Hill en fonction de k (le nombre d'excès), on cherche à trouver une région stable du graphique afin de pouvoir déterminer le seuil de grave (k pouvant directement être rattaché au seuil).

Une des conditions pour pouvoir appliquer la méthode de l'estimateur de Hill est que la distribution concernée soit à queue lourde ($\xi > 0$). On va donc utiliser nos résultats précédents quant à l'ajustement de la loi des excès à une loi de Pareto généralisée. Notre estimation par le maximum de vraisemblance du paramètre ξ pour le segment DOM est de 0,406 avec un intervalle de confiance à 95% $I = [0,145; 0,667]$, ce qui montre que la distribution de la charge pour les DOM est bien à queue lourde. De la même manière, pour le segment métropole + DOM, l'estimation du paramètre ξ est de 0,523 avec un intervalle de confiance à 95% $I = [0,490; 0,556]$ qui montre que la distribution est également à queue lourde. Dans les deux cas, on peut donc utiliser la méthode graphique de l'estimateur de Hill que nous allons voir ci-dessous :

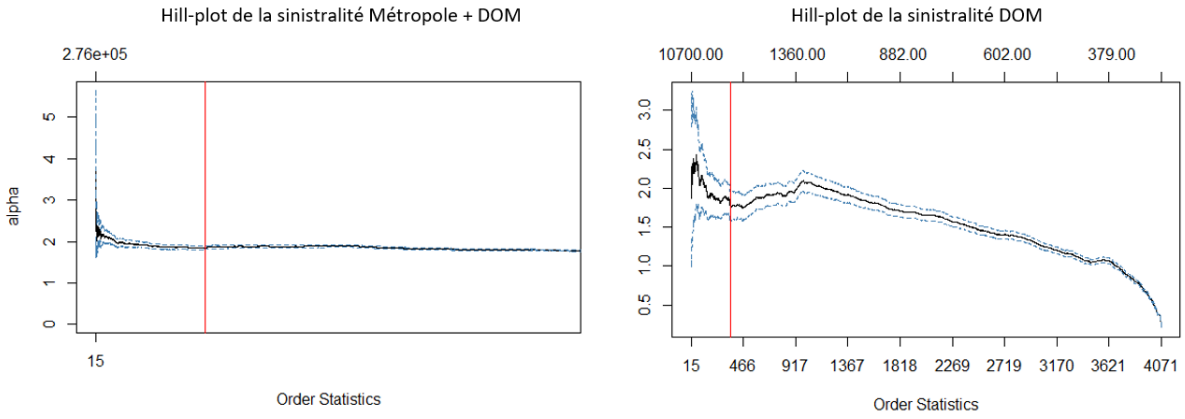


FIGURE 2.2.5 – Hill plot pour la garantie DDE dans les DOM et métropole+DOM

Le Hill plot du segment DOM permet de voir une zone de stabilisation du coefficient α à la 350^{eme} statistique d'ordre ce qui correspond environ à un seuil de 2 100€. En ce qui concerne le segment métropole + DOM, α se stabilise aux alentours de la 7000^{eme} statistique d'ordre ce qui correspond environ à un seuil de 13 000€. On remarque ainsi que les deux méthodes de seuils ne permettent pas d'obtenir les mêmes seuils, d'autant plus qu'il s'agit de deux méthodes visuelles sujettes à erreur d'interprétation. Nos résultats précédents concernant la concordance de la fonction des excès avec une loi de GPD nous font préferer la méthode du mean excess plot dans le choix du seuil. Cependant, ces techniques basées sur la théorie des valeurs extrêmes ne peuvent pas être l'unique facteur de décision pour notre seuil de graves. On doit en effet comparer les seuils obtenus avec notre sinistralité réelle. Pour cela on construit un tableau de données nous permettant de savoir quelle proportion du nombre de sinistres et quelle proportion de la somme des charges a t - on à partir d'un certain seuil. Ces tableaux nous permettant de rattacher le choix du seuil à la sinistralité ont la forme suivante :

Seuil	Prop. n	Prop. charge
1000	50,21 %	81,97%
1100	45,47 %	78,78%
1200	41,39 %	75,78%
1300	37,25 %	72,46%
1400	32,22 %	68,10%
....	...	...

TABLE 2.2.2 – Proportion de sinistres et de la charge totale au delà du seuil pour la métropole+ DOM :

Ainsi, pour les sinistres du segment métropole + DOM, 0,69 % des sinistres sont au dessus du seuil obtenu par la méthode de Hill de 13 000€ et représentent 12,10% de la charge, alors que 0,69 % des sinistres sont au dessus du seuil de 11 500€ obtenu par la méthode du mean excess et représentent environ 13,5 % de la charge totale. Les deux seuils sont donc assez similaire en terme de proportion de sinistres et de charge totale au delà du seuil. On choisira cependant le seuil de 11 500 € pour les raisons liées à la loi GPD d'une part et d'autre part dans le but de conserver un maximum de donner pour la modélisation des sinistres graves : on choisit le seuil le moins élevé des deux.

Pour le segment DOM, 9,06 % des sinistres sont au dessus du seuil obtenu par la méthode de Hill de 2 100€ et représentent 35,86% de la charge, alors que seulement 2,48 % des sinistres sont au dessus du seuil de 11 500€ obtenu par la méthode du mean excess et représentent environ 17,63 % de la charge totale. On privilégie donc ce seuil de 4 500 € qui représente moins de 5 % tout en conservant une importante proportion de la charge totale, ce qui permet ainsi de respecter les critères de faible fréquence et de forte intensité des sinistres graves.

Comme nous l'avons vu précedemment, nous obtenons deux seuils de sinistres graves différents entre le segment métropole + DOM et le segment DOM. Or, dans le cadre de notre modélisation, il est essentiel d'utiliser le même seuil afin de pouvoir comparer les sinistralités attritionelles. On choisit ainsi le seuil de grave du segment métropole + DOM de 11 500 € et on écrêtera les sinistres de nos deux segments DOM et métropole + DOM de ce montant. Le choix de ce seuil s'explique d'une part car la métropole a un poids beaucoup plus important dans le portefeuille (98 %) et d'autre part car une des conclusions possibles de l'étude et l'intégration de la tarification des DOM dans celle de la métropole et non l'inverse. Ce choix de seuil va impliquer une méthodologie légèrement différente en ce qui concerne la modélisation de la sinistralité grave pour la garantie dégâts des eaux des contrats du segment DOM (que nous détaillerons à la fin de notre étude). Ce seuil de 11 500 € est en effet particulièrement élevé pour les DOM : seuls 0,29 % des sinistres sont au delà du seuil et représentent 5,30% de la charge totale.

2.3. Vieillissement des sinistres

2.3.1 Problématique

Le coût d'un sinistre pour l'assureur évolue dans le temps. En effet, à la survenance d'un sinistre, on lui attribue au sinistre une charge forfaitaire à l'ouverture du dossier en fonction des caractéristiques du sinistre (notamment la garantie impactée). Le montant du sinistre sera alors par la suite réevalué à la hausse ou à la baisse par un expert. Pour arriver au coût final du sinistre, un délai de plusieurs années peu s'écouler selon les garanties. Or, dans le cadre de notre modélisation, les charges de sinistres n'ont pas nécéssairement atteintes leur coût final, ce qui peut fausser notre étude de sinistralité. On va donc devoir vieillir nos sinistres afin d'obtenir la vision la plus juste possible de la sinistralité finale attendue.

Dans notre cas, nous avons deux segments métropole + DOM et DOM. Bien que dans l'idéal il faudrait un vieillissement distinct entre les sinistres des deux segments, nous avons opté pour un unique vieillissement basé sur les données du segment métropole + DOM. On justifie ce choix par le manque de données du segment DOM qui ne permet pas réaliser une analyse assez robuste et on prend donc l'hypothèse que l'évolution du coût d'un sinistre ne dépend pas de sa localisation (DOM ou métropole).

Dans la section précédent, nous avons évoqué la distinction entre sinistralité attritionnelle et grave. On aura donc deux types de vieillissement selon la gravité de nos sinistres. on considère en effet que les sinistres graves correspondant à un changement dans la queue de distribution des charges, rien n'indique qu'ils aient une évolution de leurs montants équivalents à celle des sinistres attritionnels.

2.3.2 Méthodologie et résultats

Afin d'établir le coût à l'ultime (c'est à dire le coût final) de sinistres, il existe des estimations s'appuyant sur des méthodes déterministes et stochastiques. Ces méthodes sont basées sur un raisonnement par années de survenance. On notera donc dans la suite :

— i l'année de survenance, c'est à dire l'année où le sinistre est survenu ;

— j l'année de développement, soit le nombre d'années après l'année de survenance écoulée.

— $X_{i,j}$ le coût supplémentaire de l'année $(i + j)$ des sinistres survenus l'année i.

— $C_{i,j}$ le coût cumulé des sinistres survenus l'année i après un développement de j années.

On a ainsi :

$$\mathbf{C_{i,j}} = \sum_{\mathbf{h=0}}^{\mathbf{j}} \mathbf{X_{i,h}}$$

et :

$$\mathbf{X_{i,j}} = \mathbf{C_{i,j}} - \mathbf{C_{i,j-1}}$$

Ces charges cumulées de sinistres sont souvent présentées en forme de trangle de la façon suivante :

	Devpt. 0	Devpt. 1	...	Devpt. j	...	Devpt. n-1	Devpt. n
Surv. 1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,j}$	...	$C_{1,n-1}$	$C_{1,n}$
Surv. 2	$C_{2,0}$	$C_{2,1}$	...	$C_{2,j}$	...	$C_{2,n-1}$	
...	...	...	...	...	...		
Surv. i	$C_{i,0}$	$C_{i,1}$	...	$C_{i,j}$			
...	...	...	...				
Surv. i	$C_{n-1,0}$	$C_{n-1,1}$					
Surv. i	$C_{n,0}$						

TABLE 2.3.1 – Triangle de charges cumulées

Afin de vieillir nos sinistres nous allons utilisé une méthode déterministe célèbre : la méthode de Chain-Ladder. Elle va nous permettre d'obtenir les valeurs manquantes de notre triange de charges cumulées. Notons $\frac{C_{i,j+1}}{C_{i,j}}$ les rapports appelés facteurs de développement nous permettant d'obtenir ces valeurs manquantes. La méthode de Chain-Ladder repose sur l'hypothèse suivante : Les facteurs de développement sont supposés indépendants de l'année d'origine i, c'est-à-dire pour $j = 0, ..., n-1$ fixé on a pour tout i :

$$\frac{\mathbf{C_{i,j+1}}}{\mathbf{C_{i,j}}} = \mathbf{f_j}$$

Pour vérifier cette hypothèse on s'assure que :

$$\frac{\mathbf{C_{1,j+1}}}{\mathbf{C_{1,j}}} \approx ... \approx \frac{\mathbf{C_{i,j+1}}}{\mathbf{C_{i,j}}} \approx ... \approx \frac{\mathbf{C_{n,j+1}}}{\mathbf{C_{n,j}}} = \mathbf{cste}$$

Enfin, pour estimer les facteurs de développement on calcule pour tout j :

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

On peut ensuite déduire la charge cumulée finale pour tout $j = 0, ..., n$:

$$\hat{\mathbf{C}}_{\mathbf{i,j}} = \hat{\mathbf{f}}_{\mathbf{j-1}}.\hat{\mathbf{f}}_{\mathbf{j-2}}...\hat{\mathbf{f}}_{\mathbf{j-i}}.\mathbf{C_{i,j-1}}$$

On va réaliser le vieillissement de nos sinistres en utilisant un historique de 7 années : de 2014 à 2020, afin d'estimer les charges finales au mieux.

On construit donc notre base de données contenant les informations intéressantes en utilisant les bases sinistres. Pour une année n, on considère uniquement les sinistre encore ouvert au 31/12/n dans le but d'avoir des données seulement sur les sinistres susceptibles d'évoluer. De plus, pour distinguer les sinistres graves des sinistres attritionnels on va considérer comme étant grave un sinistre au delà du seuil à l'année d'observation la plus récente (ici l'année 2020), quant bien même le sinistre n'avait pas franchi le seuil l'année de survenance.

Grâce à ces données, on obtient le triangle des charges cumulées suivant pour la sinistralité attritionnelle avec les coefficients de développement associés :

Année Surv. / Dev.	N	N+1	N+2	N+3	N+4	N+5	N+6
2014	76 991 205,71 €	68 985 678,22 €	66 184 765,87 €	64 067 987,54 €	62 839 457,21 €	62 402 359,43 €	62 238 145,12 €
2015	69 232 596,70 €	61 489 216,59 €	60 717 154,95 €	59 302 899,05 €	58 321 407,94 €	57 562 460,33 €	-
2016	72 850 036,94 €	66 861 420,13 €	65 150 535,61 €	63 588 419,54 €	62 767 848,88 €	-	-
2017	76 172 101,04 €	69 602 692,42 €	67 262 339,94 €	65 695 054,19 €	-	-	-
2018	93 832 684,30 €	86 680 115,08 €	83 313 942,27 €	-	-	-	-
2019	81 668 466,83 €	75 214 920,79 €	-	-	-	-	-
2020	78 613 872,88 €	-	-	-	-	-	-
Coefficient de dév.	0,911	0,969	0,974	0,984	0,990	0,997	-
Coefficient de dev. cumul.	0,838	0,917	0,947	0,972	0,988	0,997	-

FIGURE 2.3.1 – Triangle des charges cumulées pour la sinistralité attritionnelle métropole + DOM

On s'aperçoit donc que la première année, la sinistralité attritionnelle est sur-évaluée d'environ 9% ce qui peut indiquer un coût forfaitaire d'ouverture trop élevé. Les coefficients de développement s'approchent ensuite de 1 et convergent vers cette valeur au bout de 5 ans.

On fait de même pour la sinistralité grave :

Année Surv. / Dev.	N	N+1	N+2	N+3	N+4	N+5	N+6
2014	14 692 583,55 €	19 881 982,73 €	21 540 827,81 €	23 140 283,14 €	24 086 908,34 €	24 580 918,66 €	24 602 497,01 €
2015	5 638 553,27 €	10 827 311,66 €	11 990 054,77 €	12 005 135,10 €	12 319 293,49 €	12 749 148,91 €	-
2016	6 192 835,38 €	12 755 614,25 €	13 668 233,01 €	14 257 071,76 €	14 533 642,87 €	-	-
2017	9 276 968,37 €	15 546 775,14 €	16 832 256,92 €	17 532 441,87 €	-	-	-
2018	12 401 940,54 €	20 904 567,41 €	22 595 762,57 €	-	-	-	-
2019	8 350 739,18 €	16 996 480,07 €	-	-	-	-	-
2020	10 991 042,54 €	-	-	-	-	-	-
Coefficient de dév.	1,714	1,084	1,045	1,031	1,025	1,001	-
Coefficient de dev. cumul.	2,053	1,199	1,106	1,058	1,026	1,001	-

FIGURE 2.3.2 – Triangle des charges cumulées pour la sinistralité grave métropole + DOM

On remarque que dans le cas de la sinistralité grave, la sinistralité grave est sous-évaluée d'environ 71% la première année. Cependant, de la même manière que la sinistralité attritionnelle, la sinistralité grave se stabilise au bout de 5 années.

Après avoir déterminé ces coefficients de développement, pour chaque année de survenance, on utilise les coefficients de passage (il s'agit du produit des coefficients de développement) afin d'atteindre la charge ultime. Par exemple, pour l'année 2019 (notre dernière année d'observation), il faudra appliquer un coefficient de passage de $0,969$ dans le cas d'un sinistre attritionnel et de $1,199$ dans le cas d'un sinistre grave afin d'atteindre les charges à l'ultime. Ces coefficients de passages seront uniquement appliqués aux sinistres non clos.

On remarquera que nos charges de sinistres n'ont pas été inflatées. En effet, on observe nos sinistres sur un historique relativement important (7 années ici), et les sinistres n'ont pas le même coût d'une année à l'autre du fait de l'inflation. Ainsi, on utilise généralement l'indice FFB (la Fédération Française du Bâtiment) afin de tenir en compte de l'inflation du montant des sinistres. Or cet indice n'explique pas nécéssairement la variation du coût d'un sinistre dégâts des eaux d'années en années. C'est la raison pour laquelle nous allons préferer modéliser la variabilité annuel du coût d'un sinistre (mais également de sa fréquence) à travers d'une variable "Année" nous permettant d'expliquer les évolutions temporelles. Cela permettra de prendre en compte d'autre facteurs que l'inflation pour expliquer l'évolution du montant des sinistres.

2.4. Formatage de la base de modélisation

2.4.1 Construction des bases de données

Dans cette section, nous allons utiliser toutes les informations précedents pour expliquer les étapes permettant la construction des deux bases de données servant à la modélisation. Ces deux bases de données sont la base "fréquence" et la base "coût moyen". Ces bases ont été construites grâce au logiciel SAS.

On débute avec les bases images, bases sinistres et bases clients que l'on construit comme indiqué précedemment pour chaque année de survenance de notre fenêtre d'observation (de 2009 à 2019). Pour chaque année de survenance, on joint les bases image et sinistres par numéros de contrats, qu'on joint à nouveau avec les bases clients par numéro de client : on aura une ligne par numéro de contrat x numéro de sinistre (on peut avoir plusieurs sinistre dans la même année). A chaque jointure on ne conserve que les variables jugées nécéssaires à notre modélisation. On obtient donc pour chaque année de survenance une base de données avec les informations concernant le risque, les sinistres et les clients par numéro de contrat que l'on va nommé bases jointes et qui seront au nombre de 11. En notant i l'année, n_i le nombre de contrat dans l'année et m_i le nombre de sinistre dans l'année, on obtient :

N° de contrat	Nb pièces	...	N° sinistre	Charge sin.	...	Age client	...
$c_{i,1}$	4	...	$s_{i,1}$	400 €	...	30	...
$c_{i,1}$	4	...	$s_{i,2}$	700 €	...	30	...
$c_{i,2}$	2	...	-	0 €	...	42	...
...	...	...	...	...	...	...	...
c_{i,n_i-1}	1	...	-	0 €	....	21	...
c_{i,n_i}	8	...	s_{i,m_i}	1000 €	....	67	...

TABLE 2.4.1 – Structure de la base jointe en année i

On crée ensuite les bases antécédents par années de survenance. Pour chaque année de survenance, on crée une base de données contenant la liste des numéros de contrats de l'année correspondante associée au nombre de sinistres et à la somme des charges que chacun a subi sur une fenêtre de 1, 3, 5 et 10 ans. Par exemple pour l'année 2012 les antécédent à 1 an reprendront les données de l'année 2011, ceux à 3 ans celles de 2009 à 2011 etc... On joint ensuite ces informations à notre base jointe.

N° de contrat	...	Charge sin.	...	Nb ant 1 an	...	Nb ant 10 ans	Chg 1 an	...	Chg 10 ans
$c_{i,1}$	...	400 €	...	0	...	0	0 €	...	0 €
$c_{i,1}$	...	700 €	...	0	...	0	0 €	...	0 €
$c_{i,2}$	...	0 €	...	1	...	1	200€	...	200 €
...	...	...	...	...	...	...	...	...	...
c_{i,n_i-1}	...	0 €	...	0	...	0	0 €	...	0 €
c_{i,n_i}	...	1000 €	...	1	...	4	600 €	...	1 300 €

TABLE 2.4.2 – Structure de la base jointe avec antécédent en année i

Construction de la base fréquence :

Pour construire la base fréquence, on crée tout d'abord la variable "nombre de sinistre" en comptant les numéros de sinistres (0 s'il n'y a pas de numéro de sinistre) par numéro de contrats.

Il reste une variable clé à construire : l'exposition du contrat, il s'agit de la fraction correspondant à la durée pendant laquelle un contrat est dans le portefeuille au cours de l'exercice. Par exemple, si un contrat est dans le portefeuille uniquement 3 mois dans l'année, son exposition sera de 0,25, en revanche, s'il est présent tout au long de l'année, son exposition sera de 1. On contruit ainsi cette variable à l'aide des dates d'affaire nouvelle et de résiliation pour chaque année de survenance.

Enfin, pour obtenir la base de fréquence, on concatène nos 11 bases de données, qui aura une ligne par numéro de contrat x année de survenance :

N° de contrat	...	Nb sinistre	Exposition	Année
$c_{2009,1}$	...	0	1	2009
$c_{2009,2}$	...	0	0,90	2009
...	...	...	...	...
$c_{i,1}$	...	2	1	i
$c_{i,2}$	...	0	1	i
...	...	...	...	...
c_{i,n_i}	...	1	1	i
...	...	...	...	...
$c_{2019,n_{2019}}$	...	1	1	2019

TABLE 2.4.3 – Structure de la base fréquence

Construction de la base coût moyen :

Pour la base de coût moyen, on va vieillir la charge des sinistres non clos de nos bases jointes. Pour cela on distingue les sinistres au delà de notre seuil de grave fixé à 11 500€ précédemment. On les vieilli avec les coefficients de passages calculés au préalable. Avec cette nouvelle charge vieillie, on crée une nouvelle variable "charge vieillie attritionnelle" qui prendra pour valeur la charge écrêtée du seuil et la variable "charge vieillie grave" qui prendra pour valeur le montant dépassant le seuil (ces formules étant détaillées en section 2.2). On restreint ensuite notre base de données aux contrats ayant subi un sinistre en filtrant les données avec une charge vieillie attritionnelle strictement positive. Enfin, de la même manière que pour la base "fréquence", on concatène les 11 bases obtenues pour avoir la base "coût moyen" qui aura une ligne par numéro de sinistre x année de survenance.

N° de sinistre	...	Charge attritionnelle	Charge grave	Année
$s_{2009,1}$	...	800 €	0 €	2009
$s_{2009,2}$	...	11 500 €	500 €	2009
...	...	...	...	...
$s_{i,1}$	...	400 €	0 €	i
$s_{i,2}$	...	700 €	0 €	i
...	...	...	...	...
s_{i,m_i}	...	1 000 €	0 €	i
...	...	...	...	...
$s_{2019,m_{2019}}$	...	3000 €	0 €	2019

TABLE 2.4.4 – Structure de la base coût moyen

On a ainsi obtenu nos bases fréquence et coût moyen du segment métropole + DOM, les bases image et sinistre regroupant les informations de la France métropolitaine et des territoires ultramarins à la fois. Pour obtenir les bases correspondant aux données du segment DOM il nous suffit alors de filter ces contrats.

2.4.2 Traitement des variables et corrélations

On a donc crée nos 4 bases de données servant à la modélisation. Cependant, le format de ces bases de données est encore relativement "brut" : il nous faut retraiter nos variables afin d'extraire la meilleure information possible de celles-ci lors de la modélisation :

— **Eviction des variables sensibles :**

On commence tout d'abord par retirer les variables permettant de retracer l'identité d'un assuré (numéro de contrat, numéro de sinistre et numéro de client) dans le cadre de la réglementation RGPD et d'utilisation du logiciel externe *Akur8*.

— **Gestion des valeurs aberrantes et des valeurs manquantes :**

On va considérer comme aberrantes les valeurs qu'une variable ne peut en aucun cas prendre. C'est le cas des variables :

- "âge client" quand elle prend des valeurs strictement inférieure à 18 et strictement supérieur à 100 où on remplace ces valeurs par "NA" indiquant que la valeur est inconnue.

- "ancienneté client" et "ancienneté contrat" quand elles prennent des valeurs strictement inférieure à 0 et strictement supérieur à 60 où les valeurs sont également remplacées par "NA".

- "nombre de pièce" quand elle prend des valeur égales à 0 (il n'y a pas d'habitation de 0 pièce) ou strictement supérieur à 25 (seuls les personnes ayant une habitation d'au plus 25 pièces peuvent souscrire un contrat MRH chez AXA). Si la valeur prend 0 on la remplacera par 1, si elle est supérieure à 25, on la remplacera par 25.

En ce qui concerne les valeurs manquantes, il y a deux possibilités :

- soit il s'agit d'une variable a une modalité vide qui est réellement inconnue , dans ce cas on la remplace par "NA". C'est le cas de la variable "étage".

- soit il s'agit de variables dichotomique où la valeur 0 est vide (les extractions ont besoin uniquement de la valeur 1, le reste est vide). Dans ce cas on la remplace par 0. C'est le cas des variables dichotomiques "étudiant" et "personne morale" et des variables d'antécédents.

— **Traitements sur des variables spécifiques :**

Ces traitements spécifiques concernent 5 variables :

- Avec les variables "capitaux mobilier" et "taux objet de valeur", on peut construire deux variables nous permettant d'apporter une meilleure information : on va scinder les capitaux mobiliers en deux, d'une part les capitaux mobiliers communs et d'autre part les capitaux mobiliers de valeurs.
On aura ainsi : $capitaux\ mobiliers\ de\ valeur = capitaux\ mobilier \times taux\ objet\ de\ valeur$
et : $capitaux\ mobiliers\ communs = capitaux\ mobiliers \times (1 - taux\ objet\ de\ valeur)$

- Au sein de notre base de modélisation on a les variables "type d'habitation" (appartement ou maison) et "qualité de l'occupant" (propriétaire ou locataire). On a décidé de croiser ces deux variables en une seule, d'une part afin de pouvoir modéliser le comportement de profil plus spécifique et d'autre part car c'est ainsi que nous présentons une multitude de reportings. On obtient ainsi la variable "Segment" qui aura quatre modalités : propriétaire de maison, propriétaire d'appartement, locaitaire de maison et locataire d'appartement.

- Enfin, comme détaillé en section 2.1.3, nous allons recoder la variable "Code INSEE" des contrats situés dans les DOM afin de correspondre à la réalité administrative actuelle.

— **Discrétisation et regroupement de modalités :**

La discrétisation des variables continues telles que la surface des dépendances ou encore les capitaux mobiliers, de même que le regroupement de modalité dans le cas de la variable nombre de pièce (qui est discrète) ont dux principaux avantages dans le cadre de notre modélisation : d'une part, on pourra obtenir un modèle plus interprétable et explicable (le tarif doit être facilement compréhensible pour les assurés notamment), d'autre part le pouvoir prédictif du modèle peut être amélioré en prenant en compte les effets de non-linéarité par exemple. Dans le cadre de notre étude, on réalisera une discrétisation séparée entre les segments 'DOM' et 'métropole+DOM'.

Il existe deux types de discrétisation que nous allons détailler.

On a tout d'abord la discrétisation non supervisée, qui ne prend pas compte de la variable à expliquer (nombre de sinistre et charge dans notre cas). Trois méthodes de discrétisation supervisée peuvent être appliquées :

- La discrétisation par intervalle : on sépare nos données en N intervalles de même largeur. Ainsi chaque intervalle a une largeur de : $L = \frac{valeur\ max - valeur\ min}{N}$

- La discrétisation par fréquence (appelée également quantilisation) : on sépare nos données en n intervalles contenant le même de données.

- La discrétisation par cluster qui va utiliser l'agorithme des k-means afin d'assigner nos valeur à des groupes. L'algorithme permet de s'assurer que la distance intra-roupe est minimisée (les individus d'un même groupe sont "proches"), et de maximiser la distance inter-groupe (les individus d'un groupe différent sont "éloignés").

Une notion importante dans la discrétisation est le nombre de groupe crée. Il existe de multiples méthodes permettant d'obtenir le nombre idéal de groupes (*Hunault, 2022,* **[2]**). On peut citer notamment la règle de Freedman-Diaconis où le nombre n d'intervalle pour une variable X de N données est :

$$\mathbf{N} = \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{max}(\mathbf{X}) - \mathbf{min}(\mathbf{X})) \cdot \frac{\sqrt[3]{n}}{\mathbf{Q_3}(\mathbf{X}) - \mathbf{Q_1}(\mathbf{X})}$$

ou encore la règle de Sturge :

$$\mathbf{N} = \mathbf{1} + \log_{\mathbf{2}}(\mathbf{n})$$

Cependant, ces méthodes ne seront pas utilisées pour deux raisons : d'une part, le grand nombre de données que l'on a et notamment pour le segment "métropole + DOM" implique un nombre important d'intervalle avec ces formules, et d'autre part, la principale problématique d'un point de vue commercial est la compréhension du tarif par les clients. De ce fait, le nombre d'intervalle à réaliser a été choisi de façon opérationnelle.

De la même manière, le choix de la méthode de discrétisation utilisée parmis celles citées précedemment va dépendre des intervalles formés par ces méthodes. On privilégie la méthode offrant des intervalles plus facilement compréhensible pour les assurés. Par exemple, pour la variable "surface des dépendances" qui a pour valeur 0 dans plus de 95% des cas, une quantilisation ne donnerait que deux modalités (surface égale à 0 ou surface strictement supérieure à 0) là où la méthode des k-means permet d'obtenir des intervalles plus cohérent d'un point de vue commercial.

On peut donc résumer (pour quelques variables du segment "métropole + DOM") l'étape de discrétisation avec le tableau suivant :

Variable	Méthode	Nombre groupe	Bornes des intervalles
Age client	Quantile	7	18 - 34 - 43 - 51 - 58 - 65 - 76 - 100
Anc. contrat	Quantile	8	0 - 1 - 2 - 3 - 5 - 7 - 11 - 20 - 60
...	...	...	...
Surface dép.	K-means	7	0 - 1 - 50 - 150 - 300 - 500 - 750 - 1200 - 3200

TABLE 2.4.5 – Discrétisation des variables

Enfin, on a la discrétisation supervisée qui va permettre de réaliser des groupes en fonction de la variable à expliquer. Avec ce type de discrétisation on a donc d'une part une discrétisation pour la fréquence et d'autre part une discrétisation pour le coût moyen.

Il existe plusieurs algorithmes basées principalement sur la statistique du χ^2 (*Hyunji, 2015,* ***[3]***). On peut notamment citer l'AMEVA qui est un algorithme maximisant une mesure (appelée AMEVA également) et qui génère potentiellement le nombre minimal d'intervalle discret, et en notant n le nombre d'intervalle et k le nombre de classes, on a :

$$\mathbf{AMEVA(n)} = \frac{\chi^{\mathbf{2}}\mathbf{(n)}}{\mathbf{n} * (\mathbf{k} - \mathbf{1})}$$

On n'utilisera cependant pas ces algorithmes supervisés dans notre discrétisation et on se cantonnera aux méthodes non-supervisées pour plusieurs raisons :

- Les intervalles trouvés par ces méthodes ne sont souvent pas adaptés d'un point de vue commercial.

- On aurait des modalités différentes en fonction de la variable à expliquer ce qui va avoir un impact sur la compréhensibilité du tarif final.

- Il y a une forme de sur-apprentissage à définir nos intervalles en fonction de notre variable à expliquer.

— **Analyse des corrélations entre variables :**

La dernière étape avant de réaliser la modélisation est l'étude des corrélations de notre base de données. On réalisera cette étude à partir de la base fréquence qui contient l'intégralité données, et il est également nécéssaire de distinguer les segments "DOM" et "métropole + DOM". En effet, on souhaitera conserver uniquement les variables qui n'ont pas de fortes corrélation avec une autre : la présence d'une corrélation élevée peut impacter l'interprétabilité des modèles.

Maintenant que l'on a uniquement des variables catégorielles (après discrétisation des variables conti-

nues), on va pouvoir étudier ces corrélations à l'aide du calcul du V de Cramer qui va permettre de mesurer la corrélation entre deux variable catégorielles grâce à la statistique du χ^2. Les statistiques sont définies par les formules suivantes :

$$\mathbf{V_{Cramer}} = \sqrt{\frac{\chi^2/n}{\min(k,l)-1}} \quad \text{et} \quad \chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{i,j} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

avec, en notant X et Y nos deux variables analysées :

- k et l le nombre de modalités des variables X et Y.

- $n_{i,j}$ le nombre d'observations ayant la modalité i pour X et la modalité j pour Y.

- $n_{i.}$ et $n_{.j}$ le nombre d'observations ayant la modalité i pour X et j pour Y respectivement.

- n le nombre d'observation total.

Le V de Cramer peut prendre des valeurs comprises entre 0 (aucune corrélation) et 1 (totale corrélation). On fixe un seuil de 0,85 afin de définir si deux variables sont excessivement corrélées ou non. Voici les matrices de corrélations que l'on obtient pour nos deux segments :

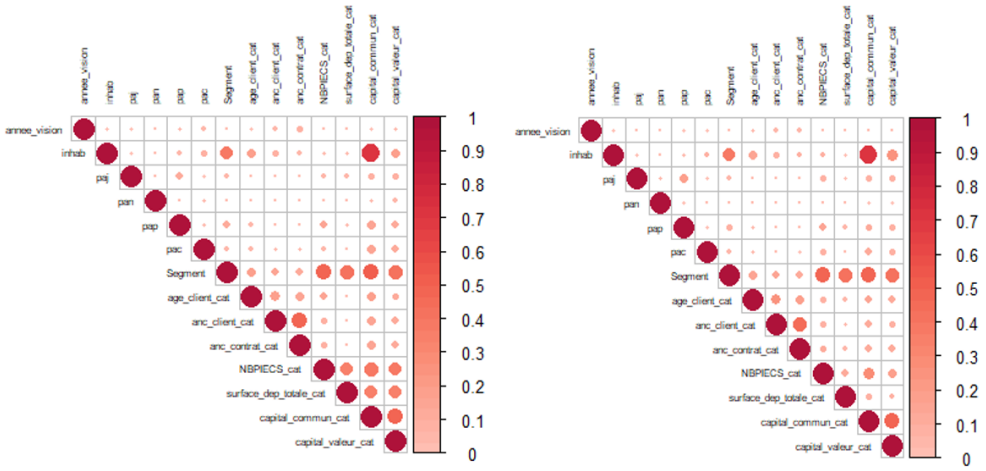


FIGURE 2.4.1 – Corrélogrammes des variables DOM (à gauche) et métropole + DOM (à droite)

On peut remarquer que des corrélogrammes des segment "métropole + DOM" et DOM sont similaires. Dans les deux cas aucune variable n'est à retirer : il n'y a pas de couple de variable ayant une corrélation supérieure à 0,85. On peut néanmoins citées les couples les plus corrélées :

- Le capital mobilier commun et le type d'occupation (logement occupé ou non) avec une corrélation de 0,70 sur le segment "DOM" et "métropole + DOM".

- Le capital mobilier commun et le segment (propriétaire / locataire d'appartement / maison) avec une corrélation de 0,50 sur le segment "DOM" et "métropole + DOM".

- Le nombre de pièces et le segment (propriétaire / locataire d'appartement / maison) avec une corrélation de 0,49 sur le segment "DOM" et de 0,50 sur le segment "métropole + DOM".

2.4.3 Analyse descriptive

Maintenant que nos 4 bases de données sont au format désiré et nettoyées, on peut réaliser une analyse descriptive de leur contenu.

— **Dimension des bases de données :**

Voici en premier lieu un tableau décrivant les dimensions de nos bases de données :

Base de donnée	N. observation	N. variable
Base fréquence métropole + DOM	39 955 372	33
Base fréquence DOM	376 714	33
Base coût moyen métropole + DOM	909 630	32
Base coût moyen DOM	3 967	32

TABLE 2.4.6 – Dimension des bases de données après formatage

On remarque ainsi que la base fréquence du segment "DOM" représente uniquement 2,27% de la base fréquence du segment "métropole+DOM". Là où la base coût moyen du segment "DOM" représente seulement 0,44 % de la base coût moyen du segment "métropole + DOM".

Les bases fréquence et coût moyen ont quasiment le même nombre variable explicative, les bases fréquence ont uniquement la variable "exposition" en plus (qui sert uniquement à la modélisation de la fréquence).

— **Localisation du risque dans les DOM :**

Comme vu précedemment, la variable "Code INSEE" a été reformatée dans les DOM afin de correspondre au découpage administratif en vigueur. Il est ainsi possible de réaliser une carte afin de mieux comprendre la localisation précise des risques dans les DOM. Il a été décidé de représenter cette carte avec les contrats actuellement en portefeuille (à mars 2022) et non des données issue de notre base de modélisation, mais les proportions auraient été les mêmes de 2009 à 2019 :

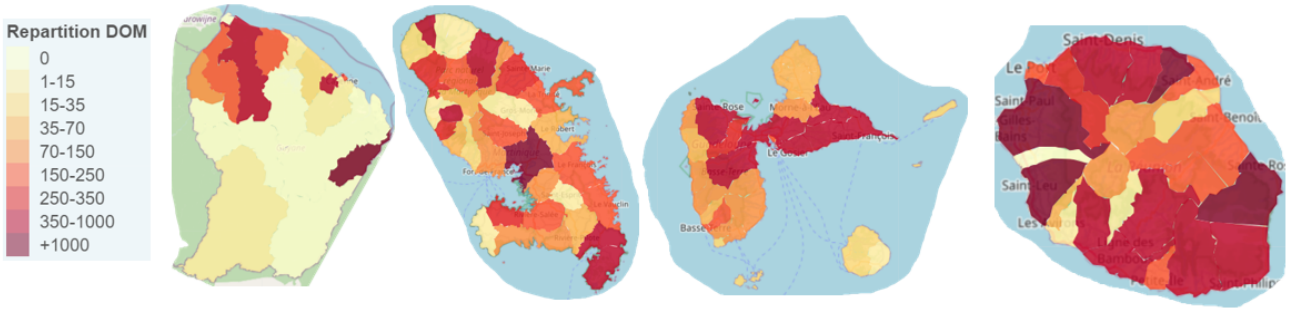


FIGURE 2.4.2 – Localisation du risque dans les DOM

Ainsi, La Réunion concentre la majorité du risque sur le littoral ouest et sud-est (zones les plus peuplées) et concentre plus de risques de manière générale par rapport aux autres DOM . La Guadeloupe concentre son risque au sud de Grande-Terre et au nord de Basse-Terre. En Martinique la localisation du risque est plus hétérogène, avec des zones à risque à la fois au nord, au sud et dans le centre. Enfin, la Guyane a son risque concentré au nord-ouest avec également un nombre important d'assurés à Saint-Georges à l'est et à Montsinéry-Tonnegrande au nord.

— **Statistiques descriptives :**

A présent, il peut être intéressant d'avoir un premier aperçu du comportement des variables à expliquer en fonction des variables explicatives.

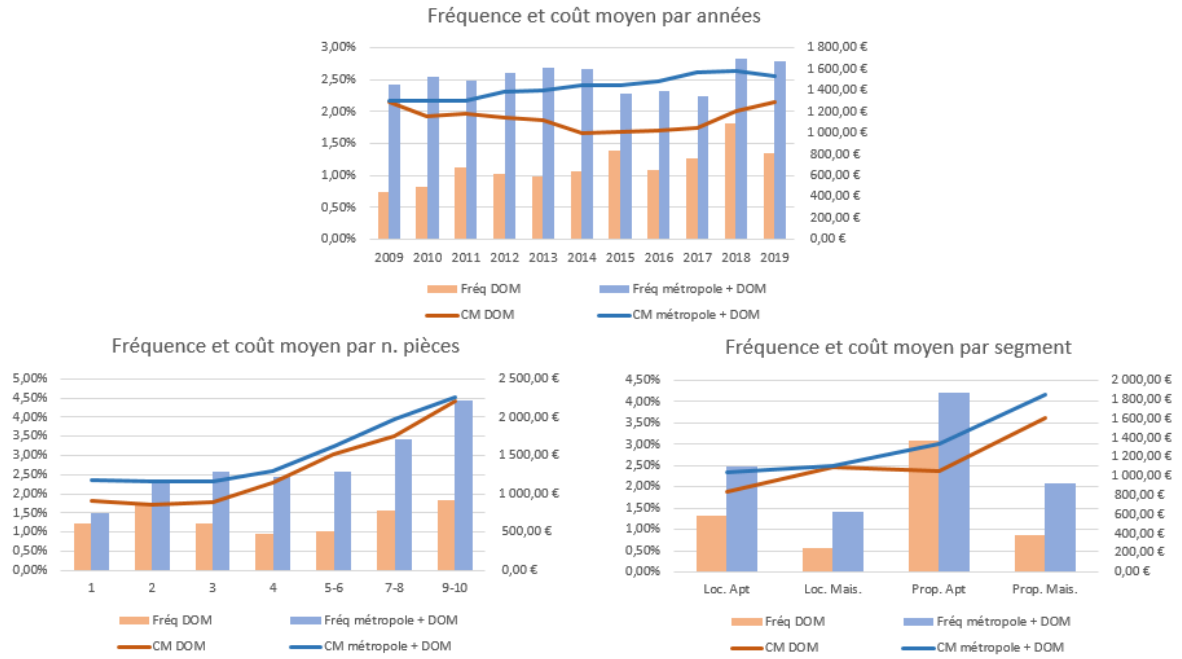


FIGURE 2.4.3 – Evolution de la fréquence et du coût moyen sur les périmètres DOM et métropole + DOM

Le premier graphique (en haut) nous montre les écarts de fréquences et de coût moyen entre le périmètre DOM et le périmètre métropole + DOM : En moyenne la fréquence de sinistre annuelle est plus de 50% plus importante sur le segment DOM que sur le segment métropole+DOM, là où le coût moyen est 20 % plus important sur le segment métropole + DOM (il s'est accentué ces 5 dernières années).

Le second graphique (à gauche) nous montre les écarts sur la variable "nombre de pièces" (uniquement sur les modalités en communs). Les écarts les plus important en fréquence sont sur les habitations de 4 et 5-6 pièces (61 % et 60% plus important sur le périmètre métropole + DOM) là où il est plus faible sur les habitation de 1 et 2 pièces (fréquence 18% et 28% supérieure "seulement"). Concernant le coût moyen, les écarts les plus importants sont sur les appartements à faible nombre de pièces (24 % plus important sur le périmètre métropole+DOM pour les 1-2-3 pièces) et les plus faibles sont sur les habitations de 9-10 pièces ("seulement" 2% plus important dans ce cas).

Enfin, le dernier graphique (à droite) présente les écarts selon le segment (locataire / propriétaire d'appartement / maison). Ici, l'écart le plus important en fréquence est sur les locataires de maison (fréquence en moyenne 60% supérieure sur le périmètre métropole + DOM) et le plus faible sur les propriétaire d'appartement (fréquence en moyenne 28% supérieure). Concernant le coût moyen l'écart le plus important se fait sur les locataires et propriétaires d'appartements (20% et 22% supérieur sur le périmètre métropole + DOM) là où l'écart le plus faible est pour le segment locataire de maison (1%

d'écart en moyenne).

Chapitre 3

Modélisation de la sinistralité hors zonier

3.1. Préliminaires

3.1.1 Méthodologie de modélisation

Dans cette section, nous allons détailler notre méthodologie de modélisation, notre objectif est de modéliser la sinistralité des contrats situés dans les DOM.

— **L'approche "coût-fréquence" :**

Comme préssenti avec la construction des bases "fréquence" d'une part et "coût moyen" d'autre part, la modélisation se basera sur une approche "coût - fréquence". Cette démarche se caractérise par le formalisme mathématique suivant :

Dans un modèle collectif, avec une classe de contrats C (les contrats en portefeuille), on s'intérèsse à S la charge sinistre de C. S est fonction du nombre de sinistres N et des montants de ceux ci $(X_i)_{i=1,...,k}$:

$$\mathbf{S} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i}$$

et si les X_i sont i.i.d et indépendant de N, on peut montrer que (*Charpentier, 2018,* **[4]**) :

$$\mathbb{E}(\mathbf{S}) = \mathbb{E}(\mathbf{N}) \ . \ \mathbb{E}(\mathbf{X_1})$$

avec :

- $\mathbb{E}(N)$ la fréquence

- $\mathbb{E}(X_1)$ le coût moyen

Ainsi, en modélisant la fréquence d'une part et le coût moyen d'autre, il sera possible de calculer la prime pure en faisant le produit des deux.

— **La décomposition des bases de données :**

Les bases de modélisation (fréquence et coût moyen) vont être séparées en trois :

- Les bases HZ (pour hors zonier) sur lesquelles on réalisera une modélisation de la sinistralité ne tenant pas compte des informations géographiques (le lieu du risque). Ces bases pèseront 40% des bases totales.

- Les bases Z (pour zonier) sur lesquelles on réalisera une modélisation de la sinistralité en tenant compte des variables géographiques (et des variables antécédents). Les bases Z représenteront également 40% de nos données.

- Les bases V (pour validation) qui vont servir à valider les modèles provenant des bases Z. Ici, il n'y aura donc pas d'apprentissage sur les bases V. Les bases V représenteront 20% de nos données.

Afin de permettre une séparation optimale es bases de données en HZ, Z et V. Les variables à expliquer vont tout d'abord être quantilisées : la variable "nombre de sinistre" sera par exemple regroupée en 5 modalités pour le périmètre métropole + DOM ("0", "1", "2", "3", et "+4"). Ensuite, on affectera chaque ligne à l'une des bases HZ, Z et V de telle sorte qu'il y ait la même proportion de chaque modalité dans ces bases. Cela va permet d'avoir à peu près les mêmes distributions de la variable à expliquer dans chaque base.

Afin de comparer les modèles issus des périmètres "DOM" et "métropole+DOM" au mieux, on fait en sorte d'avoir dans les bases de modélisation "métropole + DOM" la même proportion de contrats DOM que dans la base de modélisation "DOM". Pour cela, les bases de données "métropole + DOM" vont être séparées en "métropole" d'une part et "DOM" d'autre part, avant d'être séparées en HZ, Z, V chacune, puis reconcaténées pour obtenir la séparation finale pour le périmètre "métropole + DOM". Cette méthodologie est résumée dans la figure suivante :

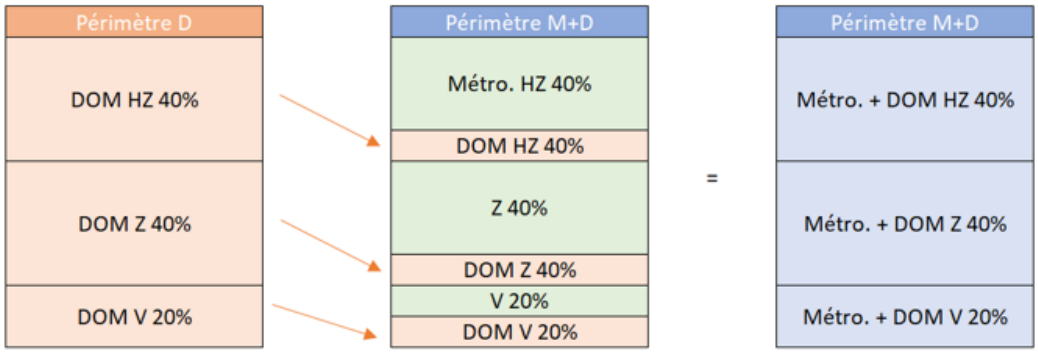


FIGURE 3.1.1 – Séparation des bases en HZ, Z , V

Notre modélisation se portera donc sur un total de 12 bases de données (2 (nombre de périmètre) * 2 (fréquence et coût moyen) * 3 (HZ / Z / V)).

— **Le processus de modélisation :**

On rappelle que l'objectif est d'étudier la faisabilité de l'intégration de la tarification des contrats situés en territoires ultramarins (assimilés au périmètre "DOM") à la tarification des contrats situés en France métropolitaine (assimilé ici au périmètre "métropole+DOM", les DOM représentant moins de 2% du portefeuille). On comparera donc des modèles issus d'une base de données "DOM" et d'une base de données "métropole + DOM" sur une base de données composée de contrats "DOM" afin d'analyser les pouvoirs explicatifs.

Le processus de modélisation se déroulera en deux principales parties (valables aussi bien pour la modélisation de la fréquence que pour celle du coût moyen) :

- La modélisation de la sinsitralité avec les variables hors zonier sur les bases de données "HZ DOM" et "HZ métropole + DOM". Plusieurs modèles seront entraînés sur les bases de modélisations et un seul sera choisi grâce à des indicateurs. On aura donc d'une part un modèle hors zonier basé sur le périmètre "métropole + DOM" et d'autre part, un modèle hors zonier basé sur le périmètre "DOM". Ces modèles seront appliqués sur la base de données "Z DOM" afin de pouvoir comparer les performances hors zonier. L'objet de ce chapitre sera la modélisation hors zonier.

Cette étape est résumée à partir de la figure suivante :

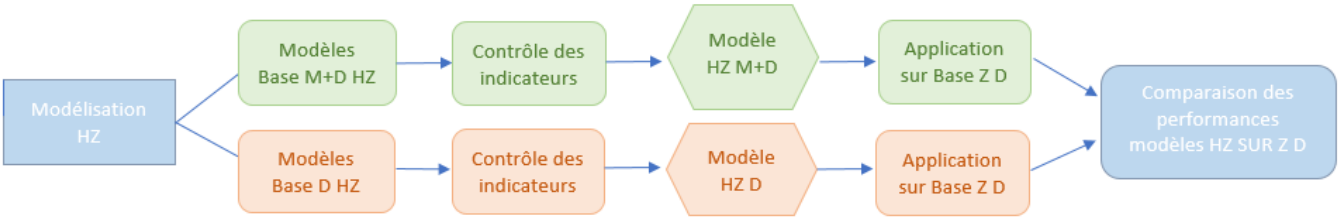


FIGURE 3.1.2 – Schéma de la modélisation hors zoniers

- La modélisation de la sinsitralité avec les variables géographiques sur les bases de données "Z DOM" et "HZ métropole + DOM". L'application des modèles hors zoniers sur les bases Z nous permet d'obtenir des résidus. Ces résidus vont être expliqué par plusieurs modèles intégrant les variables géographiques que l'on comparera avec des indicateurs. Les modèles sélectionnés seront appliqués à ces mêmes bases Z afin, cette fois, d'expliquer les résidus par les variables d'antécédents. Les modèles obtenus suite à cela seront appliqués sur la base "V DOM" dans le but de comparer leur performance sur une base sur laquelle les modèles ne seront pas entraîné.

Cette étape est résumée à partir de la figure suivante :

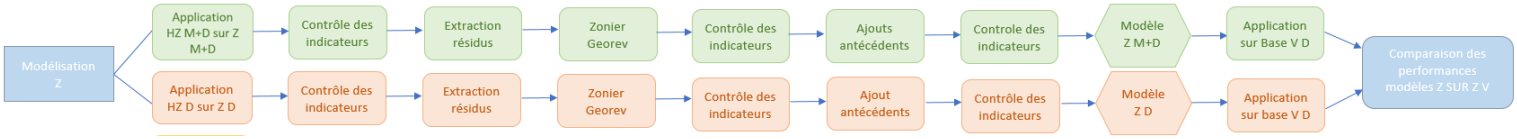


FIGURE 3.1.3 – Schéma de la modélisation avec zoniers

Ainsi, quatres modèles auront été selectionnés pour chaque variable à expliquer : le modèle hors zonier basée sur le périmètre "DOM" appelé modèle "HZ DOM", le modèle hors zonier basée sur le périmètre "métropole+DOM" appelé modèle "HZ métropole+DOM", le modèle avec zonier basée sur le périmètre "DOM" appelé modèle "Z DOM" et le modèle avec zonier basée sur le périmètre "métropole+DOM" appelé modèle "Z métropole+DOM"

3.1.2 L'évaluation des modèles

Dans cette section seront détaillés les outils ayant servi à l'évaluation de nos modèles.

— **La validation croisée :**

La validation croisée par k-folds est une méthode de validation appliquée à nos bases de données d'apprentissage. Elle sert dans notre cas à évaluer les performances de nos modèles mais peut également être utilisée dans le cadre du réglage d'hyperparamètres. Voici les étapes à réaliser :

- On commence généralement par scinder notre base de données (ici les bases HZ et Z) en k groupes (appelés folds) de manière aléatoire, et dans notre cas, chaque fold aura la même proportion de modalités de la variable à expliquer (même principe que la séparation en HZ, Z et V). Un nombre de 4 folds sera utilisé pour notre modélisation.

- $k-1$ folds sont ensuite utilisés afin d'entraîner un modèle et le dernier fold est utilisé pour l'évaluation des performances avec des indicateurs.

- Le procédé est répété k fois afin d'obtenir k estimations des indicateurs.

- On calcule ensuite la moyenne des indicateurs afin d'évaluer les performances globales du modèle.

Voici un schéma résumant la méthode des k-folds :

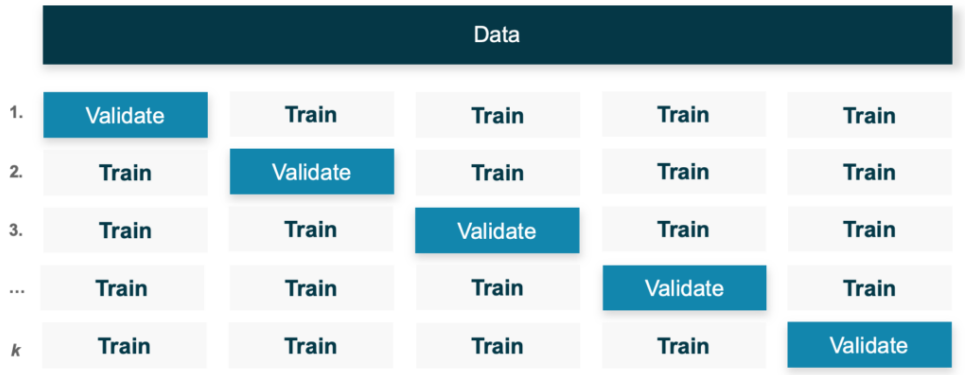


FIGURE 3.1.4 – Schéma de la validation croisée avec k-folds, source : R-Bloggers

Ainsi, on comparera les performances sur les k-folds de multiples modèles. Le modèle final que l'on va appliquer sera lui construit sur toute la base d'entraînement.

— **Les indicateurs de performances :**

La clé pour évaluer nos modèles sont les indicateurs de performances aussi appelés métriques. 6 métriques seront utilisées pour évaluer nos modèles. En notant n le nombre d'observations, y_i les valeurs réelles et $\hat{y}_i$ les valeurs prédites, elles sont définies de la manière suivante :

- La **RMSE** (signifiant Root Mean Squared Error) qui est la racine de l'erreur quadratique moyenne :

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

- La **MAE** (signifiant Mean Absolute Error) qui est l'erreur absolue moyenne :

$$\mathbf{MAE} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} |\mathbf{y_i} - \hat{\mathbf{y}_\mathbf{i}}|$$

- L'erreur globale permettant d'apprécier la qualité globale du modèle :

$$\mathbf{Erreur\ globale} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}}$$

Il y a ensuite deux mesures spécifiques au type de modèle que l'on va utiliser : la **déviance** qui permet d'évaluer la qualité de la régression (les détails de cette métrique seront expliqués dans la prochaine section). Deux types de déviance seront utilisées ici :

- La déviance de Poisson, utilisée pour le modèle de fréquence :

$$\mathbf{D_{Poisson}} = -\mathbf{2} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}.\mathbf{ln}(\frac{\mathbf{y_i}}{\hat{\mathbf{y}}_\mathbf{i}}) - (\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})$$

- La déviance de Gamma, utilisée pour le modèle de coût moyen :

$$\mathbf{D_{Gamma}} = \mathbf{2} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{ln}(\frac{\hat{\mathbf{y}}_\mathbf{i}}{\mathbf{y_i}}) - \frac{(\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y_i})}{\hat{\mathbf{y}}_\mathbf{i}}$$

Enfin, l'indice de **Gini** sera l'indicateur phare de notre modélisation au sens où il permet de mesurer la qualité de segmentation des modélisations (*Frees, Meyers, Cummings, 2012,* [5]). Il s'obtient à partir de la courbe de Lorenz ordonnée. La courbe de Lorenz ordonnée s'obtient en traçant la proportion de prime (ici nos prédictions) en fonction de la sinistralité (ici nos valeur réelles). Cette courbe se présente ainsi :

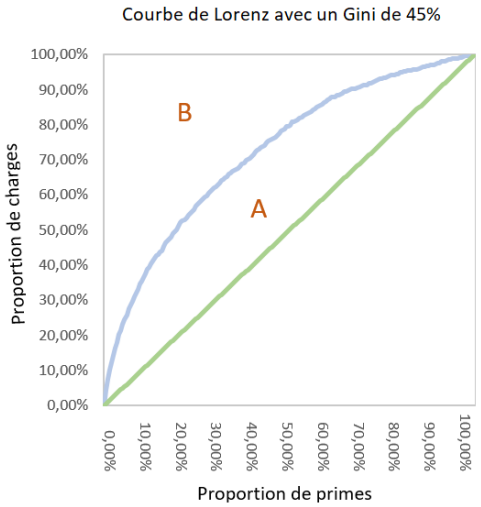


FIGURE 3.1.5 – Courbe de Lorenz ordonnée et indice de Gini

On a alors :

$$\textbf{Gini} = \frac{\textbf{Aire A}}{\textbf{Aire A} + \textbf{Aire B}}$$

Ainsi, un indice de Gini de 0, dans le cas où la courbe de Lorenz serait en fait la bissectrice (en vert), correspond à une tarification uniforme, là où un indice de Gini de 1 correspond à la tarification la plus segmentée possible.

3.2. Le cadre théorique

3.2.1 Le modèle linéaire généralisé

Le modèle linéaire généralisé, aussi appelé GLM, est le modèle classique utilisé en tarification. Il offre de bonnes performances tout en étant facilement interprétable contrairement à des modèles de machine learning. C'est donc ces modèles que nous allons mettre oeuvre pour notre modélisation. Dans cette section nous rappellerons le formalisme mathématique associé à ce type de modèle (*Thomas, 2021,* **[6]**)

Dans le cadre des modèles linéaires généralisés la notion de familles exponentielles est importante. Soit Y une variable aléatoire suivant une loin appartenant à la famille exponentielle, alors la densité de Y peut s'écrire sous la forme :

$$\mathbf{f}_{\theta,\phi}(\mathbf{y}) = \mathbf{c}_{\phi}(\mathbf{y}).\mathbf{exp}(\frac{\mathbf{y}\theta - \mathbf{a}(\theta)}{\phi})$$

où :

- θ est appelé paramètre canonique.

- ϕ est appelé paramètre de dispersion et qui est considéré comme un paramètre de nuisance n'ayant pas d'utilité directe.

- $a(\theta)$ de classe C^2 et convexe.

- $c_{\phi}(y)$ ne dépend pas de θ

on aura alors :

$$\mathbb{E}(\mathbf{y}) = \mathbf{a}'(\theta) \quad \text{et} \quad \mathbf{Var}(\mathbf{y}) = \phi\mathbf{a}''(\theta)$$

Les distributions suivantes appartiennent par exemple à la famille exponentielle :

- La loi normale de paramètre m et σ^2, avec σ^2 connu où : $\theta = m$, $a(\theta) = \sigma^2/2$ et $\phi = \sigma^2$;

- La loi gamma de paramètre k et λ, avec k connu où : $\theta = -1/\lambda$, $a(\theta) = -klog(-\theta)$ et $\phi = 1$

- la loi de Poisson de paramètre λ où : $\theta = log\lambda$, $a(\theta) = exp(\theta)$ et $\phi = 1$

En revanche la loi log-normale n'appartient par exemple pas à la famille exponentielle.

La définition d'une famille exponentielle ayant été donnée, il est désormais possible de définir le modèle linéaire généralisé.

Soit Y la variable aléatoire la variable réponse et X le vecteur aléatoire repésentant les variables explicatives, et β le vecteur des paramètres du modèles et notons $\mu(X) = \mathbb{E}(Y|X)$. Un modèle peut être qualifié de modèle linéaire généralisé s'il vérifie les deux hypothèses :

- $Y|X = x \sim \mathbb{P}_{\theta,\phi}$ appartient à une famille exponentielle.

- $g(\mu(X) = X\beta)$ pour une certaine fonction g bijective que l'on nomme fonction de lien.

Toute loi appartenant à la famille exponentielle possède une fonction de lien dite canonique. En effet, en notant $g = (a')^{-1}$, on peut écrire :

$$\mu = \mathbf{g^{-1}}(\mathbf{X}\beta)$$

Ainsi grâce aux modèle linéaire généralisés, on peut relier l'espérance de notre variable à expliquer aux variables explicatives. Pour cela, il est nécéssaire d'estimer les paramètres du modèle β. Ces estimations se font grâce à l'estimateur du maximum de vraisemblance. Dans le cas d'une famille exponentielle la log-vraisemblance se calcule de la façon suivante en considérant les Y_i indépendants :

$$\mathbf{l}(\beta) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{log\ f}(\mathbf{Y_i}, \beta, \phi) = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{log\ c}_\phi(\mathbf{Y_i}) + \frac{\mathbf{Y_i}\theta_\mathbf{i} - \mathbf{a}(\theta)}{\phi})$$

Pour maximiser la vraisemblance, l'équation suivant doit être résolue :

$$\frac{\partial \mathbf{l}}{\partial \beta_\mathbf{j}} = \mathbf{0}$$

Une fois les paramètre estimés, il est essentiel de vérifier la validité de notre modèle. Pour cela, on compare notre modèle à un modèle dit saturé qui a autant de paramètres que de variables réponses (représentant parfaitement les données). Ainsi en notant $\hat{l}$ la vraisemblance de notre modèle et $\tilde{l}$ la vraisemblance du modèle saturé, on peut comparer les modèles en calculant la déviance définie ainsi :

$$\mathbf{\Delta} = \mathbf{2}(\tilde{\mathbf{l}} - \hat{\mathbf{l}})$$

De grande valeur de Δ indique que le modèle a mal estimé les données par rapport au modèle saturé et on cherche donc à avoir la déviance la plus faible possible.

3.2.2 Une extension du GLM : le GAMLSS

— L'adéquation à une loi de la famille exponentielle :

Comme énoncé précedemment, le modèle linéaire généralisé repose sur l'hypothèse que la variable à expliquer suive une loi appartenant à la famille exponentielle (bien que certains auteurs considèrent que cette hypothèse n'est pas essentielle) . Ainsi, il est nécéssaire de vérifier que nos idées suivent bien une de ces lois.

Les lois candidates pour la modélisation de la fréquence sont les suivantes :

- la loi de Poisson, loi classique pour les données de comptage et donc définie sur $\mathbb{N}$, à 1 paramètre $\mu > 0$ (paramètre de position).
Sa fonction de masse est définie ainsi :

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mu) = \frac{\mathbf{exp}(\mu).\mu^{\mathbf{y}}}{\mathbf{y!}}$$

et a pour propriété $\mathbb{E}(Y) = Var(Y) = \mu$, l'indice de dispersion d'une loi de Poisson vaut donc 1.

- la loi binomial négative, définie sur $\mathbb{N}$ et à 2 paramètres $\mu > 0$ (paramètre de position) et $\sigma > 0$ (paramètre d'échelle). La paramétrisation donnée par *Anscombe (1950)* lui donne pour fonction de masse :

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mu, \sigma) = \frac{\mathbf{\Gamma}(\mathbf{y} + \mathbf{1}/\sigma)}{\mathbf{\Gamma}(\mathbf{1}/\sigma)\mathbf{\Gamma}(\mathbf{y} + \mathbf{1})}.\left(\frac{\sigma\mu}{\mathbf{1} + \sigma\mu}\right)^{\mathbf{y}}.\left(\frac{\mathbf{1}}{\mathbf{1} + \sigma\mu}\right)^{\mathbf{1}/\sigma}$$

Avec Γ la fonction gamma. La loi binomiale négative a pour moments : $\mathbb{E}(Y) = \mu$, $Var(Y) = \mu + \sigma\mu^2$. La loi binomiale négative est notamment utilisée lorsqu'une variable est sur-dispersée, c'est à dire si $Var(Y) > \mathbb{E}(Y)$.

Les lois candidates pour la modélisation du coût moyen sont les suivantes :

- la loi gamma, définie sur $\mathbb{R}_+^*$, qui est particulièrement adaptée pour pour les données ayant un coefficient d'asymétrie positif, a 2 paramètres $\mu > 0$ (paramètre de position) et $\sigma > 0$ (paramètre d'échelle). Il est possible de paramétrer sa densité de la façon suivante :

$$\mathbf{f_Y}(\mathbf{Y} = \mathbf{y}|\mu, \sigma) = \frac{\mathbf{y}^{\mathbf{1}/\sigma^{\mathbf{2}}-\mathbf{1}}\mathbf{exp}(-\frac{\mathbf{y}}{\sigma^{\mathbf{2}}\mu})}{(\sigma^{\mathbf{2}}\mu)^{\mathbf{1}/\sigma^{\mathbf{2}}}\mathbf{\Gamma}(\mathbf{1}/\sigma^{\mathbf{2}})}$$

La loi gamma a pour moments : $\mathbb{E}(Y) = \mu$, $Var(Y) = \sigma^2\mu^2$.

- la loi inverse gaussienne, définie sur $\mathbb{R}_+^*$, à 2 paramètres $\mu > 0$ (paramètre de position) et $\sigma > 0$ (paramètre d'échelle). Elle est également adaptée pour les données ayant un coefficient d'asymétrie positif. Il est possible de paramétrer sa densité de la façon suivante :

$$\mathbf{f_Y(Y = y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}}.exp\left[- \frac{1}{2\mu^2\sigma^2 y}(y - \mu)^2 \right]}$$

La loi inverse a pour moments : $\mathbb{E}(Y) = \mu$, $Var(Y) = \sigma^2\mu^3$.

Ainsi, afin de respecter le cadre théorique des modèles linéaires généralisés, l'adéquation de nos variables à expliquer à ces 4 lois doit être vérifiée. Pour cela, on va ajuster nos données à ces lois, puis tracer des graphiques quantile-quantile afin de comparer distributions empiriques et théoriques.

Regardons les QQ-plots pour la fréquence :

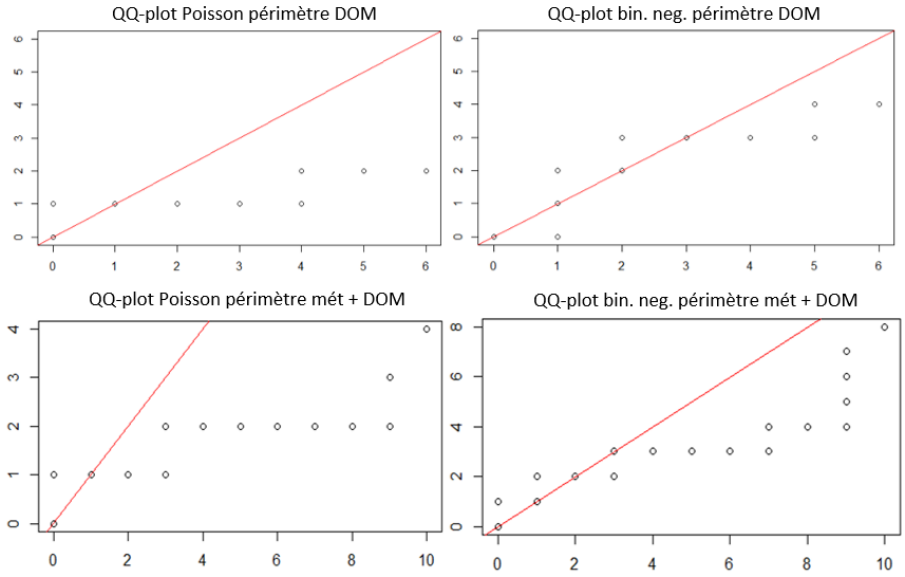


FIGURE 3.2.1 – QQ-plots du nombre de sinistre sur les périmètres DOM et métropole + DOM

D'après les QQ-plots, aussi bien sur le périmètre "DOM" que sur le périmètre "métropole+DOM", la loi binomiale négative est la plus adaptée. Cela s'explique par la sur-dispersion des données dans les deux cas. En effet, sur le périmètre "DOM" on a : $\mathbb{E}(Y) = 0,11$ et $Var(Y) = 0,33$ (importante sur-dispersion), là où sur le périmètre "métropole+DOM", on a : $\mathbb{E}(Y) = 0,24$ et $Var(Y) = 0,29$ (légère sur-disperion).

Regardons à présent les QQ-plots pour le coût moyen :

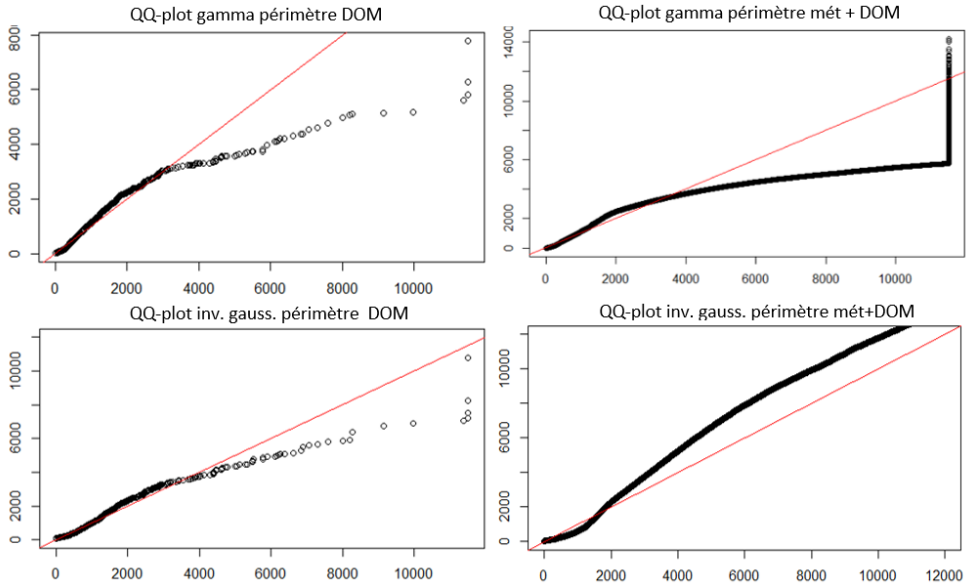


FIGURE 3.2.2 – QQ-plots de la charge sinistre sur les périmètres DOM et métropole + DOM

Sur le périmètre "DOM", la loi inverse gaussienne est la plus adaptée. Alors que c'est l'inverse sur le périmètre "métropole+DOL".

— **Introduction aux GAMLSS :**

Aussi bien pour la fréquence que pour le coût moyen, et sur chaque périmètre, on peut remarquer que nos lois ne semblent pas tout à fait adaptée. Cela est principalement dû au fait que nous avons des données fortement asymétriques et de la sur-dispersion pour la fréquence. Ainsi, nous avons besoin d'un modèle tenant compte de la particularité de nos données.

Ainsi, nous introduisons les modèles *GAMLSS* (pour Generalized Additive Models for Location, Scale and Shape) développés par *Rigby et Stasinopoulos (2001, 2005)* ayant pour objectif de contourner les limitations des GLM et des GAM (modèles additifs généralisés) et se placent donc comme une extension de ceux-ci.

Le modèle GAMLSS relâche l'hypothèse selon laquelle la variable réponse doit appartenir à la famille exponentielle, et la remplace par une famille de distribution plus générale qui inclue les distributions fortement asymétrique (relatif au coefficient d'asymétrie ou skewness) et/ou avec un excès d'aplatissement

(relatif au kurtosis). Enfin le modèle GAMLSS permet non seulement de modéliser le paramètre de position noté μ (la moyenne dans la plupart des cas) mais aussi le paramètre d'échelle noté σ (dirigeant la dispersion) et les paramètres de formes notés ν et τ (régissant la forme de la distribution, donc relatifs au skewness et au kurtosis). C'est notamment le paramètre μ qui va nous intéresser dans notre étude. Ainsi, le modèle GAMLSS peut en théorie répondre à nos problèmes de sur-dispersion et d'asymétrie (*Stasinopoulos, Rigby, 2017,* **[7]**).

Le modèle statistique du GAMLSS présuppose d'avoir des observations indépendantes $(y_i)_{i=1,..,n}$ de densité $f(y_i|\mu i, \sigma_i, \nu_i, \tau i)$. Chacun de ces paramètres peut être fonction des variables explicatives. En effet, en notant $y^T = (y_i)_{i=1,..,n}$ le vecteur de longueur n des variables réponses, et pour $k = 1, 2, 3, 4$: g_k la fonction lien du k-ème paramètre, β_k les vecteurs de paramètres de longeur J_k, X_k la matrice des variables explicatives et $h_{j,k}$ une fonction lisse et non-paramétrique (aspect GAM) des variables explicatives $x_{j,k}$ pour $j = 1, 2, ..., J_k$, les paramètres se modélisent de la façon suivante :

$$\mathbf{g_1}(\mu) = \mathbf{X_1}\beta_\mathbf{1} + \sum_{\mathbf{j=1}}^{\mathbf{J_1}} \mathbf{h_{j,1}}(\mathbf{x_{j,1}})$$

$$\mathbf{g_2}(\sigma) = \mathbf{X_2}\beta_\mathbf{2} + \sum_{\mathbf{j=2}}^{\mathbf{J_2}} \mathbf{h_{j,2}}(\mathbf{x_{j,2}})$$

$$\mathbf{g_3}(\nu) = \mathbf{X_3}\beta_\mathbf{3} + \sum_{\mathbf{j=3}}^{\mathbf{J_3}} \mathbf{h_{j,3}}(\mathbf{x_{j,3}})$$

$$\mathbf{g_4}(\tau) = \mathbf{X_4}\beta_\mathbf{4} + \sum_{\mathbf{j=4}}^{\mathbf{J_4}} \mathbf{h_{j,4}}(\mathbf{x_{j,4}})$$

De même que pour le GLM, les paramètres β_k sont estimés grâce à l'estimateur du maximum de vraisemblance.

Enfin, la sélection de modèle dans le cadre des GAMLSS se fait relativement différemment des GLM. Là où le GLM va utiliser la déviance, pour les GAMLSS, on a deux métriques utilisées :

- La **déviance globale** :

$$\mathbf{GDEV} = \mathbf{-2log}(\hat{\mathbf{l}})$$

avec $\hat{l}$ la vraisemblance de notre modèle.

- L'**AIC généralisé** :

$$\mathbf{GAIC} = -\mathbf{2log}(\hat{\mathbf{l}}) + \kappa.\mathbf{d}$$

avec κ la pénalité, $\kappa = 2$ correspondant à l'AIC par exemple, et d le nombre de degré de liberté.

L'intérêts de ces deux mesures et de pouvoir comparer différentes distributions. Ainsi, on les utilisera dans le but de choisir la distribution adéquate plutôt que pour nous renseigner sur le pouvoir prédictif de nos modèles. Dans ce dernier cas, on utilisera les indicateurs précédemment évoqués en section 3.1.2.

— **Les distributions du package 'gamlss' :**

Un des grand avantage des GAMLSS est sans aucun doute le large choix de distribution disponible dans le package R associé (nommé "gamlss"). Il en existe plus de 100 (et il est même possible d'en créer de nouvelles).

De ce fait, il est essentiel de savoir si un nouveau type de distribution pourrait être mieux adapté à nos données. Pour cela, on va utiliser la fonction *Fitdist* du package "gamlss" qui va ajuster nos données pour chaque type de distribution (discrètes, binomiales, continues ou continues positives) et calculer la deviance global et l'AIC généralisé résultant. Quatre lois ont été retenues suite à cela, 1 pour la fréquence et 3 pour le coût moyen :

- La loi de Sichel qui a été obtenue aussi bien pour la fréquence du périmètre "DOM" que pour celle du périmètre "métropole+DOM". C'est une loi à 3 paramètres $\mu > 0$, $\sigma > 0$ et ν dans $\mathbb{R}$. En notant : $\alpha^2 = 1/\sigma^2 + 2\mu/\sigma$ et $K_\lambda(t) = 1/2 \int_0^\infty x^{\lambda-1} exp(-1/2t(x + x^{-1})dx$, sa fonction de masse s'écrit :

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mu, \sigma, \nu) = \frac{\mu^{\mathbf{y}} \mathbf{K}_{\mathbf{y}+\nu}(\alpha)}{(\alpha\sigma)^{\mathbf{y}+\nu} \mathbf{y}! \mathbf{K}_\nu(\mathbf{1}/\sigma)}$$

Cette loi a pour moments : $\mathbb{E}(Y) = \mu$ et $Var(Y) = \mu + \mu^2 g_1$ avec $g_1 = 2\sigma(\nu + 1)/b + 1/b^2$ et $b = K_{\nu+1}(\sigma^{-1})/K_\nu(\sigma^{-1})$

- La loi log-normale, qui a été retenue pour le coût moyen des deux périmètres "DOM" et "métropole+DOM". Elle a deux paramètres μ dans $\mathbb{R}$ et σ dans R_+^*. Il est possible de paramétrer sa densité de la façon suivante :

$$\mathbf{f_Y(Y = y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}exp\left\{ - \frac{(logy - \mu)^2}{2\sigma^2} \right\}}$$

La loi log-normale a pour moments : $\mathbb{E}(Y) = exp(\mu + \sigma^2/2)$ et $Var(Y) = exp(2\mu + \sigma^2).(exp(\sigma^2) - 1)$

- La loi Beta généralisée de type 2. Elle a été retenue pour le coût moyen du périmètre "métropole+DOM". C'est une loi à 4 paramètres (μ, σ, ν, τ) dans $(\mathbb{R}_+^*)^4$. Il est possible de paramétrer sa densité de la façon suivante :

$$\mathbf{f_Y(Y = y|\mu, \sigma, \nu, \tau) = \frac{\Gamma(\nu + \tau)}{\Gamma(\nu)\Gamma(\tau)}.\frac{|\sigma|(y/\mu)^{\sigma\nu}}{y[1 + (y/\mu)^{\sigma}]^{\nu+\tau}}}$$

Les moments de cette loi seront détaillés en annexe 1.

- La loi Box cox t, qui a été retenue pour le coût moyen du périmètre "DOM". C'est aussi une loi a 4 paramètres $\mu > 0$, $\sigma > 0$, ν dans $\mathbb{R}$ et $\tau > 0$. Elle est définie à partir d'une variable aléatoire Z suivant une distribution tronquée t (détails de cette distribution en annexe 2) de degré de liberté τ. Sa densité est paramétrée de la façon suivante :

$$\mathbf{f_Y(Y = y|\mu, \sigma, \nu, \tau) = \frac{y^{\nu-1}f_T(z)}{\mu^{\nu}\sigma F_T(1/(\sigma|\nu|))}}$$

où $f_T(t)$ et $F_T(t)$ sont respectivement la densité et la fonction de répartition d'une variable aléatoire T suivant une distribution tronquée t de degré de liberté τ et z définit par :

$$\mathbf{Z} = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\mu} - 1\right] & si\ \nu \neq 0 \\ \frac{1}{\sigma}log(\frac{Y}{\mu}) & si\ \nu = 0 \end{cases}$$

Il est désormais possible de tracer les QQ-plots de la charge et de la fréquence en fonction des nouvelles distributions afin de vérifier si elles sont plus adaptées à nos données :

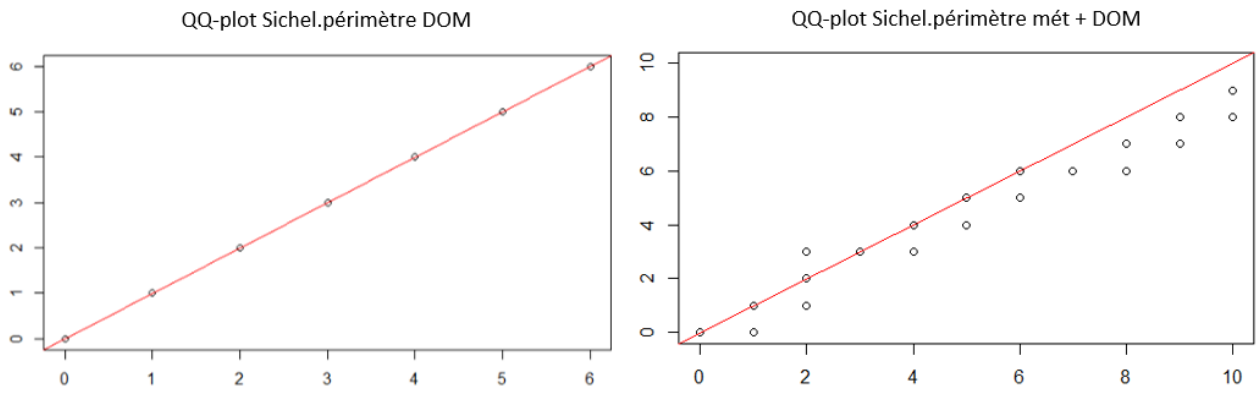


FIGURE 3.2.3 – QQ-plots du nombre de sinistre avec la loi de Sichel

Comparés aux lois de Poisson et binomiale négative, la loi de Sichel semble bien plus adaptée, notamment pour le périmètre DOM où l'on retrouve bien l'équation $y = x$ dans le QQ-plot.

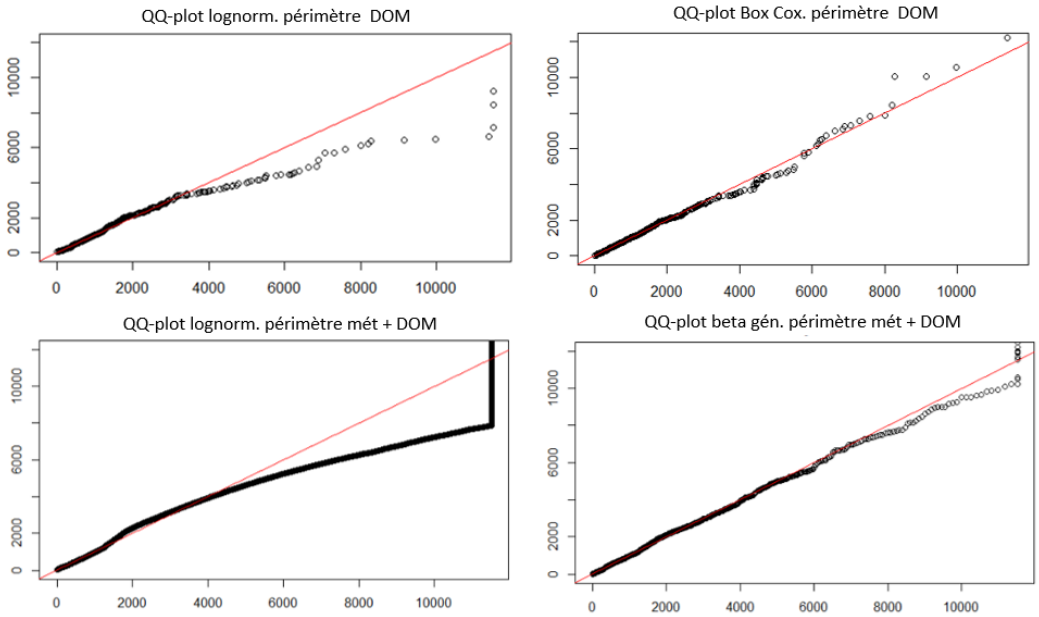


FIGURE 3.2.4 – QQ-plots de la charge sinistre avec les lois lognormale, beta généralisée et Box Cox

Sur le périmètre "DOM", la distribution log-normale semble proche de la distribution inverse-gaussienne d'après les QQ-plots. Sur les deux périmètres, elle semble mieux ajuster les données que la loi Gamma. Enfin, les lois Box-Cox et Beta généralisée ajuste presque parfaitement nos données.

Cependant, l'analyse des QQ-plots ne suffit pas, il nous faut calculer la déviance globale et/ou l'AIC généralisé afin de déterminer quelle loi candidate ajuste au mieux nos données.

Regardons tout d'abord le cas de la fréquence :

Distribution	GDEV D	GAIC D	GDEV D	GAIC M+D
Poisson	19 929	19 931	381 896	381 898
Binomiale négative	18 486	18 490	363 399	363 403
Sichel	18 325	18 331	361 191	361 197

TABLE 3.2.1 – Détails des GDEV et GAIC pour les distributions candidates de la fréquence

Grâce aux métriques, il est confirmé que la loi de Sichel est la loi ajustant le mieux nos données. Cependant, la déviance globale et l'AIC généralisé des lois de Sichel n'est pas si éloignés de ceux de la loi binomiale négative , là où la loi de Poisson n'est clairement pas adaptée.

On présente ensuite un tableau similaire pour les distributions candidates du coût moyen :

Distribution	GDEV D	GAIC D	GDEV D	GAIC M+D
Gamma	25 208	25 212	5 995 090	5 995 094
Inverse gaussienne	25 216	25 220	6 293 070	6 293 074
Log-normale	24 990	24 994	5 941 894	5 941 898
Box Cox t	24 898	24 906	-	-
Beta généralisée 2	-	-	5 931 410	5 931 418

TABLE 3.2.2 – Détails des GDEV et GAIC pour les distributions candidates du coût moyen

Sur le périmètre "DOM", là distribution Box Cox t est la plus adaptée, vient ensuite la distribution log-normale, la distribution gamma et enfin la distribution inverse gaussienne. Les conclusions sont les mêmes sur le périmètre "métropole + DOM" où la loi beta généralisée est la plus adaptée au contraire de la loi inverse gaussienne qui ajuste encore moins bien les données que sur le périmètre "DOM".

Enfin, dans l'environnememnt GAMLSS, il existe un dernier outil graphique dans l'analyse de l'ajustement à une distribution. Il s'agit des worm plots (introduits par *Buuren et Fredriks (2001)* , qui sont une forme de QQ-plot dont la tendance a été retirée afin de mieux voir les régions où l'ajustement n'est pas adéquat. Le worm plot représente ainsi la déviation des résidus normalisés à une loi normale standardisée.

On présente ci-dessous les graphiques worm plots (pour certaines distributions) de la fréquence et du coût moyen pour le périmètres "DOM" :

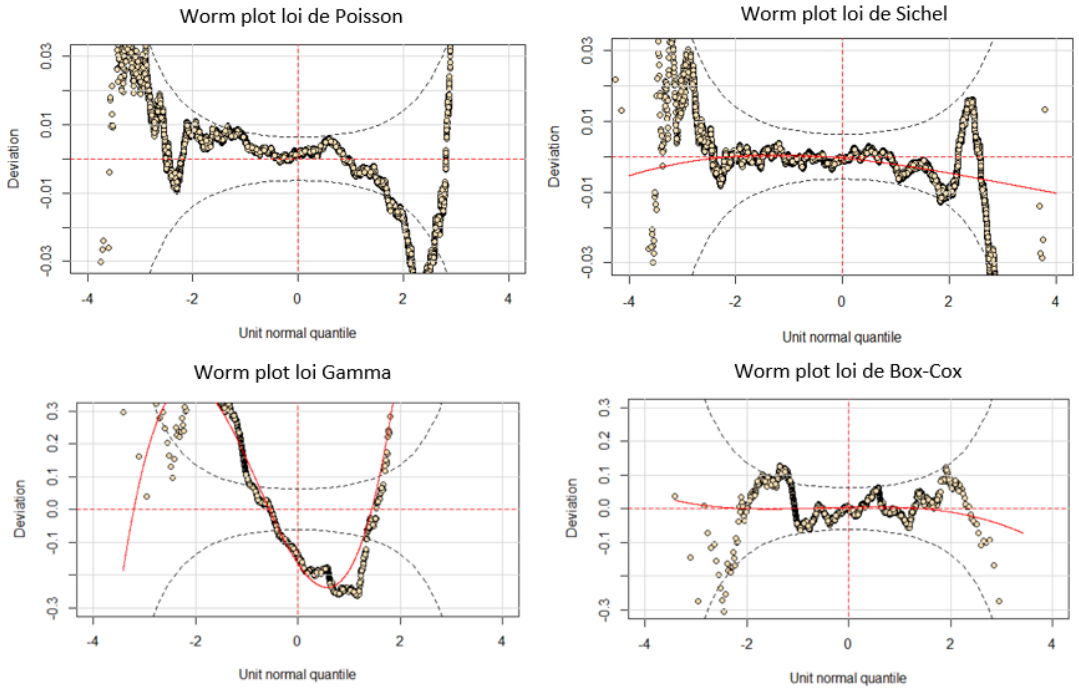


FIGURE 3.2.5 – Worm plots du nombre de sinistre et de la charge pour le périmètre DOM

Les points des graphiques montre à quel points les résidus (ordonnés) sont éloignés des valeurs attendues représentée par la ligne horizontale. Les courbes elliptiques représentent les intervalles de confiance à 95%. Si le modèle est correct, 95% des points devraient se trouver entre les deux courbes.

Ces worm plots, confirment bien que la loi de Sichel est bien plus adaptée que la loi de Poisson pour ajuster la fréquence sur le périmètre "DOM" : sa tendance (courbe rouge) est en effet plus proche de la ligne horizontale que la loi de Sichel, et les points sont globalement tous dans l'intervalle de confiance. Le constat est encore plus clair pour le coût moyen où la loi de Box-Cox ajuste mieux les données au vu du nuage de points se situant bien dans l'intervalle de confiance, contrairement à la loi Gamma où de nombreux points ne sont pas situés dans cette intervalle et avec une tendance s'écartant fortement de la ligne horizontale.

— **Les étapes de la construction d'un modèle GAMLSS :**

Enfin, nous allons détailler les étapes permettant d'obtenir le modèle GAMLSS le plus abouti possible (au sens de la qualité de l'adéquation via l'AIC généralisé). La construction d'un modèle GAMLSS se fait ainsi en quatre étapes :

- La sélection de la distribution : pour une variable réponse, on commence généralement par définir quelle est la loi la plus adéquate. Cependant, dans le cadre de notre étude, des modèles pour toutes les lois mentionnés seront ajustés. On pourra ainsi voir si les lois ajustant le mieux nos données permettent d'obtenir le meilleur pouvoir prédictif.

- La sélection de la fonction de lien pour chacun des paramètres. La fonction utilisée dans nos modèles sera toujours le *log* dans le but d'obtenir un modèle multiplicatif. En effet, on préfèrera ce type de modèle dans un contexte de prime pure où les variations relatives d'un coefficient à un autre seront plus interprétables que les variations absolues (*Ruoyan, (2004),* **[8]**).

- La sélection des termes apparaissant dans notre modèle, c'est à dire la sélection de variables.

- La sélection des fonctions de lissages à appliquer.

Toutes ces étapes auront pour objectif d'améliorer notre modèle de proche en proche. Les étapes de sélections de variables et de sélections de lissages et les aspects théoriques associés seront en particulier détaillés dans la section suivante.

3.3. La modélisation de la fréquence hors zonier

3.3.1 Modélisation de la fréquence sur le périmètre "DOM"

Dans cette section, nous allons nous attarder sur la modélisation de la fréquence, hors zonier, du périmètre "DOM" . Nous allons mettre en oeuvre un modèle permettant de prédire le nombre moyen de sinistre (la fréquence) qu'aura un assuré sur 1 année. La variable réponse est donc ici le nombre de sinistre corrigé de l'exposition N (avec $N = nombre\ de\ sinistre/exposition$). Notons qu'il est aussi possible de modéliser uniquement le nombre de sinistre (non corrigé) et de tenir compte de l'exposition grâce à une variable de décalage introduite dans nos modèles (variable offset). Ces deux méthodes sont équivalentes.

Nos données seront ajustées aux trois lois de probabilité candidates pour la fréquence : la loi de Poisson, la loi binomiale négative et la loi de Sichel. C'est en particulier le paramètre μ, correspondant à la moyenne des distributions considérées, qui sera modélisé avec la fonction de lien *log*. A chaque étape de notre modélisation, à l'aide de la validation croisées à 4-folds, on mesurera d'une part l'adéquation de notre modèle aux données avec la déviance globale et d'autre part le pouvoir prédictif des modèles avec les métriques cités en section 3.1.2 et notamment l'indice de Gini.

— **Modèles sans sélection de variables :**

Trois modèles basés sur nos 3 distributions, sans aucune sélection de variables, c'est à dire utilisant toutes les variables explicatives disponibles, sont d'abord implémenter afin d'ajuster nos données. Les mesures d'adéquation suivantes sont obtenues :

Distribution	Test. GDEV
Poisson	19 448
Binomiale négative	18 206
Sichel	17 692

TABLE 3.3.1 – Déviance globale test des modèles sans sélection de variable pour la fréquence sur "HZ DOM"

La déviance globale test est obtenue en calculant la moyenne des déviances globales sur les folds tests (sur lesquels le modèle n'a pas été entraîné). Ici, il est clair que la loi de Sichel est la plus adaptée. Les conclusions déjà établies sur la distribution de la variable réponse sont donc les mêmes avec sa distribution conditionnelle. Mais il est important de vérifier que ce modèle ait un bon pouvoir prédictif :

Métriques	Poisson	Bin. Neg	Sichel
Gini Train K-Fold	41,85%	42,67%	39,46%
Gini Test K-Fold	33,14%	33,97%	34,67%
RMSE Train K-Fold	0,1807	0,1807	0,1808
RMSE Test K -Fold	0,1674	0,1673	0,1674
MAE Train K-Fold	0,023	0,023	0,0229
MAE Test K-Fold	0,0231	0,0231	0,0230
Poisson deviance Train -Fold	0,1037	0,1038	0,1051
Poisson deviance Test K-Fold	0,10803	0,10873	0,10773
Erreur globale Train K-Fold	0,000%	-0,1894%	0,4926%
Erreur globale Test K-Fold	-0,6954%	-0,9051%	-0,1615%

TABLE 3.3.2 – Performances sans sélection de variable pour la fréquence sur "HZ DOM"

Les résultats sur les "tests k-folds" sont ceux nous intéréssant particulièrement. L'indice de Gini le plus élevé de 34,67% correspond à la distribution de Sichel. De plus, les distributions de Poisson et binomiale négative ont un indice de Gini plus élevé sur les "train k-folds" plus élevé, cela indique que la modélisation avec une distribution de Sichel permet d'obtenir un sur-apprentissage plus faible.

— **Modèles avec sélection de variables :**

L'étape précédente a permis de s'assurer que la distribution de Sichel est mieux adaptée à nos données. A présent nous devons réaliser une sélection qui permettra d'une part de retenir uniquement les variables servant à expliquer la variable réponse et d'autre part à améliorer le pouvoir prédictif de nos modèles.

Nous allons utiliser deux méthodes pour la sélection de nos variables :

- La première est la pénalisation par LASSO (pour Least Absolute Shrinkage and Selection Operator) qui va ajouter des contraintes à l'estimation des coefficients et en reprenant les notations précédentes sur les GLM (et GAMLSS) :

$$\beta_{\mathbf{LASSO}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \beta' \mathbf{X_i})^{\mathbf{2}} + \lambda \sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_{\mathbf{j}}| \quad \text{avec } \lambda \in \mathbb{R}^+$$

Le LASSO permet donc de favoriser la parcimonie du modèle en forçant les variables les moins importantes à avoir un coefficient nul (notamment pour λ élevé). Les coefficients λ optimaux sont obtenus par validation croisée en testant un maximum de possibilité (technique du "grid search").

Avec cette méthode, on obtient les déviances globales suivantes :

Distribution	Test. GDEV
Poisson	19 228
Binomiale négative	17 953
Sichel	17 464

TABLE 3.3.3 – Déviance globale test avec sélection de variable LASSO pour la fréquence sur "HZ DOM"

La sélection de variable par LASSO a permis de réduire les déviances globales des trois distributions de 250 environ, les modèles ont donc été améliorés.

Regardons donc à présent les performances de ces modèles avec la sélection de variables par LASSO :

Métriques	Poisson	Bin. Neg	Sichel
Gini Train K-Fold	41,38%	41,92%	40,37%
Gini Test K-Fold	36,21%	36,65%	36,71%
RMSE Train K-Fold	0,1807	0,1808	0,1808
RMSE Test K -Fold	0,1673	0,1673	0,1672
MAE Train K-Fold	0,0230	0,0231	0,0229
MAE Test K-Fold	0,0231	0,0231	0,0229
Poisson deviance Train -Fold	0,1041	0,1042	0,1049
Poisson deviance Test K-Fold	0,1065	0,1064	0,1062
Erreur globale Train K-Fold	0,000%	-0,2357%	0,1021%
Erreur globale Test K-Fold	-0,6662%	-0,9111%	0,4470%

TABLE 3.3.4 – Performances avec sélection de variable LASSO pour la fréquence sur "HZ DOM"

L'amélioration des modèles est confirmée par les indicateurs de performances. Les indices de Gini ont gagné environ 2 points sur les "tests k-folds" (ce qui est important), et la déviance de poisson a diminué. On remarque notamment que l'indice de Gini du modèle utilisant la loi négative binomiale (Gini de 36,65 %) devient très proche de celui du modèle utilisant la loi de Sichel (Gini de 36,71%).

- La seconde méthode de sélection de variables est l'approche par optimisation de critère. Le critère à maximiser dans notre cas sera l'AIC généralisé. C'est la méthode *backward* qui va être appliquée : on commence par l'ajustement d'un modèle avec toutes les variables d'intérêt incluses. La variable la moins significative est exclue, la significativité étant basée sur le critère du GAIC. On continue en réajustant successivement les modèles réduits jusqu'à ce que toutes les variables restantes dans le modèle soient significatives (c'est à dire qu'il n'est plus possible de réduire le GAIC du modèle).

Nous allons couplée cette approche à la validation croisée. On va ainsi appliquer la méthode backward

à chacun de nos 4 folds. Chacun des quatre modèles aura alors retenu les variables les plus significatives. Pour déterminer les variables finales à conserver, nous allons compter le nombre de fois où les variables ont été retenues, et nous conserverons uniquement celles apparaissant au moins 2 fois.

Ainsi, pour nos distributions, on relance les modèles avec les variables retenues dans chacun des cas. On obtient les déviances globales suivantes :

Distribution	Test. GDEV
Poisson	19 137
Binomiale négative	17 832
Sichel	17 314

TABLE 3.3.5 – Déviance globale test avec sélection de variable backward pour la fréquence sur "HZ DOM"

De la même manière que pour la sélection de variables avec LASSO, les déviance globales ont diminué et sont même légèrement meilleures que celles obtenues avec LASSO.

Regardons donc à présent les performances de ces modèles avec la sélection de variables par approche backward :

Métriques	Poisson	Bin. Neg	Sichel
Gini Train K-Fold	40,62%	41,01%	39,84%
Gini Test K-Fold	36,10%	36,41%	36,86%
RMSE Train K-Fold	0,1808	0,1808	0,1808
RMSE Test K -Fold	0,1673	0,1672	0,1672
MAE Train K-Fold	0,0231	0,0231	0,0229
MAE Test K-Fold	0,0231	0,0231	0,0229
Poisson deviance Train -Fold	0,1045	0,1045	0,1051
Poisson deviance Test K-Fold	0,1066	0,1067	0,1065
Erreur globale Train K-Fold	0,000%	-0,1886%	0,1341%
Erreur globale Test K-Fold	-0,6402%	-0,8471%	0,8016%

TABLE 3.3.6 – Performances avec sélection de variable backward pour la fréquence sur "HZ DOM"

En comparant les métriques sur le "test k-folds" entre la sélection de variable par LASSO et celle par approche backward on observe un indice de Gini plus important pour l'approche backward avec la distribution de Sichel (Gini de 36,86 %). De plus les écart de Gini entre le "train k-folds" et le "test k-folds" indiquent que la sélection de variables par approche backward permet de diminuer le sur-apprentissage, bien que les déviances Poisson soient en faveur de la sélection de variable par LASSO.

Au final, ce sont les variables sélectionnées par l'approche backward qui ont été retenues.

Les variables suivantes sont ainsi utilisées dans notre modèle :"Année", "Segment", "Age client", "Option piscine", "Option nouvelles énergies", "Personne morale", "Etudiant", "Boursier", "Etage", "Type d'occupation", "Nombre de pièces" et "Capital objets communs".

— **Modèles avec fonctions de lissage :**

La dernière étape dans la construction d'un modèle GAMLSS est le lissage. Le lissage sert généralement à réduire les irrégularités dans une courbe afin de s'approcher de la tendance. Ici, nous somme dans le cas particulier où nous avons uniquement des variables catégorielles, la notion de lissage n'a donc pas le même sens ici car il n'y a pas de courbe, on va donc plutôt s'intéresser aux coefficients attribués à nos différentes modalités. Notons que le choix des variables sur lesquelles appliquer ces fonctions de lissage a été déterminé "manuellement" en testant plusieurs combinaison et en retenant celles offrant les meilleures performances (cette méthode a été employée à cause d'une comlexité en temps trop élevée). De plus, lors de cette étape, seule la distribution de Sichel servira à ajuster nos données, étant la plus performantes d'un point de vue prédictif.

Nous présentons ici deux méthodes implémentés dans l'environnement GAMLSS permettant de réaliser une forme de lissage avec ce type de variable (on n'abordera pas le formalisme mathématiques associé, présenté dans *Flexible Regression and Smoothing Using GAMLSS in R (2017)*) :

- Tout d'abord, la fonction *pcat*() servant à réduire les modalités des variables catégorielles. L'objectif de cette fonction est de fusionner les modalités ayant des coefficients proches. Cela peut être considéré comme un moyen de classer les différentes modalité de nos variables en groupes de modalités ayant des caractéristiques similaires. Dans notre cas, ce lissage sera appliquer aux variables : "Année", "Nombre de pièces", "Capital objets communs", "Capital objets de valeurs ", "Age du client" et "Ancienneté du contrat".

- Il y a ensuite la fonction *random*() qui permet principalement de réduire les coefficients de chaque modalité autour de la moyenne des coefficients. Ainsi, même si l'estimation des coefficients des modalités sera toujours différente, cette différence sera réduite. Dans notre cas, ce lissage sera appliquer aux variables : "Année", "Nombre de pièces", "Capital objets communs" et "Segment".

Avec ces lissages catégorielles, on obtient les déviances globales suivantes :

Le lissage catgéoriel avec la fonction *pcat*() détériore la déviance globale alors que lissage avec la fonction *random*() améliore ce critère.

Il nous faut regarder à présent les performances de ces modèles :

Distribution	Lissage	Test. GDEV
Sichel	pcat()	17 508
Sichel	random()	17 202

TABLE 3.3.7 – Déviance globale test avec lissage de variable pour la fréquence sur "HZ DOM"

Métriques	Sichel random()	Sichel pcat()
Gini Train K-Fold	39,86%	39,77%
Gini Test K-Fold	36,99%	35,95%
RMSE Train K-Fold	0,1808	0,1808
RMSE Test K -Fold	0,1672	0,1672
MAE Train K-Fold	0,0229	0,0228
MAE Test K-Fold	0,0229	0,0228
Poisson deviance Train -Fold	0,1051	0,1051
Poisson deviance Test K-Fold	0,10640	0,10670
Erreur globale Train K-Fold	1,1966%	2,1952%
Erreur globale Test K-Fold	0,6760%	1,7417%

TABLE 3.3.8 – Performances avec lissage de variable pour la fréquence sur "HZ DOM"

Le lissage avec *pcat*() dégrade effectivement les performances du modèle, l'indice de Gini passant de 36,86 % à 35,95%, la déviance de Poisson augmentant de même que l'erreur globale. A l'inverse le lissage avec *random*() améliore bien le modèle l'indice de Gini étant à présent de 36,99%.

— **Modélisation des autres paramètres :**

En introduction de section, nous avons indiqué vouloir modéliser uniquement le paramètre de position μ. Nous pouvons nous demander si la modélisation des paamètres σ et ν peut nous aider à améliorer la prédiction de μ. Nous allons ainsi modéliser le paramètre σ en fonction des variables "segment" et "personne morale" et le paramètre ν en fonction de la variable "étage". Ces variables ont été sélectionnées manuellement après avoir essayé de multiples combinaisons.

En modélisant ces deux paramètres, notre modèle offre les performances suivantes :

Métriques	Sichel avec modélisation de $\sigma et \nu$
Gini Train K-Fold	40,00%
Gini Test K-Fold	37,15%
RMSE Train K-Fold	0,1808
RMSE Test K -Fold	0,1672
MAE Train K-Fold	0,0229
MAE Test K-Fold	0,0229
Poisson deviance Train -Fold	0,1050
Poisson deviance Test K-Fold	0,1063
Erreur globale Train K-Fold	0,9579%
Erreur globale Test K-Fold	0,3477%

TABLE 3.3.9 – Performances avec modélisation de σ et ν pour la fréquence sur "HZ DOM"

La modélisation des paramètres σ et ν permet en effet d'améliorer les performances du modèle : l'indice de Gini passant de 36,99 % à 37,15 % et l'erreur globale passant d'environ 0,68% à 0,35%. De proche en proche le modèle a donc été grandement amélioré (l'indice de Gini passant notamment de 34,67 % à 37,15 %).

— **Application du modèle sélectionné à la base de données "Z DOM"**

Nous avons selectionné le modèle offrant les meilleures performances grâce à une validation croisée. Le modèle est lui, entraîné sur la totalité de la base "HZ DOM".

En résumé, le modèle final a les caractéristiques suivantes :

- les données sont ajustées grâce à la loi de Sichel

- la fonction de lien utilisée est le *log*.

- les termes additifs sont les variables suivantes : "Année", "Segment", "Age client", "Option piscine", "Option nouvelles énergies", "Personne morale", "Etudiant", "Boursier", "Etage", "Type d'occupation", "Nombre de pièces" et "Capital objets communs".

- la fonction de lissage $random()$ est appliquée sur les variables "Année", "Segment", "Nombre de pièces" et "Capital objets communs".

- le paramètre σ est modélisé avec les variables "Segment" et "Personne morale" et le paramètre ν avec la variable "Etage".

Avec le modèle, les performances sur les bases "HZ DOM" et "Z DOM" sont les suivantes :

Métriques	Modèle final
Gini HZ DOM	39,54%
Gini Z DOM	36,33%
RMSE HZ DOM	0,1838
RMSE Z DOM	0,1474
MAE HZ DOM	0,0229
MAE Z DOM	0,0230
Poisson deviance HZ DOM	0,1052
Poisson deviance Z DOM	0,1054
Erreur globale HZ DOM	1,3645%
Erreur globale Z DOM	2,1099%

TABLE 3.3.10 – Performances du modèle hors zonier final fréquence DOM

Les performances sur la base de données "Z DOM" sont bonnes avec un indice de Gini de 36,33%. Il n'y a donc pas de détérioration des performances en appliquant notre modèle sur bases de données sur laquelle il n'a pas apprise.

3.3.2 Comparaison avec la modélisation de la fréquence "métropole+DOM"

Dans cette section nous allons exposer le modèle de fréquence retenu pour le périmètre "métropole+DOM" sans détailler toutes les étapes comme précédemment, la démarche restant la même. Nous comparerons ensuite les performances de ce modèle avec celle du modèle retenu sur le périmètre "DOM".

— **Le modèle retenu pour le périmètre "métropole+DOM"**

Le modèle retenu pour la modélisation de la fréquence sur le périmètre "métropole+DOM" possède les caractéristiques suivantes :

- les données sont ajustées grâce à la loi de Sichel

- la fonction de lien utilisée est le *log*.

- les termes additifs sont les variables suivantes : "Année", "Segment", "Age client", "Ancienneté du contrat", "Option piscine", "Réseau de distribution","Surface dépendances","Ancienneté du logement", "Type d'occupation", "Nombre de pièces", "Capital objets de valeur" et "Capital objets communs".

Ce modèle a obtenu les performances suivantes sur les bases "HZ métropole+DOM", "Z métropole+DOM" :

Métriques	Modèle final fréquence DOM
Gini HZ MET+ DOM	32,26%
Gini Z MET+ DOM	31,65%
RMSE HZ MET + DOM	0,2426
RMSE Z MET + DOM	0,2748
MAE HZ MET + DOM	0,0476
MAE Z MET + DOM	0,0483
Poisson deviance HZ MET + DOM	0,1852
Poisson deviance Z MET + DOM	0,1908
Erreur globale HZ MET+DOM	0,3908%
Erreur globale Z MET+DOM	2,9547%

TABLE 3.3.11 – Performances du modèle final fréquence métropole+DOM

Bien que les métriques ne soit pas comparable de manière absolue, le modèle de fréquence "DOM" segmente mieux les données "DOM" que le modèle de fréquence "métropole+DOM" sur les données "métropole+DOM" en regardant les indices de Gini. Le modèle de fréquence "métropole+DOM" surapprend en revanche moins. Pour pouvoir comparer les deux modèles, on doit regarder leur performances sur la même base de données.

— **Comparaison des performances des deux modèles**

Les performances des modèles fréquence DOM et métropole + DOM sur la base "Z DOM" sont résumées dans le tableau suivant :

Indicateurs	Modèle fréquence DOM	Modèle fréquence Mét+DOM
Nb variables	12	12
Gini	36,33%	25,18%
RMSE	0,1474	0,1475
MAE	0,0230	0,0229
Avg Deviance	0,1054	0,1078
Erreur globale	2,110%	3,675%

TABLE 3.3.12 – Comparaison des performances des modèles fréquence hors zonier sur la base "Z DOM"

Le modèle basé sur le périmètre "DOM" offre de bien meilleur performances : l'indice de Gini est supérieur de 11 points (Gini de 36,33%) par rapport à celui du modèle basé sur le périmètre "métropole+DOM", cet écart est très important. Toutes les métriques hormis la MAE (quasiment égale des deux côtés) vont dans le sens d'une modélisation hors zonier basée sur le périmètre "DOM".

— **Analyse des écarts**

Nous allons à présent tenter d'expliquer les écarts en étudiant les coefficients associés aux variables hors zonier.

Pour cela, nous allons introduire la notion de *spread*, une mesure de l'importance des variables dans le cas où l'on a des variables catégorielles. Dans le cadre d'un modèle multiplicatif tel que le notre, il se définit, en fonction des différents coefficients associés aux modalités d'une variable, comme suit :

$$\mathbf{Spread} = \frac{\max\limits_{\mathbf{i}} \mathbf{C_{relatif_i}} + \mathbf{1}}{\min\limits_{\mathbf{i}} \mathbf{C_{relatif_i}} + \mathbf{1}} - \mathbf{1}$$

où les $C_{relatif_i}$ sont les coefficients relatifs associés au coefficient absolus C_i définis en fonction d'un coefficient de réference C_{ref} (qui peut être choisi arbitrairement parmis les C_i comme dans notre cas) par :

$$\mathbf{C_{relatif_i}} = \frac{\mathbf{C_i}}{\mathbf{C_{ref}}} - \mathbf{1}$$

Nous présentons ci-dessous un tableau regroupant les spreads associés à chaque variable utilisée dans nos modèles :

Variables	Modèle HZ DOM	Variables	Modèle HZ MET+DOM
Etage	633,08%	Nombre de pièces	863,96%
Boursier	604,65%	Segment	375,87%
Occupation du logement	265,77%	Capital objets communs	205,16%
Capital objets commun	212,04%	Surface des dépendances	187,88%
Segment	173,98%	Capital objets valeur	146,18%
Année	150,65%	Réseau de distribution	105,59%
Age du client	133,13%	Occupation du logement	80,60%
Nombre de pièces	123,08%	Ancienneté du logement	75,65%
Option piscine	86,98%	Option piscine	73,00%
Option Nouvelles Energies	60,50%	Age du client	45,65%
Personne morale	8,07%	Ancienneté du contrat	38,72%
Etudiant	3,19%	Année	27,49%

TABLE 3.3.13 – Spread des variables hors zonier des modèles fréquence

L'importance des variables n'est globalement pas la même entre le modèle fréquence "DOM" et le modèle fréquence "métropole+DOM". Les variables ayant les écarts en valeur absolue de spread les plus importants sont :

- la variable "Nombre de pièces" avec un écart de spread de 740 points.

- la variable "Segment" avec un écart de 200 points.

- la variable "Occupation" du logement avec un écart de spread de 160 points.

- la variable "Année" avec un écart de spread de 125 points.

Traçons les coefficients relatifs des variables "Nombre de pièces" et "Segment" afin de mieux comprendre les écarts.

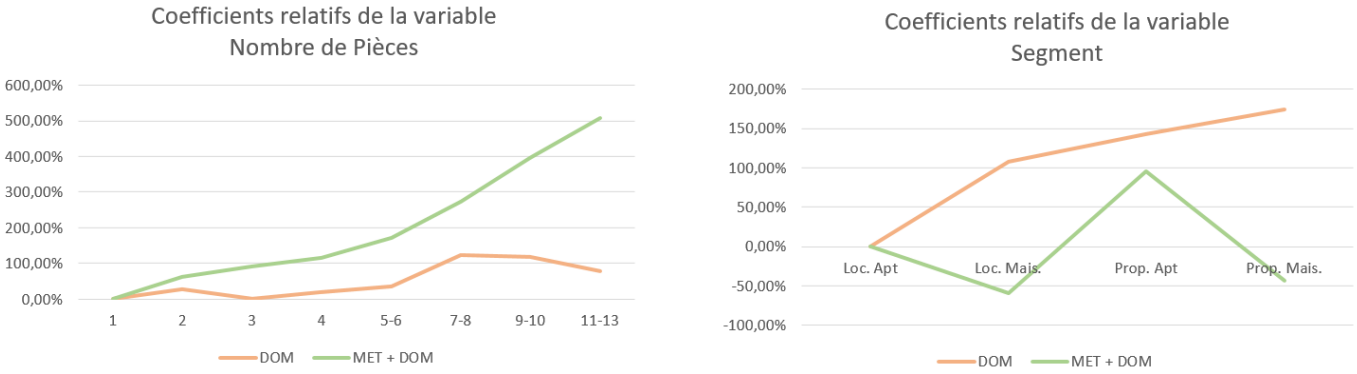


FIGURE 3.3.1 – Comparaison des coefficients relatifs pour les modèles fréquence

Pour la variable "Nombre de pièces", les coefficients relatifs à la modalité "1 pièce" sont bien plus élevés dans le cas du modèle fréquence "métropole+DOM", ce modèle va donc estimer à la hausse le risque associé à cette variable par rapport au modèle fréquence "DOM". Concernant la variable "Segment", les coefficients relatifs à la modalité "locataire d'appartement" sont plus élevés sur le modèle fréquence "DOM", il discrimine donc à la hausse les modalités de la variable "Segment" par rapport au modèle fréquence "métropole+DOM". De plus, les coefficients relatifs de ce modèle ne suivent pas du tout la même tendance que les coefficients relatifs du modèle fréquence "métropole+DOM" (principalement du fait du lissage avec la fonction *random*()). L'étude de ces deux variables fournit donc une piste sur l'explication de la différence de performances entre les deux modèles.

3.4. La modélisation du coût moyen hors zonier

3.4.1 Modélisation du coût moyen segment "DOM"

Dans cette section, la modélisation du coût moyen hors zonier sur le périmètre "DOM" sera détaillée. Un modèle permettant de prédire la charge moyenne d'un sinistre (la coût moyen) pour un risque donné sur une année. Ainsi, la variable réponse est ici la charge vieillie attritionnelle.

Les données seronts ajustées aux 4 distributions candidates pour la modélisation du coût moyen sur le périmètre "DOM". Elles sont les suivantes : la loi gamma, la loi inverse gaussienne, la loi log-normale et la loi de Box-Cox t. De la même manière que pour la fréquence, la qualité de l'ajustement aux distributions candidates sera dans un premier temps mesurée avec les déviances globales. Enfin, le pouvoir prédictif des modèles sera évalué avec les indicateurs de performance et notamment l'indice de Gini.

— **Modèles sans sélection de variable :**

La modélisation du coût moyen hors zonier et sans sélection de variables produit les déviances globales suivantes :

Distribution	Test. GDEV
Gamma	25 074
Inverse Gaussienne	25 154
Log-normale	24 956
Box-Cox	24 881

TABLE 3.4.1 – Déviance globale test avec sélection de variable pour le coût moyen sur "HZ DOM"

La loi de Box-Cox est celle ajustant le mieux nos données (GDEV de 24 881) suivie de la loi log-normale (GDEV de 24 956). Les conclusions déjà établies sur la distribution de la variable réponse sont donc les mêmes avec sa distribution conditionnelle. Il faut cependant vérifier les performances de ces modèles sur le "test k-folds".

Métriques	Gamma	Inv. Gauss	Log-norm	Box-cox
Gini Train K-Fold	20,71%	20,90%	19,76%	19,55%
Gini Test K-Fold	14,16%	14,76%	14,17%	14,57%
RMSE Train K-Fold	1058,85	1107,34	1113,62	1110,65
RMSE Test K -Fold	1189,65	1236,97	1203,93	1213,18
MAE Train K-Fold	627,09	642,51	584,44	580,59
MAE Test K-Fold	678,34	695,12	631,13	633,59
Gamma deviance Train -Fold	0,4974	0,5015	0,5835	0,5760
Gamma deviance Test K-Fold	0,5848	0,5839	0,6656	0,6477
Erreur globale Train K-Fold	0,13%	-1,47%	23,48%	21,86%
Erreur globale Test K-Fold	-1,30%	-2,96%	22,43%	20,72%

TABLE 3.4.2 – Performances sans sélection de variable pour le coût moyen sur "HZ DOM"

Les résultats sont ici particulièrement intéressant. En effet, l'indice de Gini le plus élevé (14,76%) est celui de la distribution inverse gaussienne qui ajustait moins bien la distribution des données, elle offre cependant la RMSE la plus élevée (à 1 237). La loi de gamma a l'indice de Gini le moins élevé mais la meilleure déviance moyenne (la plus faible à 0,5848). Enfin bien que les lois log-normale et Box Cox étaient les plus adaptées pour ajuster la loi de nos données (d'après les déviances globales), les erreurs globales associés sont particulièrement importantes à environ 20 % d'écart là où les erreurs globales des lois gamma et inverse gaussienne sont bien plus faibles. Seule la MAE est favorable aux distribution log-normale et Box-Cox. Au final, il est à cette étape relativement complexe de déterminer quelle est la loi la plus adaptée à nos données.

— **Modèles avec sélection de variable :**

La méthode de sélection de variable utilisée pour la modélisation du coût moyen est l'approche par réduction de GAIC backward. De la même manière que pour la fréquence, cette méthode est appliquée sur chacun des 4 folds et les variables les plus représentées seront retenues. Après sélection des variables avec cette approche, les modèles proposent les déviances globales suivantes :

Distribution	Test. GDEV
Gamma	25 010
Inverse Gaussienne	25 115
Log-normale	24 916
Box-Cox	24 846

TABLE 3.4.3 – Déviance globale test avec sélection de variable backward et k-folds pour le coût moyen sur "HZ DOM"

Les déviances globales sont ici légèrement améliorées. Il faut à présenter regarder les performances de ces modèles :

Métriques	Gamma	Inv. Gauss	Log-norm	Box-cox
Gini Train K-Fold	18,26%	19,12%	18,80%	16,25%
Gini Test K-Fold	15,49%	15,91%	15,26%	13,92%
RMSE Train K-Fold	1127,10	1129,61	1164,51	1188,35
RMSE Test K -Fold	1161,47	1166,50	1190,00	1199,78
MAE Train K-Fold	651,80	653,03	596,58	603,65
MAE Test K-Fold	669,41	673,84	616,95	617,51
Gamma deviance Train -Fold	0,5358	0,5276	0,6102	0,6458
Gamma deviance Test K-Fold	0,5697	0,5678	0,6611	0,6752
Erreur globale Train K-Fold	-0,07%	-0,52%	24,57%	21,85%
Erreur globale Test K-Fold	-0,41%	-0,69%	24,39%	24,83%

TABLE 3.4.4 – Performances avec sélection de variable backward et k-folds pour le coût moyen sur "HZ DOM"

La loi de Box Cox, bien qu'ajustant le mieux la distribution de nos données (du point de vue de la déviance globale), a à présent l'indice de Gini le plus faible à 13,92% et l'erreur globale la plus élevée à 24,83 %. Les gains de Gini sur les autres distributions sont en revanche plus importants, la distribution inverse gaussienne permettant maintenant d'obtenir un indice de Gini de 15,91%.

Il est clair ici que la méthode de sélection de variable utilisée jusqu'à présent ne permet pas de mettre en valeur la distribution Box-Cox (elle perd beaucoup de son pouvoir explicatif). En effet, en sélectionnant les données avec des k-folds, on réduit la quantité de données sur lesquelles les modèles apprennent. D'autant plus que la base de données coût moyen "HZ DOM" ne contient qu'environ 1 500 lignes. Pour pallier à ce problème, il est décidé d'appliquer l'approche backward à toute la base de donnée coût moyen "HZ DOM" dans le but d'obtenir un maximum de données.

Les déviances globales suivantes sont ainsi obtenues avec cette méthode :

Distribution	Test. GDEV
Gamma	24 992
Inverse Gaussienne	25 105
Log-normale	24 898
Box-Cox	24 811

TABLE 3.4.5 – Déviance globale test avec sélection de variables backward sans k-folds pour le coût moyen sur "HZ DOM"

Par rapport à la sélection de variables backward avec k-folds, les déviances globales globales sont davantages réduites ce qui indique que cette sélection de variables permet de mieux ajuster la distribution de la variable réponse. Il faut maintenant vérifier les performances de ces modèles :

Métriques	Gamma	Inv. Gauss	Log-norm	Box-cox
Gini Train K-Fold	19,80%	19,00%	18,08%	18,72%
Gini Test K-Fold	15,98%	16,15%	15,16%	16,51%
RMSE Train K-Fold	1072,33	1127,72	1140,32	1126,61
RMSE Test K -Fold	1198,89	1238,09	1189,48	1220,82
MAE Train K-Fold	632,81	651,64	595,26	590,63
MAE Test K-Fold	672,53	680,81	620,31	625,65
Gamma deviance Train -Fold	0,5091	0,5224	0,6070	0,5903
Gamma deviance Test K-Fold	0,5615	0,5593	0,6497	0,6230
Erreur globale Train K-Fold	0,05%	-1,03%	24,12%	22,41%
Erreur globale Test K-Fold	-1,55%	-2,32%	23,35%	20,82%

TABLE 3.4.6 – Performances avec sélection de variable backward sans k-folds pour le coût moyen sur "HZ DOM"

Notre intuition était la bonne, cette sélection de variables permet de grandement améliorer l'indice de Gini pour la distribution Box-Cox avec un Gini de 16,51%. Et, tous les indices de Gini sauf celui de la loi log-normale ont été améliorés par rapport à la sélection de variable backward avec k-folds.

Ainsi, les variables suivantes ont été sélectionnées à cette étape : "Année", "Etudiant", "Option Nouvelles Energies", "Option Piscine", "Segment", "Nombre de pièces", "Surface des dépendances" et "Capital objets de valeur".

— **Modèles avec lissage et ajustement des autres paramètres :**

De la même manière que pour la modélisation de la fréquence, des fonctions de lissage seront appliquées à certaines des variables (sélectionnées manuellement) afin de réduire les différences entre les coefficients de chacune de leurs modalités. L'ajustement des autres paramètres σ, ν et τ sera réalisé afin d'améliorer davantage le modèle :

- La fonction de lissage utilisée est $random()$ qui, pour rappel, force les coefficients à se positionner autour de leur moyenne. La fonction sera appliquée "Année".

- Seul le paramètre σ sera ajusté étant l'unique paramètre à apporter un gain de performance à nos modèles. Ce paramètre est ajusté en fonction des variables "Année" et "Capital objets de valeurs".

Au vu des performances de la loi log-normale à l'étape, l'ajustement de cette distribution a été écarté. Voici les déviances globales associées à ces nouveaux modèles :

Distribution	Test. GDEV
Gamma	24 969
Inverse Gaussienne	25 099
Box-Cox	24 795

TABLE 3.4.7 – Déviance globale test avec lissage et ajustement de σ pour le coût moyen sur "HZ DOM"

Là aussi, les déviances globales ont été légèrement améliorées et la loi Box-Cox est toujours celle ajustant le mieux la distribution de nos données.

Voici les performances des nouveaux modèles :

Métriques	Gamma	Inv. Gauss	Box Cox
Gini Train K-Fold	18,98%	19,16%	18,78%
Gini Test K-Fold	16,51%	16,57%	17,15%
RMSE Train K-Fold	1082,07	1095,00	1133,12
RMSE Test K -Fold	1201,52	1205,54	1187,73
MAE Train K-Fold	635,86	639,25	592,50
MAE Test K-Fold	670,45	672,41	619,63
Gamma deviance Train -Fold	0,5195	0,5191	0,5920
Gamma deviance Test K-Fold	0,5543	0,5539	0,6164
Erreur globale Train K-Fold	0,29%	0,02%	22,48%
Erreur globale Test K-Fold	-1,45%	-1,62%	21,40%

TABLE 3.4.8 – Performances avec lissage et ajustement de σ pour le coût moyen sur "HZ DOM"

Les tendances sont conservées, la loi de Box-Cox offre toujours l'indice de Gini le plus important à 17,15% ainsi que la RMSE la plus faible, là où l'indice de Gini de la loi gamma (16,51 %) devient quasiment égal à celui de la loi inverse gaussienne. En revanche l'erreur globale est largement en défaveur de la loi de Box Cox. C'est pourquoi le choix du modèle final reste difficile et des modèles finaux basés sur ces trois distributions seront implémentés afin de poursuivre les comparaisons.

— **Application des modèles finaux à la base de données "Z DOM"**

Les modèles les plus aboutis ont été construits d'étape en étape grâce à la validation croisée. Les modèles finaux seront eux implémentés sur toute la base coût moyen "HZ DOM".

En résumé, les modèles finaux ont les caractéristiques suivantes :

- les données sont ajustées grâce aux lois gamma, inverse gaussienne et Box-Cox.

- l fonction de lien utilisée dans les 3 cas est le *log*.

- les termes additifs sont les variables suivantes : "Année", "Etudiant", "Option Nouvelles Energies", "Option Piscine", "Segment", "Nombre de pièces", "Surface des dépendances" et "Capital objets de valeur".

- la fonction de lissage *random*() est appliquée sur la variable "Année".

- le paramètre σ est ajusté avec les variables "Année" et "Capital objets de valeurs".

Ainsi, les performance sur les bases coût moyen "HZ DOM" et "Z DOM" pour nos 3 distributions sont les suivantes :

Métriques	Gamma	Inv. Gauss	Box Cox
Gini HZ DOM	18,76%	18,95%	18,65%
Gini Z DOM	18,06%	18,02%	17,65%
RMSE HZ DOM	1088,76	1095,64	1131,88
RMSE Z DOM	1211,69	1210,03	1248,31
MAE HZ DOM	637,64	639,51	593,38
MAE Z DOM	661,29	661,47	620,16
Poisson deviance HZ DOM	0,5229	0,5223	0,5894
Poisson deviance Z DOM	0,5627	0,5623	0,6422
Erreur globale HZ DOM	0,30%	0,23%	21,85%
Erreur globale Z DOM	2,24%	2,24%	23,30%

TABLE 3.4.9 – Performances des modèles finaux coût moyen hors zonier DOM

Là aussi les résultats sont particulièrement intéressant. La distribution Box-Cox qui ajustait au mieux la distribution de nos données et qui offrait les meilleures performances offre à présent les performances les plus faibles d'un point de vue prédictif avec la RMSE la plus élevée et un indice de Gini de 17,65 % sur la base "Z DOM" là où celui de la distribution gamma est de 18,06%. De plus le problème de l'erreur globale très élevée avec la distribution Box-Cox est toujours présent.

De ce fait, il peut être pertinent d'analyser en détail la distribution des résidus normalisés (résidus obtenus de la base "Z DOM") entre les lois gamma et Box-Cox grâce à un graphique des densités, d'un QQ-plot et d'un worm plot :

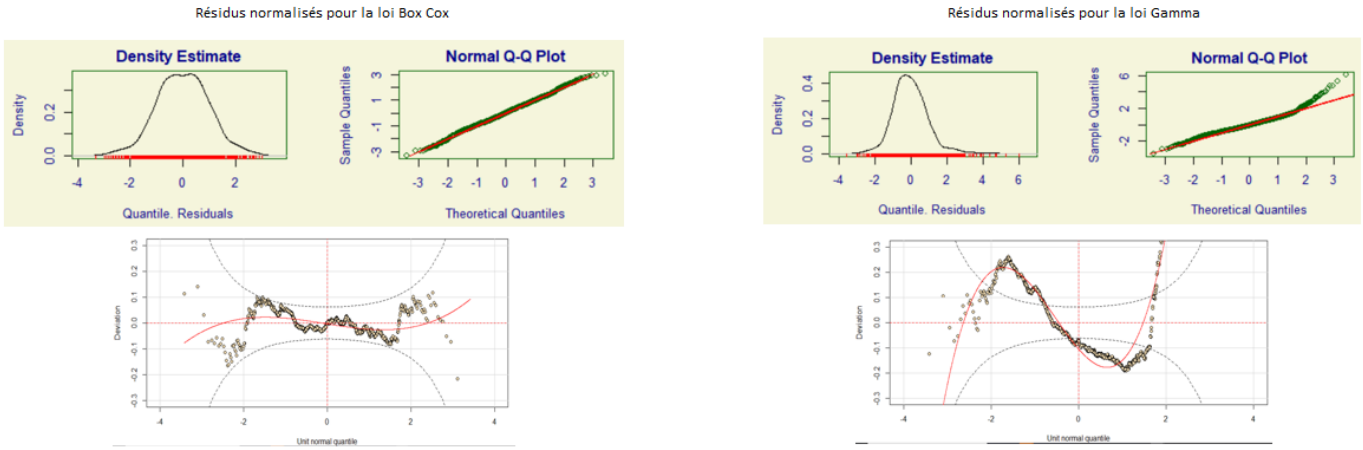


FIGURE 3.4.1 – Analyse des résidus normalisés des lois Box Cox et gamma pour le coût moyen DOM

Les résidus normalisés sont plus ajustés à une loi normale dans le cas de la distribution Box-Cox où le QQ-plot indique une très bonne adéquation et le worm plot étant entre les deux courbes elliptiques. Cela n'est pas du tout le cas le la loi Gamma. Il y a donc ici une situation étonnante, où le meilleur modèle n'est pas celui ajustant le mieux la distribution des données. Dans notre étude nous somme davantage interessé par le pouvoir prédictif, c'est pour le modèle utilisant la distribution de gamma est le plus adapté. Nous allons cependant poursuivre les comparaisons entre la loi gamma et la loi Box-Cox dans la suite afin de pouvoir établir un verdict final.

3.4.2 Comparaison avec la modélisation du coût moyen "métropole+DOM"

Dans cette section le modèle de coût retenu pour le périmètre "métropole+DOM" sera exposé sans détailler toutes les étapes comme précédemment, la démarche restant la même. Les performances seront ensuite comparées avec celles du modèle retenu sur le périmètre "DOM".

— Le modèle de coût moyen retenu sur le périmètre "MET+DOM"

Le modèle retenu pour la modélisation du coût moyen sur le périmètre "métropole+DOM" possède les caractéristiques suivante :

- les données ont été ajustées avec la distribution gamma (une comparaison avec la loi Beta généralisée sera proposée)

- la fonction de lien utilisé est le *log*

- les termes additifs sont les variables suivantes : "Nombre de pièces", "Segment", "Capital objets communs", "Capital objets de valeur", "Réseau de distribution", "Type d'occupation", "Année", "Age du client", "Ancienneté du contrat", "Ancienneté du logement" et "Etudiant".

Ce modèle a obtenu les performances suivantes sur les bases coût moyen "HZ métropole+DOM" et "Z métropole + DOM" :

Métriques	Gamma	Beta gen.
Gini HZ MET+DOM	16,21%	16,10%
Gini Z MET+DOM	16,20%	16,11%
RMSE HZ MET +DOM	1591,75	1653,43
RMSE Z MET + DOM	1593,18	1654,50
MAE HZ MET +DOM	933,08	852,92
MAE Z MET + DOM	935,14	854,11
Gamma deviance HZ MET +DOM	0,70730	0,82220
Gamma deviance Z MET +DOM	0,70740	0,82180
Erreur globale HZ MET+DOM	0,01%	27,91%
Erreur globale Z MET+DOM	-0,04%	27,88%

TABLE 3.4.10 – Performances des modèles finaux coût moyen métropole+DOM

La distribution Beta généralisée est dans le même cas que la distribution Box-Cox sur le périmètre "DOM" au sens où elle ajuste mieux la distribution des données mais a un moins bon pouvoir prédictif global.

Bien que les métriques ne soit pas comparables de manière absolue, le modèle de coût moyen "DOM" segmente mieux les données "DOM" que le modèle de coût moyen "métropole+DOM" sur les données "métropole+DOM" en regardant les indices de Gini. Le modèle de coût "métropole+DOM" sur-apprend en revanche moins. Cela s'explique notamment par le fait que le segment "métropole+DOM" a beaucoup plus de données. Pour pouvoir comparer les deux modèles, on doit regarder leur performance sur la même base de données.

— **Comparaison des performances des deux modèles**

Les modèles qui vont être comparés sont ici ceux issus de la distribution gamma, qui offre les meilleures performances sur les deux segments.

Les performances des modèles coût moyen DOM et métropole + DOM sur la base coût moyen "Z DOM" sont résumées dans le tableau suivant :

Indicateurs	Modèle coût moyen DOM	Modèle coût moyen métropole + DOM
Nb variables	8	11
Gini	18,06%	15,96%
RMSE	1211,69	1219,88
MAE	661,29	667,67
Avg Deviance	0,56270	0,58540
Erreur globale	2,24%	1,72%

TABLE 3.4.11 – Comparaison des modèles coût moyen hors zonier sur la base "Z DOM"

Le modèle basé sur le périmètre "DOM" offre de meilleures performances avec un indice de Gini supérieur de 2 points (18,06% contre 15,96%). De même toutes les métriques sauf l'erreur globale (qui reste proche dans les deux cas), sont en faveur du modèle coût moyen DOM.

— **Analyse des écarts**

Nous tenterons d'expliquer ces écarts de performances avec les spreads de nos variables.

Les deux modèles fournissent les spreads suivant :

Variables	Modèle HZ DOM	Variables	Modèle HZ MET+DOM
Surface des dépendances	557,87%	Nombre de pièces	103,25%
Capital objets valeur	509,52%	Segment	45,64%
Option nouvelle énergie	81,40%	Capital objets valeur	34,25%
Nombre de pièces	79,40%	Réseau de distribution	30,11%
Segment	36,98%	Capital objets commun	27,46%
Etudiant	32,19%	Occupation du logement	22,92%
Année	21,18%	Année	21,10%
Option piscine	19,98%	Age client	15,27%
-	-	Ancienneté contrat	6,58%
-	-	Etudiant	5,72%
-	-	Ancienneté logement	5,31%

TABLE 3.4.12 – Spread des variables hors zonier des modèles coût moyen

Le modèle coût moyen DOM segmente davantage les variables "Surface des dépendances" et "Capital objets de valeurs" là où le modèle coût moyen métropole + DOM discrimine davantage la variable "Nombre de pièces"

Les écarts de spread les plus importants sont sur les variables :

- "Capital objets de valeur" avec un écart de plus de 450 points. (Bien que les écart de spread, en omettant les valeurs extrêmes, soient presque les mêmes)

- "Nombre de pièces" avec un écart de spread d'environ 70 points.

- "Etudiant" avec un écart d'environ 30 points.

- "Segment" avec un écart d'environ 10 points.

De la même manière que pour la fréquence, traçons les coefficients des modalités des variables "Nombre de pièces" et "Segment" afin de mieux comprendre ces écarts :

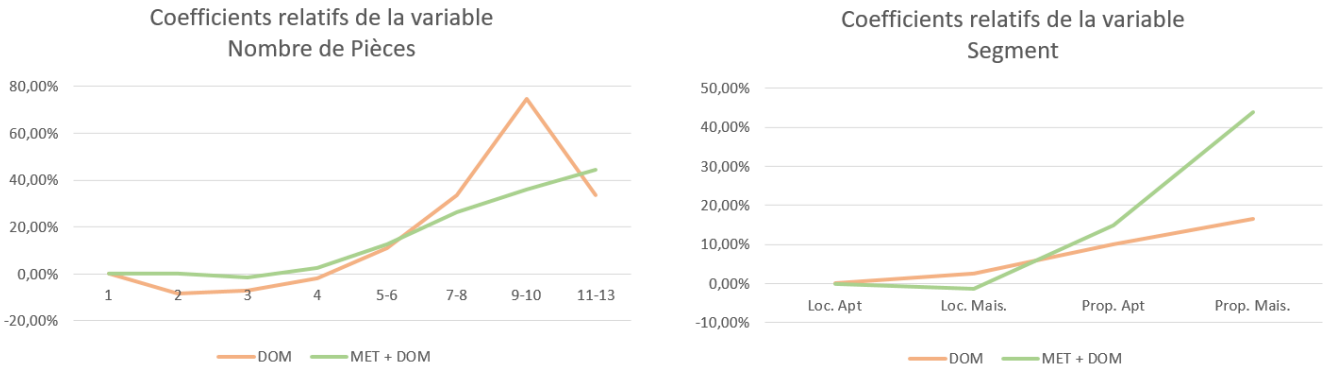


FIGURE 3.4.2 – Comparaison des coefficients relatifs pour les modèles coût moyen

Le modèle coût DOM discrimine davantage, à la hausse, les logements de 9 à 10 pièces et légèrement à la baisse les logement de moins de 5 pièces par rapport au modèle coût moyen métropole + DOM. Concernant la variable "Segment" le modèle coût moyen DOM discrimine à la baisse les propriétaires par rapport au modèle de coût moyen.

Comparativement aux modèles de fréquence les écarts de spread sont globalement moins important dans les modèles de coût moyen ce qui peut notamment expliquer le plus faible écart de l'indice de Gini (2 points d'écart pour le coût contre 11 points pour la fréquence).

Chapitre 4

Modélisation de la sinistralité finale

4.1. La modélisation géographique

4.1.1 Méthodologie de modélisation avec zoniers

Dans cette section sera détaillée la modélisation des variables réponses avec zonier.

Jusque là, les modèles ont été réalisés en tenant compte des caractéristiques liées aux assurés (tels que son âge et l'ancienneté de son contrat) et à leur bien (tels que le nombre de pièces et les capitaux mobiliers). Nos modèles ne tenaient donc pas compte de la localisation du risque. Or, il s'agit, à priori d'une variable pouvant expliquer une part ce risque.

— **Méthodologie**

Ainsi, une nouvelle variable géographique va être ajoutée à nos modèle. Cette variable correspondra au risque associé à une certaine zone géographique. La méthode de modélisation suivante est ainsi proposée (*Perrin, (2021),* **[9]**) :

- Les résidus sont récupérés sur la prédiction des variables réponses sur les bases "Z DOM" et "Z Métropole + DOM". (1)

Ce résidu est défini comme :

$$\textbf{residu}_{\textbf{modele hors zonier}} = \textbf{valeur reelle} - \textbf{valeur predite}_{\textbf{modele hors zonier}}$$

Ce résidu contient donc une part de signal géographique de telle sorte que :

$$\textbf{residu}_{\textbf{modele hors zonier}} = \textbf{valeur reelle}_{\textbf{geographique}} - \textbf{residu}_{\textbf{modele zonier}}$$

en décomposant $valeur\ reelle = valeur\ reelle_{non\ geographique} + valeur\ reelle_{geographique}$

- Les résidus sont ensuite agrégés à une certaine maille géographique (nous détaillerons ces différentes maille gégraphique plus loin). (2)

- Des zones sont déterminées en découpant les résidus en zones géographiques (3).

- Les résidus sont lissés afin de gommer les irrégularités (4).
Une nouvelle variable associant une valeur de risque à chaque zone est ainsi obtenue.

- Cette nouvelle variable est intégrée en tant que variable explicative à nos modèles. Un coefficient sera ainsi associé à chaque zone géographique. (5)

Les étapes (1) et (2) seront realisées grâce à l'algorithme Georev que nous présenterons dans la prochaine section.

La modélisation des variables antécédents interviendra ensuite selon le même procédé : les résidus seront expliqués en fonction des nouvelles variables. La modélisation des antécédents intervient à cette étape car ceux-ci sont en réalité définis à la maille contrat (les antécédents sont agrégés par contrat, voir section 2.4.1), or les déménagements étant généralement assez peu nombreux, il est cohérent de considérer que ces antécédents sont en fait à la maille adresse. Ainsi, les antécédents contiennent une part d'information géographique et modéliser ces variables en dernier permettra de mettre l'accent sur l'importance de la variable "Zonier" (qui serait captée par les variables "Antécédents" autrement).

Le résumé de la modélisation des variables géographiques et des antécédents est disponible en section 3.1.1 (figure 29 en particulier).

— **Les différentes maille géographiques**

Il existe de multiples mailles géographiques. Dans le cadre de la tarification au sein de l'équipe MRH d'AXA France, 3 d'entre elles sont utilisées. Présentons les par granularité croissante :

- La maille *INSEE* : il s'agit du découpage communal classique utilisé notamment pour répertorier les différentes communes de France. Il en existe actuellement 37 026 dont 112 dans les départements ultramarins de Guadeloupe, Martinique Guyane et Réunion.

- la maille *IRIS* : c'est une extension du découpage INSEE permettant un découpage infra-communal pour certaine commune. La France comporte 50 800 IRIS dont 700 dans les territoires ultramarins. Seules les communes de plus de 10 000 habitants et la plupart des communes de 5 000 à 10 000 habitants sont découpées. Il reste ainsi 34 800 IRIS : conserve en fait le découpage à la maille INSEE.

- la maille *Voronoï* : il s'agit d'un découpage encore plus fin que la maille IRIS. Les adresses des contrats situés en portefeuille sont découpées en polygones : les points contenus dans les polygones sont plus proche de cette adresses que n'importe quel autre.

Dans le cadre de notre étude, seule la maille INSEE servira à la construction des zoniers. Une modélisation plus fine avec la maille IRIS poserait problème au regard de la faible quantité disponible.

4.1.2 Construction du zonier avec Georev

Dans cette section nous allons détailler brièvement le fonctionnement de l'algorithme Georev (*Toesca, (2020),* **[10]**).

La construction du zonier avec l'algorithme Georev se fait de la manière suivante :

- la première étape est l'agrégation des résidus à la maille INSEE. Chaque code INSEE sera donc associé à une valeur de résidu moyen. Les résidus moyen sont ensuite quantilisés en N quantiles en fonction de l'exposition. Et un coefficient initial sera attribué aux INSEE en fonction du quantile dans lequel ils sont. Le nombre de zone de risque (quantiles) N est un des paramètre du modèle.

- la seconde étape consiste en le lissage de la carte. L'objectif est d'assigner une zone à chaque INSEE (même ceux sans contrats) et de lisser la carte obtenue afin d'éviter le sur-apprentissage. Le coefficient d'un INSEE sera calculé grâce à coefficients mêlant les notions de crédibilité (on va croire les INSEE ayant le plus d'exposition) et de similarité (on va croire les zones voisines ayant une exposition similaire). Le coefficient associé à un INSEE i se calcule ainsi :

$$\mathbf{c_{final,i}} = \mathbf{c_{r,i}} \times \mathbf{c_{initial,i}} + (\mathbf{1} - \mathbf{c_{r,i}}) \times \frac{\sum_{\mathbf{v=1}}^{\mathbf{n}} \mathbf{f(v)}.\mathbf{c_{initial,v}}}{\sum_{\mathbf{v=1}}^{\mathbf{n}} \mathbf{f(v)}}$$

où :

- $c_{r,i}$ est le coefficient de crédibilité associé à l'INSEE i.

- $c_{initial,i}$ le coefficient initial attribué à l'INSEE i après la quantilisation.

- f une fonction de pondération permettant d'expliquer le coefficient de l'INSEE i par celui de ses proches voisins qui lui ressemblent en terme d'exposition.

Le nombre de plus proches voisins (la notion de proche voisin étant définie par distance euclidienne ou de Harvesine), doit respecter deux contraintes. On souhaite que le nombre de voisins soit :

$$\mathbf{n} = \mathbf{max}(\mathbf{n_{min}}, \mathbf{n_e})$$

où n_{min} est le nombre minimal de voisin que l'on souhaite considéré et n_e est le nombre minimal de voisins permettant de considérer une exposition de e. n et e sont des paramètres d'entrée de l'algorithme.

Le coefficient de crédibilité se calcule par défaut comme :

$$\mathbf{c_{r,i}} = \min\left(\mathbf{1}\ ,\ \frac{\mathbf{exposition_i^3}}{\mathbf{quantile^3_{exposition,99,5\%}}}\right)$$

Ainsi, on croira totalement (coefficient de crédibilité de 1) les INSEE ayant une exposition supérieure ou égale au quantile 99,5% de l'exposition. Et le coefficicient croît rapidement avec l'exposition du fait du polynôme de degré 3.

La fonction de pondération f se décompose en réalité en deux parties. En effet un coefficient semblable doit être assigné à deux variables "similaire". Cette notion de similarité se décompose en deux aspects : similaire du point de vue de la distance et similaire du point de vue de l'exposition. Ainsi :

$$\mathbf{f(v)} = \mathbf{f_d(v)} \times \mathbf{f_e(v)}$$

f_d est de la forme suivante avec x la distance :

$$\mathbf{f_d(x)} = \mathbf{min}\left(\mathbf{1}\ ,\ \frac{\mathbf{1}}{\sqrt[p]{x}}\right)$$

On prendra par défaut $p = 3$, ainsi la fonction tend vers rapidement vers 0 quand la distance augmente.

Et, f_e est de la forme suivante avec x l'exposition :

$$\mathbf{f_e(x)} = \mathbf{min}\left(\mathbf{1}\ ,\ \frac{\mathbf{1}}{\sqrt[3]{|x-m|+1}}\right)$$

où m est l'exposition de l'INSEE considéré. Il s'agit donc d'une fonction symétrique en m et décroissante à mesure que l'on s'éloigne de ce point.

Avec ces fonctions, il est désormais possible de calculer $c_{final,i}$ pour chaque INSEE.

- La dernière étape consiste donc à quantiliser les $c_{final,i}$ en fonction de l'exposition.
Au final une zone de risque de 1 à N est obtenue.

4.2. La modélisation de la fréquence avec zonier

Dans cette partie sera étudiée la modélisation de la fréquence en intégrant une variable géographique : la note du zonier Georev intitulée "Zonier".

4.2.1 Modélisation du zonier fréquence sur le périmètre "DOM"

— **Modélisation des variables zoniers**

Lors du chapitre précédent, un modèle de type GAMLSS utilisant entre autres la loi de Sichel a été obtenu sur la base fréquence "HZ DOM" et appliqué à la base fréquence "Z DOM". Cette loi sera encore utilisée pour modéliser les variables géographiques.

Les résidus obtenus vont ainsi nous permettre de contruire la variable "Zonier". L'objectif est en réalité de construire plusieurs variables géographiques basées sur des paramètres différents de l'algorithme Georev. 4 combinaisons vont être réalisées sur le périmètre "DOM" avec les paramètres suivant :

- Une exposition minimale de 5 000, un nombre de voisin minimal de 10 et 10 zones de risques, qui vont permettre de construire la variable "Zonier 1".

- Une exposition minimale de 3 000, un nombre de voisin minimal de 10 et 10 zones de risques, qui vont permettre de construire la variable "Zonier 2".

- Une exposition minimale de 5 000, un nombre de voisin minimal de 20 et 10 zones de risques, qui vont permettre de construire la variable "Zonier 3".

- Une exposition minimale de 5 000, un nombre de voisin minimal de 10 et 20 zones de risques, qui vont permettre de construire la variable "Zonier 4".

En effet, l'algorithme Georev étant relativement récent, il n'y a pas d'historique permettant de connaître les valeurs des paramètres à utiliser.

De la même manière que pour la modélisation de la fréquence hors zonier, les modèles vont être validé par validation croisée avec 4-folds.

La subtilité de réaliser un modèle en deux étapes (modélisation hors zonier puis modélisation avec zonier), est de devoir modéliser notre variable réponse en intégrant un offset ω (ou décalage) correspondant à la prédiction du modèle hors zonier sélectionné de manière à modéliser uniquement le résidu du modèle précédent. La modélisation avec zonier de μ devient avec l'offset (pour une fonction de lien log) :

$$\mu(\mathbf{x}) = \mathbf{exp}(\mathbf{x}\beta_{\mathbf{zonier}}).\mathbf{exp}(\omega)$$

De ce fait, les coefficients associés aux variables "hors zonier" du modèle avec zonier seront les mêmes que ceux du modèle hors zonier.

Ci-dessous le tableau résumant les performances des modèles pour les différentes variables zonier :

Métriques	Zonier 1	Zonier 2	Zonier 3	Zonier 4
Gini Train K-Fold	42,92%	42,85%	42,83%	43,43%
Gini Test K-Fold	42,52%	42,38%	42,22%	42,49%
RMSE Train K-Fold	0,1396	0,1396	0,1396	0,1396
RMSE Test K -Fold	0,1359	0,1359	0,1359	0,1359
MAE Train K-Fold	0,0224	0,0224	0,0224	0,0224
MAE Test K-Fold	0,0224	0,0224	0,0224	0,0224
Poisson deviance Train -Fold	0,09863	0,09865	0,09865	0,09848
Poisson deviance Test K-Fold	0,09880	0,09882	0,09890	0,09878
Erreur globale Train K-Fold	-1,9687%	-1,8865%	-1,8000%	-1,8850%
Erreur globale Test K-Fold	-1,9979%	-1,9601%	-1,8910%	-1,9469%

TABLE 4.2.1 – Performances avec variables zoniers pour la fréquence sur "Z DOM"

Les 4 variables offrent presque les mêmes performances. Le Gini le plus élevé a été obtenu avec la variable "Zonier 1" (Gini de 42,52 % sur le "test k-folds"), suivi de près de la variable "Zonier 4" (Gini de 42,52 % sur le "test k-folds") qui offre en plus la déviance la plus faible. Cependant , la variable "Zonier" a un écart de plus d'1 point de Gini entre le "train k-folds" et "test-folds" là où l'écart n'est que de 0,4 point pour la variable "Zonier 1". Ainsi, c'est cette dernière qui va être sélectionnée.

Cependant, le spread de 404% associé à cette variable "Zonier 1" est très élevé. Cela signifie qu'au sein des DOM, le tarif peut être multiplié par 5 selon l'endroit où le risque est localisé. Cette différence n'est pas acceptable d'un point de vue commercial. C'est pourquoi on va utiliser une pénalisation *Ridge*. Il s'agit d'une pénalisation utilisée dans les modèles linéaire ressemblant fortement au LASSO vu en section 3.3.1, mais à la différence de celui-ci, la pénalisation Ridge va utiliser les norme L_2 pour contraindre les coefficients. Ainsi, là où la pénalisation LASSO va forcer les coefficients des variables ne servant pas à expliquer le modèle à être nuls, la pénalisation Ridge va réduire l'amplitude des coefficients des variables présentes dans le modèle. Les paramètres d'une régression Ridge se calculent de la façon suivante :

$$\beta_{\mathbf{Ridge}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \beta' \mathbf{X_i})^{\mathbf{2}} + \lambda \sum_{\mathbf{j=1}}^{\mathbf{p}} (\beta_{\mathbf{j}})^{\mathbf{2}} \quad \text{avec } \lambda \in \mathbb{R}^+$$

Avec l'introduction de la pénalisation Ridge pour la variable "Zonier 1", les performances deviennent alors :

Métriques	Zonier 1 sans Ridge	Zonier 1 avec Ridge
Gini Train K-Fold	42,92%	42,90%
Gini Test K-Fold	42,52%	42,93%
RMSE Train K-Fold	0,1396	0,1396
RMSE Test K -Fold	0,1359	0,1359
MAE Train K-Fold	0,0224	0,0223
MAE Test K-Fold	0,0224	0,0222
Poisson deviance Train -Fold	0,09863	0,09895
Poisson deviance Test K-Fold	0,09880	0,09852
Erreur globale Train K-Fold	-1,9687%	-0,4987%
Erreur globale Test K-Fold	-1,9979%	-0,5915%

TABLE 4.2.2 – Performances avec variables zoniers et Ridge pour la fréquence sur "Z DOM"

Du point de vue des performances, la régression Ridge permet d'améliorer le coefficient de Gini de 0,4 points et de réduire d'1 point l'erreur globale. Et concernant la problématique du spread de la variable "Zonier 1", celui-ci passe de 404% à 60 % ce qui est bien plus cohérent d'un point de vue commercial.

— **Modélisation des variables antécédents**

Maintenant que la variable "Zonier 1" a été intégrée au modèle, il nous faut à présent intégrer les variables antécédents : "Nombre de sinistre sur 1 an", "Nombre de sinistre sur 3 ans", "Nombre de sinistre sur 5 ans", "Nombre de sinistre sur 10 ans", "Charge sinistre sur 1 an", "Charge sinistre sur 3 ans", "Charge sinistre sur 5 ans", "Charge sinistre sur 10 ans".

Ces 8 variables étant fortement corrélées, l'intégration de chacune d'entre elles dans notre modèle va rendre les coefficients associés "illisible". Ainsi, il a été décidé d'intégrer uniquement la variable antécédent apportant le maximum de performance à notre modèle. Il s'agit dans notre cas de la variable "Nombre de sinistre sur 10 ans".

Pour intégrer cette variable, de la même manière que pour l'intégration des variables zoniers, il faut utiliser les prédictions obtenues avec le modèle fréquence intégrant la variable "Zonier 1" (le modèle précédent) sur la base "Z DOM". Ces prédictions seront donc intégrées en tant que variable *offset* dans notre modèle.

Les performances de notre modèle en intégrant cette nouvelle variable sont les suivante :

Métriques	Zonier 1 + N. sinistre 10 ans
Gini Train K-Fold	45,88%
Gini Test K-Fold	45,74%
RMSE Train K-Fold	0,1396
RMSE Test K -Fold	0,1359
MAE Train K-Fold	0,0224
MAE Test K-Fold	0,0224
c Poisson deviance Train -Fold	0,09720
Poisson deviance Test K-Fold	0,09743
Erreur globale Train K-Fold	-3,4271%
Erreur globale Test K-Fold	-3,5093%

TABLE 4.2.3 – Performances avec variables zoniers et antécédents pour la fréquence sur "Z DOM"

L'ajout de la variable "Nombre de sinistre sur 10 ans" permet d'améliorer significativement les performances de notre modèle. En effet, le modèle a gagné près de 3 points de Gini (à 45,74%) au prix d'une détoriation de 3 points de l'erreur globale. Enfin, la différence des indicateurs entre le "test k-folds" et le "train k-folds" toujours faible indique que le modèle a conservé sa capacité à ne pas sur-apprendre.

— **Application du modèle à la base "V DOM"**

Il est maintenant impératif de valider notre modèle final grâce à la base fréquence "V DOM". Et voici les performances sur cette base comparée à celles sur la base "Z DOM" :

Métriques	Modèle final fréquence sur le périmètre "DOM"
Gini Z DOM	45,93%
Gini V DOM	42,36%
RMSE Z DOM	0,1402
RMSE V DOM	0,1301
MAE Z DOM	0,0224
MAE V DOM	0,0230
Poisson deviance Z DOM	0,09730
Poisson deviance V DOM	0,10260
Erreur globale Z DOM	-3,5769%
Erreur globale V DOM	-0,5426%

TABLE 4.2.4 – Performances du modèle avec zonier final fréquence DOM

Concentrons nous tous d'abord sur les performances de notre modèle sur la base "Z DOM". Le Gini y est de 45,93% ce qui est élevé, d'autant plus qu'à partir du modèle fréquence hors zonier sur le périmètre "DOM", nous obtenions un Gini de 36,33 %. Ainsi, l'ajout successif des variables de zoniers et d'antécedents a permis d'augmenter le Gini de près de 10 points ce qui est là aussi très important. Concernant les performances sur la base fréquence "V DOM", elles sont là aussi plutôt bonnes : l'indice de Gini est élevé, à 42,36% et l'erreur globale est faible.

4.2.2 Comparaison avec le zonier fréquence "métropole+DOM"

Dans cette section , nous exposerons le modèle retenu pour la modélisation du zonier fréquence sur le périmètre "métropole+DOM" sans les détails, la démarche étant la même que pour le périmètre "DOM". Les performances seront ensuite comparées au modèle retenu pour la modélisation de la fréquence du périmètre "DOM" avec zonier.

— Le modèle de fréquence avec variable zoniers et antécédents retenu sur le périmètre "MET+DOM"

Nous rappelons que le modèle de fréquence réalisé sur le périmètre "métropole+DOM" utilise notamment la distribution de Sichel.

Les nouvelles variables intégrées à ce modèle sont les suivantes :

- la variable "Zonier" construite avec l'algorithme Georev en prenant une exposition minimale de 15 000, un nombre de voisin minimal de 100 et 20 zones de risques.

- la variable "Nombre d'antécédents sur 10 ans"

Ce modèle a obtenu les performance suivantes sur les bases fréquence "Z métropole + DOM" et "V métropole + DOM" :

Métriques	Modèle final fréquence sur le périmètre "MET+DOM"
Gini Z MET+DOM	40,69%
Gini V MET+DOM	40,88%
RMSE Z MET+DOM	0,3688
RMSE V MET+DOM	0,2031
MAE Z MET+DOM	0,0469
MAE V MET+DOM	0,0462
Poisson deviance Z MET+DOM	0,1785
Poisson deviance V MET+DOM	0,1736
Erreur globale Z MET+DOM	-2,5836%
Erreur globale V MET+DOM	-3,4405%

TABLE 4.2.5 – Performances du modèle avec zonier final fréquence métropole+DOM

Sur la base fréquence "Z métropole+DOM", l'indice de Gini est de 40,69%. Ainsi, en ajoutant les variables de zoniers et d'antécédents le Gini a gagné près de 8 points sur cette base. De plus, la valeur de

ce même indicateur sur la base fréquence "V métropole+DOM" (Gini de 40,88%) indique qu'il y a peu de sur-apprentissage.

— **Comparaison des performances et analyse**

Les performances des modèles fréquence finaux avec zonier sur la base "V DOM" sont résumées dans le tableau suivant :

Indicateurs	Modèle finale fréquence périmètre "DOM"	Modèle finale fréquence périmètre "MET+DOM"
Nb variables	16	16
Gini	42,37%	32,53%
RMSE	0,1306	0,1298
MAE	0,0231	0,0212
Abg. Deviance	0,10280	0,10470
Erreur globale	-1,892%	6,182%

TABLE 4.2.6 – Comparaison des performances des modèles fréquence avec zonier sur la base "V DOM"

Ainsi le modèle de fréquence final basé sur le périmètre "DOM" permet d'obtenir un indice de Gini supérieur de 10 points sur la base de validation fréquence "V DOM" par rapport au modèle basé sur le périmètre "métropole + DOM", ce modèle permet donc de mieux segmenter le risque des DOM. Cependant, la RMSE et la MAE sont moins élevées avec la modélisation "métropole+DOM".

Afin de comparer plus en profondeur le pouvoir prédictif de nos deux modèles, nous introduisons les graphiques *Lift Curve*. Ce graphique permet d'analyser les écarts entre les prédictions et les valeurs réelles. Pour cela, les prédictions sont triées et rangées dans 20 groupes représentant chacun 5% des prédictions. Pour chaque groupe, on affiche les prédictions moyennes et les valeurs réelles moyennes. Comparons donc les lift curves de nos deux modèles sur la base "V DOM" :

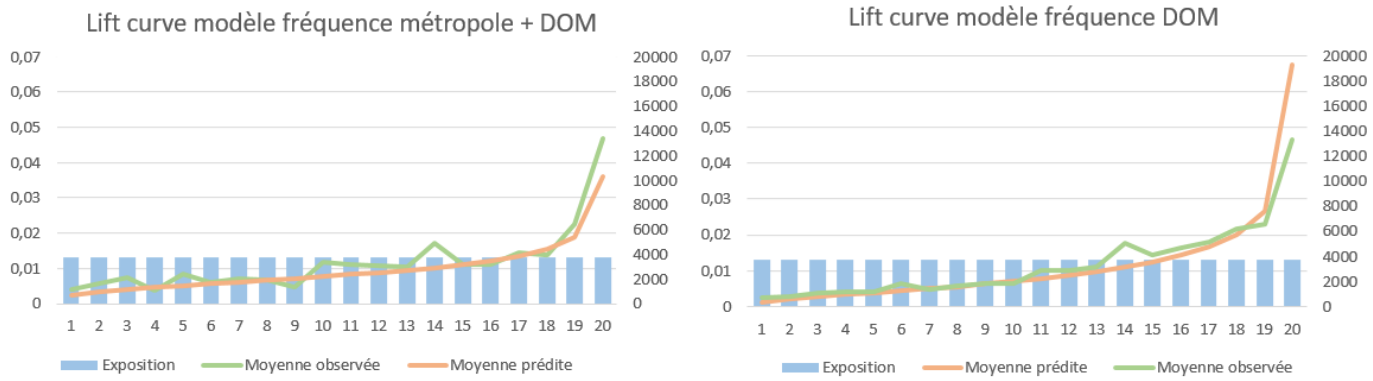


FIGURE 4.2.1 – Lift Curves des modèles finaux fréquence

En analysant ces lift curves, nous pouvons remarquer qu'il y a globalement moins d'écart entre la prédiction moyenne et l'observée moyenne, notamment sur les 10 premiers quantiles, avec le modèle DOM qu'avec le modèle "métropole+DOM". Cependant , le quantile le plus risqué est mieux prédit par le modèle "métropole+DOM".

Détaillons à présent les différences dans la modélisation de la variable "Zonier". Nous rappelons que le spread de cette variable avec le modèle fréquence "DOM" est de 60,63% alors qu'il est de 28,35% avec le modèle fréquence "métropole+DOM".

Nous allons afficher les cartes permettant de localiser les zones de risques déterminées par nos modèles puis nous allons analyser les spreads de la variable "Zonier" sur nos 4 DOM pour chacun de ceux-ci.

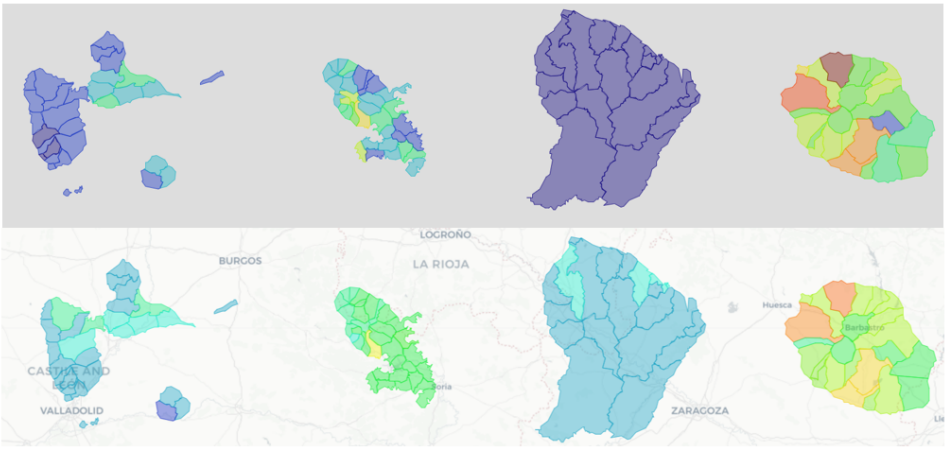


FIGURE 4.2.2 – Zones de risques déterminées par les modèles "DOM" (en haut) et "métropole+DOM"

Pour analyser ces cartes, nous avons besoin des spreads associés à chaque DOM présentés dans le tableau suivant :

DOM	Modèle MET+DOM	Modèle DOM	Exposition
Réunion	13,18%	30,43%	56,85%
Martinique	9,72%	14,54%	20,94%
Guadeloupe	4,72%	24,81%	13,83%
Guyane	1,66%	0,00%	8,38%
Tout DOM	16,12%	60,63%	100,00%

TABLE 4.2.7 – Spread de la variable "Zonier" par DOM

Ainsi, le spread de la variable "Zonier" du modèle fréquence "métropole+DOM" sur les DOM est seulement de 16,12 % alors qu'il est de 60,63% avec le modèle "DOM". Ainsi le modèle "DOM" segmente plus le risque sur ce périmètre alors que le modèle "métropole+DOM" considère le risque de ce périmètre comme étant plus uniforme. Cela explique donc pourquoi le gain de Gini lié à l'intégration de la variable géographique sur le périmètre "DOM" est bien moindre pour le modèle "métropole+DOM". En regardant en détail les spreads sur chaque DOM, on remarque que le risque prédit de La Réunion est le plus segmenté là où celui en Guyane est le même pour chaque commune. Cela s'explique notamment par le fait que La Réunion représente plus de la moitié de l'exposition alors que la Guyane n'en représente que 8%.

— **Modélisation hybride**

Avec les modélisations précédentes, nous avons vu que sur le périmètre "DOM" :

- il y a une différence de 10 points de Gini en faveur de la modélisation hors zonier basée sur le périmètre "DOM" par rapport à la modélisation hors zonier basée sur le périmètre "métropole+DOM" (voir section 3.3.2)

- l'intégration des variables zoniers permet d'apporter environ 11 points de Gini avec le modèle "DOM" contre environ 6 points avec le modèle "métropole+DOM".

Ainsi, nous décidons d'implémenter un modèle "hybride" avec les coefficients hors zonier du modèle basé sur le segment "métropole+DOM" mais dans lequel les coefficients des variables zoniers seront eux basés sur le périmètre "DOM". Cela nous permettra de voir si la réalisation d'un zonier concentré sur le périmètre "DOM" pourra compenser le fait d'utiliser des variables "hors zonier" provenant du modèle "métropole+DOM".

On résume notre méthodologie par la figure suivante (mise à jour de la figure 29 de la section 3.1.1) :

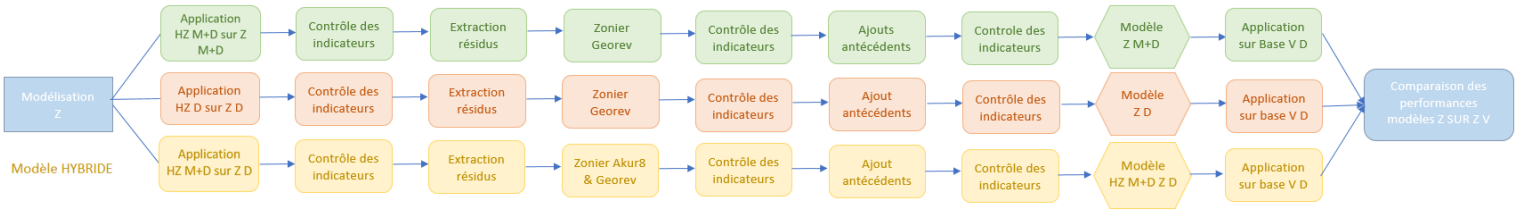


FIGURE 4.2.3 – Schéma de la modélisation avec zonier avec le modèle hybride

Nous partons donc du modèle fréquence hors zonier retenu sur le périmètre "métropole+DOM". Nous allons ensuite modéliser la variable "Zonier 1" en intégrant la prédiction sur la base fréquence "Z DOM" en tant que variable offset. Le même procédé est ensuite appliqué aux variables antécédents.

Les performances du modèle hybride ainsi obtenu sur les bases fréquence "Z DOM" et "V DOM" sont les suivantes :

Métriques	Modèle hybride
Gini Z DOM	40,32%
Gini V DOM	39,29%
RMSE Z DOM	0,1473
RMSE V DOM	0,1305
MAE Z DOM	0,0235
MAE V DOM	0,0236
Poisson deviance Z DOM	0,1039
Poisson deviance V DOM	0,1039
Erreur globale Z DOM	-4,412%
Erreur globale V DOM	-5,966%

TABLE 4.2.8 – Performance du modèle hybride fréquence sur le périmètre "DOM"

Avec le modèle hybride, sur la base fréquence "V DOM", l'indice de Gini a gagné environ 7 points sur le périmètre "DOM" (à 39,29%) par rapport au modèle base sur le périmètre "métropole+DOM" (Gini de 32,53%) et n'a plus que 3 points de différence avec le modèle basé sur le périmètre "DOM" (Gini de 42,37%).

La dernière étape consiste à vérifier que les cartes et les spreads sont globalement les mêmes entre le modèle basé sur le périmètre "DOM" et le modèle hybride, ce dernier devant avoir une cartographie des risques similaire à celle du modèle offrant les meilleures performances sur les DOM. En effet, cela voudra dire que le risque sera segmenté de la même manière d'un point de vue géographique.

La cartographie des risques prédite par le modèle hybride est la suivante :

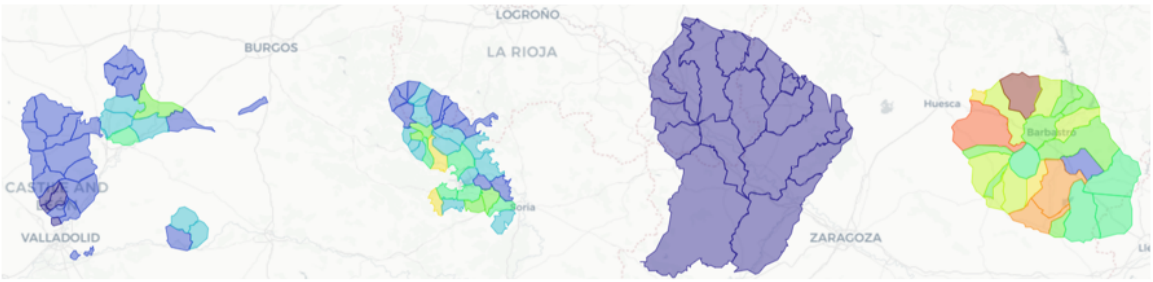


FIGURE 4.2.4 – Zones de risques déterminées par le modèle hybride

Nous pouvons déjà remarquer que la carte est similaire à celle obtenue avec le modèle basé sur le périmètre "DOM", ce qui est rassurant. Pour être sûr de la similarité nous devons regarder les spreads :

DOM	Modèle Hybride	Modèle DOM
Réunion	34,63%	30,43%
Martinique	10,05%	14,54%
Guadeloupe	33,27%	24,81%
Guyane	0,00%	0,00%
Tout DOM	69,47%	60,63%

TABLE 4.2.9 – Spread de la variable "Zonier" par DOM avec modèle hybride

Nous remarquons ainsi que les risques sont segmentés quasiment de la même manière entre les deux modèles, les ordres de grandeur étant les mêmes. Les spreads sont en effet presque identiques, le modèle hybride ayant un spread de 69,47% sur la variable "Zonier" contre 60,63% sur le modèle "DOM". Le modèle hybride segmente ainsi encore davantage.

Enfin, la conclusion quant au choix du modèle final pour la fréquence sera donnée en fin de chapitre.

4.3. La modélisation du coût moyen avec zonier

4.3.1 Modélisation du zonier coût moyen sur le périmètre "DOM"

Dans cette section sera détaillée l'intégration des variable géographiques et d'antécédents pour le coût moyen sur le périmètre "DOM". Lors de la modélisation du coût moyen hors zonier sur ce périmètre, nous avons vu que le choix de la distribution à adopter était relativement complexe entre la loi gamma et la loi de Box-Cox. C'est pourquoi nous allons implémenter des modèles pour chacune de ces deux distributions à chaque étape.

— **Modélisation des variables zoniers**

Ici, la variable zonier sera construite avec l'algorithme Georev en sélectionnant les deux combinaisons de paramètres suivantes :

- Une exposition minimale de 100, un nombre de voisin minimal de 10, et 10 zones de risque, qui va permettre de construire la variable "Zonier 1"

- Une exposition minimale de 100, un nombre de voisin minimal de 10, et 5 zones de risque, qui va permettre de construire la variable "Zonier 2".

Ci-dessous les performances de nos modèles pour les différentes variables zonier :

Métriques	Zonier 1 Gamma	Zonier 2 Gamma	Zonier 1 Box Cox	Zonier 2 Box-Cox
Gini Train K-Fold	18,09%	18,12%	18,11%	18,03%
Gini Test K-Fold	17,81%	17,87%	17,78%	17,83%
RMSE Train K-Fold	1211,44	1212,23	1245,50	1244,75
RMSE Test K -Fold	1208,13	1210,14	1240,12	1238,78
MAE Train K-Fold	662,67	664,67	617,37	618,84
MAE Test K-Fold	662,28	664,91	619,27	618,14
Gamma deviance Train -Fold	0,55582	0,55571	0,64050	0,63970
Gamma deviance Test K-Fold	0,56325	0,56567	0,64387	0,64265
Erreur globale Train K-Fold	2,11%	2,01%	23,67%	23,48%
Erreur globale Test K-Fold	1,09%	0,99%	23,65%	23,47%

TABLE 4.3.1 – Performances avec variables zoniers pour le coût moyen sur "Z DOM"

La variable "Zonier 2" offre des légèrement plus élevées avec chaque distribution. L'écart des métriques entre le "train k-folds" et le "test k-folds" montre que cette variable permet de moins sur-apprendre par rapport à la variable "Zonier 1". Ces performances meilleures peuvent s'expliquer par le plus faible nombre de modalité de la variable "Zonier 2" (5 zones) par rapport à la variable "Zonier 1" (10 zones). En effet, nous avons peu de données dans nos base coût moyen sur le périmètre, ce qui empêche une segmentation fine du risque.

Ainsi, nous avons l'intuition que le remplacement du zonier par une variable "Département DOM" ayant pour modalité "Réunion", "Guadeloupe", "Martinique", et "Guyane" pourrait permettre d'uniformiser la prédiction en associant un coefficient global pour chaque DOM afin d'améliorer les performances du modèle. Nous allons ainsi intégrer cette variable à laquelle on aura appliqué la fonction de lissage *random*().

Voici les performances de nos modèles avec cette nouvelle variable :

Métriques	Gamma variable "Département DOM"	Box Cox variable "Département DOM"
Gini Train K-Fold	18,71%	18,34%
Gini Test K-Fold	18,17%	18,15%
RMSE Train K-Fold	1207,62	1244,36
RMSE Test K -Fold	1202,85	1238,87
MAE Train K-Fold	660,63	612,91
MAE Test K-Fold	662,22	614,40
Gamma deviance Train -Fold	0,55552	0,64015
Gamma deviance Test K-Fold	0,56025	0,64292
Erreur globale Train K-Fold	1,39%	23,78%
Erreur globale Test K-Fold	1,34%	23,77%

TABLE 4.3.2 – Performances avec la variable "Département DOM" pour le coût moyen sur "Z DOM"

L'utilisation de la variable "Département DOM" permet en effet d'améliorer légèrement le modèle en terme de Gini, l'indicateur gagne en effet environ 0,3 points. De plus, les lois gamma et Box-Cox ont, à cette étape, le même pouvoir de segmentation au vu de leur indice de Gini respectif. La cartographie à la maille sera cependant présentée en annexe 3.

— **Modélisation des variables antécédents**

Pour la modélisation, des antécédents, la variable "Nombre de sinistre sur 10 ans" va être intégrée à notre modèle.

Les performances avec cette nouvelle variable sont les suivantes :

Métriques	Gamma variable antécédent	Box Cox variable antécédent
Gini Train K-Fold	19,07%	18,61%
Gini Test K-Fold	18,69%	18,31%
RMSE Train K-Fold	1206,22	1241,38
RMSE Test K -Fold	1199,22	1234,63
MAE Train K-Fold	663,28	611,55
MAE Test K-Fold	663,90	612,73
Gamma deviance Train -Fold	0,55182	0,63130
Gamma deviance Test K-Fold	0,55470	0,63360
Erreur globale Train K-Fold	0,03%	23,10%
Erreur globale Test K-Fold	0,00%	23,04%

TABLE 4.3.3 – Performances avec la variable antécédent pour le coût moyen sur "Z DOM"

Pour la distribution gamma, l'intégration de la variable d'antécédent permet de gagner 0,5 point de Gini, là où le gain en Gini n'est que de 0,15 point pour la distribution Box-Cox. C'est donc à cette étape que la distribution gamma se démarque en terme de segmentation de la variable réponse. De plus l'erreur globale devient quasiment nulle pour cette dernière là où elle reste très élevée pour la loi Box-Cox.

— **Application du modèle à la base "V DOM"**

Les performances suivantes sont obtenues en appliquant notre modèle aux bases "Z DOM" et "V DOM" :

Métriques	Modèle final Gamma	Modèle final Box-Cox
Gini Z DOM	19,03%	18,62%
Gini V DOM	17,84%	16,33%
RMSE Z DOM	1207,36	1242,57
RMSE V DOM	1327,87	1330,88
MAE Z DOM	663,29	611,91
MAE V DOM	705,82	638,28
Gamma deviance Z DOM	0,5522	0,6313
Gama deviance V DOM	0,5990	0,7066
Erreur globale Z DOM	0,02%	23,12%
Erreur globale V DOM	0,94%	24,71%

TABLE 4.3.4 – Performances des modèles finaux "coût moyen avec zonier" DOM

Concentrons nous d'abord sur les performances sur la base coût moyen "Z DOM" : pour chaque distribution le Gini permet de gagner environ 1 point de Gini. Le gain de Gini avec la modélisation du coût moyen, après intégration des variables zoniers et antécédents, est bien plus faible que celui avec la modélisation de la fréquence (11 points de gagné). Sur la base "V DOM", là aussi, la distribution gamma se démarque fortement avec un Gini supérieur de 1,5 points par rapport à celui obtenu avec la modélisation Box-Cox. De plus l'erreur globale avec la loi gamma reste très faible comparativement à celle de la loi Box-Cox.

4.3.2 Comparaison avec le zonier coût moyen "métropole+DOM"

Dans cette section, nous exposerons le modèle retenu pour la modélisation du coût moyen intégrant les variable géographique sur le périmètre "métropole+DOM" puis nous comparerons les performances avec le modèle de coût moyen du périmètre "DOM".

— Le modèle retenu sur le périmètre "métropole+DOM"

Rappelons que le modèle de coût moyen utilise notamment la distribution gamma, mais nous comparerons cette loi avec la distribution beta généralisée pour les mêmes raisons que l'utilisation de la loi Box-Cox sur le périmètre "DOM".

De même que pour le périmètre DOM, la variable géographique intégrée à notre modèle est "département DOM". Notons que la modélisation de cette variable peut se faire aussi bien sur le périmètre "métropole+DOM" que sur le périmètre "DOM". En effet, dans le cas où la modélisation est réalisée sur le segment "métropole+DOM", la variable "département DOM" aura pour nouvelle modalité "Métropole" qui aura son propre coefficient et les coefficients associés aux modalités "Réunion", "Guadeloupe", "Martinique" et "Guyane" seront donc finalement calculés uniquement sur les données DOM (les seules à avoir ces modalités). En ce sens il s'agit d'une forme de modèle hybride : les variables "hors zonier" provenant du modèle coût moyen hors zonier et la variable géographique ayant été ajustée uniquement sur les données du périmètre "DOM."

Enfin la variable antécédent ajoutée est "Nombre de sinistre sur 10 ans"

Nous pouvons donc ainsi analyser les performances de ce modèle directement sur les bases coût moyen "Z DOM" et "V DOM" :

Métriques	Modèle hybride Gamma	Modèle hybride Beta gen.
Gini Z DOM	16,28%	17,19%
Gini V DOM	18,03%	18,08%
RMSE Z DOM	1252,08	1209,88
RMSE V DOM	1288,56	1243,68
MAE Z DOM	608,49	661,09
MAE V DOM	614,54	675,13
Gamma deviance Z DOM	0,65880	0,57270
Gama deviance V DOM	0,67650	0,58450
Erreur globale Z DOM	24,11%	0,61%
Erreur globale V DOM	25,55%	2,30%

TABLE 4.3.5 – Performances du modèle hybride coût moyen sur le périmètre "DOM"

Commençons tout d'abord à comparer les performances de ces modèles avec les modèles issus de la modélisation hors zonier sur la base "Z DOM". La modélisation avec la loi Beta généralisée permet de gagner 0,7 point de Gini (de 15,58% à 16,28%) là où la loi de Gamma permet un gain de 1,2 point de Gini (de 15,96% à 17,19%). Concernant les résultats sur la base "V DOM", les indices de Gini sont semblables pour les deux distributions. En revanche, les autre métriques (et notamment l'erreur globale, 2% contre 26%) en dehors de la MAE sont en faveur de la modélisation avec la loi gamma. Ainsi, les conclusions sont les mêmes que pour la modélisation du coût moyen sur le périmètre "DOM" entre la loi gamma et la loi Box-Cox.

— **Comparaison des performances et analyse**

Les performances de nos 2 modèle coût moyen sur la base "V DOM" sont les suivantes :

Indicateurs	Modèle final coût moyen DOM	Modèle hybride coût moyen
Nb variables	10	13
Gini	17,84%	18,08%
RMSE	1327,87	1243,68
MAE	705,82	675,13
Avg Deviance	0,5990	0,5845
Erreur globale	0,940%	2,300%

TABLE 4.3.6 – Comparaison des performances des modèles coût moyen avec zonier sur la base "V DOM"

Le modèle hybride coût moyen offre de meilleures performances que le modèle coût moyen DOM. En effet, l'indice de Gini du modèle hybride est supérieur de 0,2 point, la déviance, la MAE et la RMSE sont plus faibles également. Seul l'erreur globale est lègerement plus élevée.

Regardons également comment le risque géographique a été classé entre les deu modèles :

DOM	Modèle Hybride	Modèle DOM
Réunion	93,51%	92,41%
Martinique	113,47%	109,23%
Guadeloupe	119,67%	117,25%
Guyane	102,86%	100,51%
Spread	27,97%	29,89%

TABLE 4.3.7 – Coefficient multiplicatif de la variable "Département" par DOM

Nous pouvons donc remarquer que dans les deux cas, le classement des risques est très proche, le modèle hybride permet donc d'obtenir la même segmentation géographique que le modèle basé uniquement sur le périmètre "DOM". Nous voyons également que pour le coût moyen, le département Guadeloupe est le plus risqué alors que la Réunion porte moins de risque.

Pour finir notre analyse, nous allons tracer les lift curves pour comprendre les écarts. Ces lift curves vont être tracées pour les distribution gamma, mais aussi pour les distributions Box-Cox et Beta généralisée afin de donner une conclusion finale sur le choix de la distribution pour la modélisation du coût moyen :

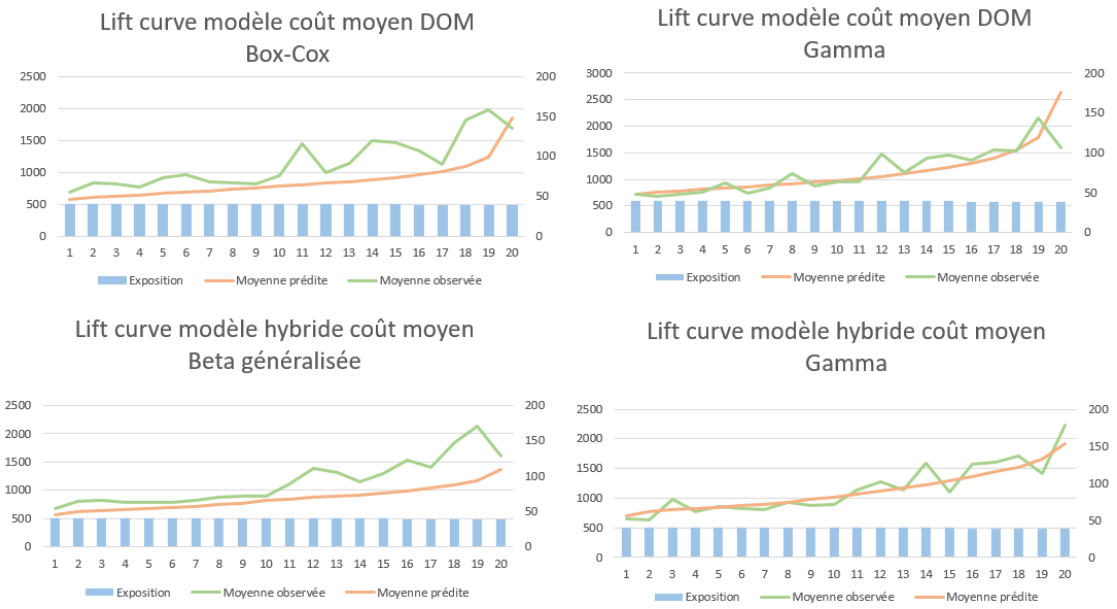


FIGURE 4.3.1 – Lift curve des modèles finaux coût moyen

La modélisation avec la distribution de Box-Cox issue du modèle coût moyen "DOM" a le moins bon pouvoir prédictif, les prédictions sont bien en dessous des valeurs réelles. Il en est presque de même pour la distribution beta généralisée issue du modèle hybride, les 10 premiers quantiles étant mieux estimés en moyenne. Concernant les distributions gamma, les lift curves sont proches, mais celle issue du modèle "DOM" a de moins bonnes prédictions concernant les profils les plus risqués (les 2 derniers quantiles).

4.4. Choix des modèles attritionnels et modélisation des graves

4.4.1 Conclusions sur la modélisation de la sinistralité attritionnelle

Tout le long de ce chapitre, nous avons peaufiné nos modèles au fur et à mesure afin d'aboutir aux modèles offrants les meilleurs performances possibles sur le périmètre "DOM". Ainsi, nous pouvons conclure sur la pertinence de l'intégration de la modélisation de la sinistralité attritionnelle dégât des eaux du "périmètre DOM" au périmètre "métropole".

Concernant la modélisation du coût moyen, le modèle offrant les meilleures performances sur la base de validation (base coût moyen "V DOM") est le modèle hybride : avec des variables hors zonier provenant d'un modèle "métropole+DOM" et une variable géographique "Département DOM" construite sur les données DOM. L'intégration de la modélisation du coût moyen du périmètre "DOM" est donc tout à fait possible. Et, dans ce cas, pour les contrats situés en métropole, la variable "Département DOM" aura un coefficient de 1 afin de ne pas affecter leur tarif. Ces contrats métropolitains auront leur propre coefficients de zonier (zonier construit avec Georev) et de la même manière ce coefficient sera de 1 pour les contrats ultramarins.

Enfin, la conclusion sur la distribution à utiliser est que la distribution Gamma est la plus adaptée à la modélisation du coût moyen grâce à un meilleur pouvoir prédictif comparé aux lois Box-Cox et Beta généralisée (qui sont cependant mieux adaptée à la distribution intrinsèque des données).

Concernant la modélisation de la fréquence, le modèle offrant les meilleures performances sur la base de validation (base fréquence "V DOM") est le modèle basé sur le périmètre "DOM". Cependant, le modèle hybride offre également de bonnes performances comparativement au modèle "métropole+DOM". D'autant plus que le zonier associé au modèle hybride est quasiment le même que celui du modèle "DOM" (il est important d'avoir la meilleure segmentation géographique possible). Ce modèle va donc être choisi afin d'avoir un équilibre entre les aspects techniques et opérationnels (maintenance des modèles).

Enfin, concernant la disbution à choisir, nous avons vu que la loi de Sichel était la plus adaptée, mais dans un environnement sans possibilité d'implémenter les modèles GAMLSS, il faut privilégier la loi négative binomiale par rapport à la Poisson.

Ainsi aussi bien pour la fréquence que pour le coût moyen, l'intégration de la modélisation de la sinistralité des DOM à la métropole est validée pour la garantie dégât des eaux.

4.4.2 Modélisation des sinistres graves

Jusqu'à présent, nous avons traité de la modélisation de la sinistralité attritionnelle. Dans cette section, nous allons détailler la modélisation de la sinistralité grave avec une approche de mutalisation des sinistres graves.

Rappelons que la charge de sinistre grave a été obtenue en écrêtant la charge totale avec un seuil de grave de 11 500 €. Ce seuil de grave a été déterminé à partir des données du périmètre "métropole+DOM" (voir section 2.2.2) et été utilisé aussi bien sur le périmètre "métropole+DOM" que sur le périmètre "DOM" pour la modélisation de la sinistralité attritionnelle. Cela signifie que les sinistres graves des DOM seront mutualisés avec les sinistres graves de la métropole, et inversement.

Définissons donc la prime pure grave qui est la partie de la prime pure correspondant à la mutualisation des sinistres graves :

$$\textbf{prime pure grave} = \frac{\textbf{sur} - \textbf{crete}}{\textbf{exposition}}$$

Il serait possible de mutualiser les sinistres graves avec tout le portefeuille uniformément. Cependant, nous allons réaliser une mutualisation légèrement segmentée en fonction des variables "Nombre de pièces", "Qualité de l'occupant", "Type d'habitation" et "Capitaux mobiliers".

Afin de déterminer les segments sur lesquels mutualiser nos sinistres graves, nous allons réaliser un arbre de regréssion afin d'expliquer la charge écrêtée en fonction de ces variables. Nous ferons en sorte d'obtenir un arbre de régréssion de faible complexité afin d'avoir un nombre contenu de segment.

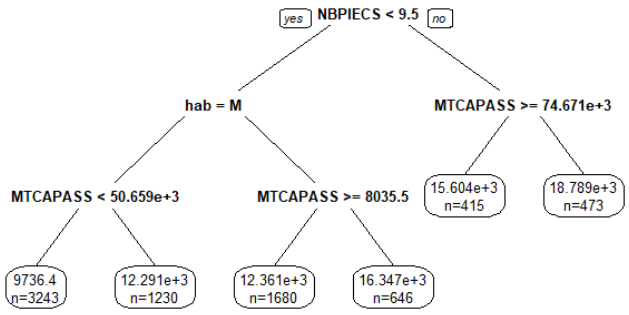


FIGURE 4.4.1 – Arbre de régression pour la prime pure grave

Grâce à cet arbre de régression, il est possible de déterminer 6 segments :

- Les maisons de moins de 9 pièces avec des capitaux mobilier supérieurs à 50 000€.

- Les maisons de moins de 9 pièces avec des capitaux mobilier inférieurs à 50 000€.

- Les appartements de moins de 9 pièces avec des capitaux mobilier supérieurs à 8 000€.

- Les appartements de moins de 9 pièces avec des capitaux mobilier inférieurs à 8 000€.

- Les logements de plus de 10 pièces avec des capitaux mobiliers supérieurs à 75 000 €.

- Les logements de plus de 10 pièces avec des capitaux mobiliers inférieurs à 75 000 €.

Pour chacun de ces 6 segments, la prime pure grave est calculée selon la formule indiquée précedemment.

Voici ces données résumées dans un tableau :

Segment	Exposition %	Charge grave totale	Prime pure grave
M - 9P capitaux >50 000	45,70%	31 284 622,00 €	1,97 €
M - 9P capitaux <50 000	7,12%	15 407 995,00 €	6,24 €
A - 9P capitaux > 8000	14,28%	10 095 173,00 €	2,04 €
A - 9P capitaux < 8000	31,38%	21 230 386,00 €	1,95 €
A/M + 10P capitaux < 75 000	1,05%	8 918 094,00 €	24,49 €
A/M + 10P capitaux > 75 000	0,47%	6 444 956,00 €	39,43 €
Total	100,00%	93 381 226,00 €	2,69 €

TABLE 4.4.1 – Prime pure grave par segment

Ainsi, chaque assuré du périmètre "métropole+DOM" aura une prime pure grave correspondant à son segment.

Cependant, avec cette méthodologie, les assurés situés dans les DOM paieront une part de prime pure grave trop importante par rapport à la sinistralité grave réellement observée dans les DOM. Il faut donc corriger leur primes pures graves avec un certain coefficient afin de s'aligner avec cette sinistralité observée.

La charge totale des graves dans les DOM est de 118 385 € alors que la prime pure grave dans les DOM déterminée avec notre méthode est de 867 497 €. Le coefficient de correction est ainsi obtenu en faisant le rapport de ces deux valeurs. Ce coefficient est donc égal à 0,136. Il nous suffit alors de multiplier les primes pures graves par ce coefficient pour obtenir la prime pure grave des DOM.

Cette correction est détaillée par segment dans le tableau suivant :

Segment	Prime pure grave	Prime pure grave corrigée
M -9P capitaux >50 000	1,97 €	0,27 €
M -9P capitaux <50 000	6,24 €	0,85 €
A - 9P capitaux > 8000	2,04 €	0,28 €
A - 9P capitaux < 8000	1,95 €	0,27 €
A/M + 10P capitaux < 75 000	24,49 €	3,34 €
A/M + 10P capitaux > 75 000	39,43 €	5,38 €
Total	2,69 €	0,37 €

TABLE 4.4.2 – Prime pure grave DOM par segment après correction

Effectivement, les montants de prime pure grave deviennent environ 85 % plus faible sur le périmètre "DOM" ce qui permet d'être plus accolé à la sinistralité réelle.

Chapitre 5

Analyse de la prime commerciale DOM

5.1. Construction de la prime commerciale

5.1.1 Synthèse sur la modélisation des autres garanties

Dans cette section seront détaillées les conclusions de l'intégration de la tarification du périmètre "DOM" à la tarification "métropole". En effet, la prime commerciale qui sera étudiée dans ce chapitre est basée sur les primes pures de chacune des garanties.

Les garanties suivantes ont été modélisées :

- la garantie "vol" en scindant les périmètres "appartement" et "maison" avec un modèle coût fréquence.

- la garantie "bris de glaces" en scindant les périmètres "appartement" et "maison" avec un modèle coût fréquence.

- la garantie "incendie en scindant les périmètres "appartement" et "maison" avec un modèle coût-fréquence.

- la garantie "climatique" avec un modèle de prime pure.

- la garantie "responsabilité civile" avec un modèle de prime pure.

Ainsi, ces garanties ont été modélisées en suivant la même méthodologie que pour la garantie "dégât des eaux" grâce au logiciel *Akur8*.

Pour chaque garantie, présentons le type de modélisation retenue et le type de modélisation géographique choisie :

Garantie	Modèle prime pure	Modèle freq	Modèle cm	Variable géographique
Vol	-	Hybride	Hybride	Zonier INSEE fréquence et DOM coût moyen
Bris de glace	-	Hybride	Hybride	Zonier INSEE fréquence et DOM coût moyen
Incendie	-	Hybride	Hybride	Zonier INSEE fréquence et DOM coût moyen
Climatique	"Métropole+DOM"	-	-	Pas de modélisation géographique
Responsabilité civile	"Métropole+DOM"	-	-	Pas de modélisation géographique

TABLE 5.1.1 – Sélection de modèle par garantie

Ainsi, les garanties pour lesquelles un modèle coût-fréquence a été utilisé, c'est le modèle hybride qui a été retenu pour le coût moyen et pour la fréquence. En effet, les performances étaient proches d'une modélisation basée uniquement sur le périmètre "DOM" et ces modèles ont donc été retenus pour les mêmes raisons que pour la garantie dégât des eaux. Et de même que pour la garantie dégâts des eaux, un zonier INSEE a été implémenté pour la fréquence alors que pour le coût moyen seul un coefficient pour chaque DOM a été ajusté.

Pour les garanties modélisées avec un modèle de prime pure, aucune modélisation géographique n'a été introduite, et la modélisation hors zonier s'est basée sur le périmètre "métropole+DOM".

Pour résumer, toutes les garanties ont été modélisées avec un modèle hors zonier basé sur le périmètre "métropole+DOM", et lorsque c'est possible, une modélisation géographique sur le périmètre "DOM" constituant alors une modélisation "hybride" permettant de concilier les aspects techniques et opérationnels. L'intégration du périmètre "DOM" à la tarification de France métropolitaine est donc validée.

5.1.2 Construction de la prime commerciale DOM

Dans cette section sera détaillée les étapes permettant de construire la prime commerciale :

- La première étape consiste à calculer la prime pure des garanties constituant le socle des garanties obligatoires (toutes les garanties vues jusqu'à présent). Pour chacune de ces garanties, ces primes pures sont obtenus après modélisation de la prime pure attritionnelle qui sera sommée à la prime pure grave selon les garanties où elle doit être modélisée (c'est à dire les garanties "dégâts des eaux", "vol" et "incendie"). C'est ce qui a été fait jusqu'à présent. La prime pure de la garantie "catastrophe naturelle" correspondant à 12% des primes "dommages" est également ajoutée. En sommant ces primes pures, la première couche de la prime commerciale est calculée, et la prime obtenue est appelé P_0.

- La seconde étape consiste à réaliser un *base level adjustment*, il s'agit de recoller à la sinistralité observée pour chaque garantie grâce à un coefficient de proportionnalité c_p qui sera à appliqué à la prime pure de chaque contrat. Ce coefficient se calcule au global (sur tout le portefeuille et non ligne à ligne) sur une certaine fenêtre d'observation F de la façon suivante /.

$$\mathbf{c_p} = \frac{\sum \mathbf{Charge_F}}{\sum \mathbf{Prime\ pure_F}}$$

A cette étape est également ajoutée la prime correspondant aux options (des garanties falcutatives) telle que l'option "dommage électrique". En sommant ces primes pures, la seconde couche de la prime pure est obtenue. Elle correspond à la sinistralité "brute" et est nommée P_1.

- La troisième consiste à ajouter les frais généraux (correspondant entre autres aux coûts de gestions et d'acquisitions des contrats), les commissions (part des primes destinés aux agents et courtiers) et enfin les frais de réassurance f_{reass} calculé comme suit :

$$\mathbf{f_{reass}} = \mathbf{P_{reass}} - \mathbf{S_{reass}} + \mathbf{C_{reass}}$$

avec P_{reass} les primes cédées, S_{reass} les sinistres cédées et C_{reass} les commissions de réassurance.
En ajoutant ces frais, la prime P_2 est obtenue.

- Enfin, la dernière consiste à prendre en compte le crédit commercial permettant de réduire la prime dans certains cas (selon le budget alloué). La prime commerciale P_3 effectivement payée par les assurés est ainsi calculée.

Ces différentes étapes ont été appliquées afin d'obtenir la prime commerciale sur le périmètre "DOM".

5.2. Résultats sur les affaires nouvelles DOM

5.2.1 Analyse de la prime commerciale

Dans cette section, nous allons analyser la prime commerciale P_3 obtenue pour les affaires nouvelles du 01/01/2020 au 31/06/2022 sur le périmètre "DOM".

Pour le périmètre "DOM" dans son entièreté, la nouvelle prime commerciale moyenne est de 242,65 € contre 187,27 € avant, soit une augmentation de près de 30 %.

Regardons en détails l'évolution globale de la prime commerciale par département :

DOM	Evolution globale de la prime commerciale
Guadeloupe	18,90%
Guyane	40,41%
Martinique	17,26%
Réunion	36,16%

TABLE 5.2.1 – Evolution globale de prime la commerciale par DOM

Ainsi, la nouvelle prime commerciale est bien plus élevée qu'avant, notamment à la Réunion et en Guyane. Un changement de prime aussi important peut engendrer une hausse des résiliations. Notre solution est d'appliquer un coefficient c_{tarif} dans le but de s'aligner à l'ancien tarif, ce coefficient est calculé sur chaque département D et est défini comme suit :

$$\mathbf{c_{tarif,D}} = \frac{\sum \mathbf{Prime_{ancienne,D}}}{\sum \mathbf{Prime_{nouvelle,D}}}$$

En appliquant ce coefficient à chaque contrat, nous alignons les tarifs à l'ancien, tout en conservant la segmentation du risque sur chaque département. De cette manière, il n'y a plus d'écarts globaux (les deux primes moyennes sont les mêmes). Observons plutôt les écarts relatifs moyens pour un contrat :

DOM	Ecart moyen par contrat avant correction	Ecart moyen par contrat après correction
Guadeloupe	24,70%	4,88%
Guyane	42,00%	1,14%
Martinique	21,48%	3,60%
Réunion	39,63%	2,55%
Tout DOM	34,10%	2,99%

TABLE 5.2.2 – Ecarts relatifs moyen de la prime commerciale par contrat

Nous remarquons alors que l'écart de prime moyen est beaucoup plus faible après correction. Ce nouveau tarif devient alors bien plus acceptable commercialement.

Regardons à présent la distribution des écarts relatifs moyens :

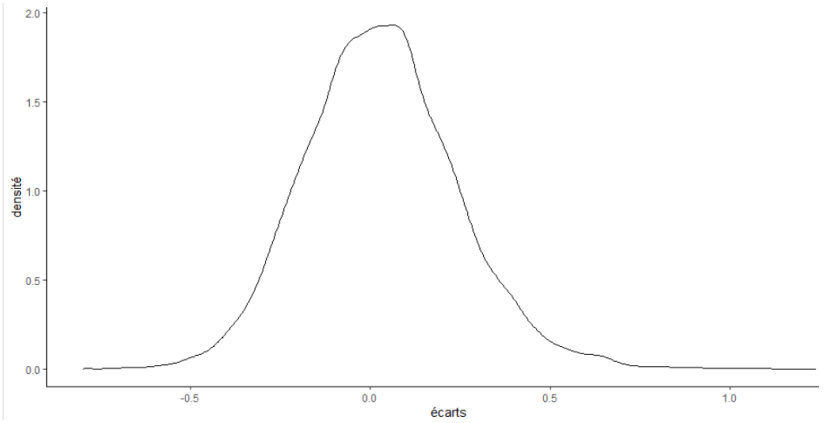


FIGURE 5.2.1 – Distribution des écarts moyens de la prime commerciale

Nous voyons alors que les écarts moyens sont globalement symétrique en 0. De plus, le calcul des quantiles montre que 50 % des contrats ont vu leur prime commerciale varier en - 15 % et + 15 %.

Enfin, regardons quelles sont les catégories de contrats où la variation de prime commerciale est la plus significative. Pour cela nous avons les écarts moyens relatifs en fonction de différentes variables grâce à un arbre de régression. L'arbre de régression obtenu est le suivant :

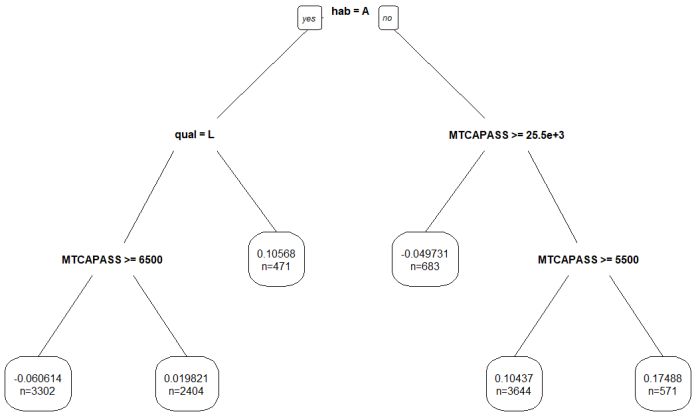


FIGURE 5.2.2 – Arbre de régression pour les écarts moyens relatifs de prime commerciale

Ainsi, les segments expliquant le plus les écarts sont les suivantes :

- Les locataires d'appartements avec des capitaux mobiliers supérieurs à 6 500 € qui ont en moyenne une prime commerciale plus faible de 6%.

- Les locataires d'appartements avec des capitaux mobiliers inférieurs à 6 500 € qui ont en moyenne une prime commerciale plus élevées de 2%.

- Les propriétaires d'appartement qui ont vu leur prime augmenter d'environ 10,5%.

- Les assurés avec une maison et des capitaux mobiliers de plus de 25 500 € ont vu leur prime diminuer de 5% environ.

- Les assurés avec une maison et des capitaux mobiliers compris entre 5 500 € et 25 500 € ont vu leur prime augmenter de 10% environ.

- Les assurés avec une maison et des capitaux mobiliers inférieurs à 5 500€ ont vu leur prime augmenter de 17% environ.

5.2.2 Analyse de la rentabilité technique

Dans cette section, nous allons étudier la rentabilité technique (hors frais) sur la même fenêtre d'observation et le même périmètre qu'en section 5.2.2.

Nous allons calculer la rentabilité technique grâce à l'indicateur *Expected Loss Ratio* appelé ELR. Cette indicateur se calcule de la façon suivante :

$$\textbf{ELR} = \frac{\textbf{Prime pure}}{\textbf{Prime commerciale}}$$

Dans notre cas, l'ELR sera calculé avec la nouvelle prime commerciale sur les garanties composant le socle des garanties obligatoires.

L'ELR sur notre périmètre des affaires nouvelles est de 76,41 %, ce qui est une valeur classique pour les affaires nouvelles (l'ELR étant ensuite amélioré au fur et à mesure des majorations annuelles).

Regardons les ELR globaux par DOM :

DOM	ELR
Guadeloupe	78,10%
Guyane	66,70%
Martinique	75,66%
Réunion	77,66%
Tout DOM	76,41%

TABLE 5.2.3 – ELR Globaux par DOM

Nous voyons ainsi que le département Guyane a l'ELR le plus faible à 66,70 % là où les autres départements ont tous environ le même ELR.

Enfin, de la même manière que pour la prime commerciale, nous allons regarder quelles sont les catégories d'assurés permettant d'obtenir la meilleure segmentation de l'ELR en utilisant un arbre de régression.

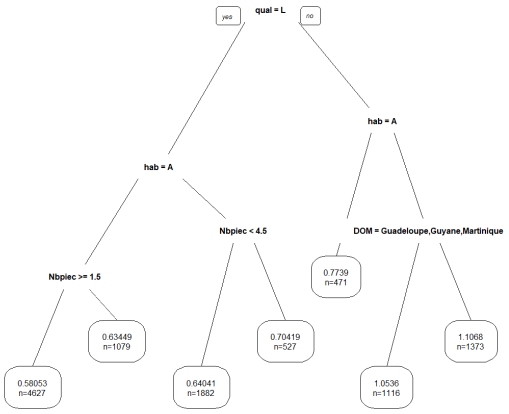


FIGURE 5.2.3 – Arbre de régression des ELR DOM

Ainsi, les catégories d'assurés segmentant le plus l'ELR sont les suivantes :

- les locataires d'appartement d'une pièce avec un ELR moyen de 58%

- les locataires d'appartement de plus d'une pièce avec un ELR moyen de 63%

- les locataires de maison de moins de 4 pièces avec un ELR moyen de 64%.

- les locataires de maison de plus de 5 pièces avec un ELR moyen de 70%.

- les propriétaires d'appartements avec un ELR moyen de 77%.

- les proriétaires de maison situées à la Réunion avec un ELR moyen de 105 %.

- les proriétaires de maison situées dans les autres DOM avec un ELR moyen de 111 %.

Ainsi nous avons déterminé les catégories d'assurés expliquant le plus notre ELR sur le périmètre des affaires nouvelles "DOM".

Conclusion

Tout au long de ce mémoire, nous avons mis en avant les différentes étapes amenant à la meilleure modélisation possible du périmètre "DOM" permettant de savoir si l'intégration de la tarification de ce périmètre à la tarification du périmètre "métropole" est possible sans dégradation importante des performances.

Nous avons en premier lieu explicité la préparation des données servant à la modélisation, à travers notamment la justification d'un choix de seuil de grave commun entre les périmètre "DOM" et "métropole+DOM" et le détail des différentes méthodes de discrétisation des variables quantitatives en variables qualitatives.

Dans un second temps, nous avons détaillé la modélisation de la sinistralité hors variables géographiques. C'est là que nous avons introduit les modèles de type GAMLSS permettant notamment de modéliser des lois de probabilités inusuelles mais plus adaptées à la distribution des variables réponses que nous souhaitions modéliser. A cette étape, nous avons ainsi observé que, les performances des modèles hors zonier basés sur le périmètre "DOM" sont biens meilleures que celles des modèles hors zonier basés sur le périmètre "DOM" dans l'explication de la sinistralité DOM.

Nous avons ensuite intégré une variable géographique à nos modéles afin d'améliorer leur pouvoir prédictif. En constatant les gains de performances conséquents sur la fréquence en intégrant un zonier basé sur le périmètre "DOM" au modèle hors zonier DOM, nous avons décidé de construire un modèle "hybride" utilisant d'une part le modèle hors zonier métropole+DOM et une modélisation géographique basé sur le périmètre "DOM" d'autre part. Ce modèle hybride a montré de bonnes performances en rattrapant notamment le pouvoir prédictif moins probant du modèle métropole+DOM : il est en effet meilleur dans le cas de la modélisation du coût moyen tout en offrant des performances très correctes sur la fréquence (qui restent relativement moins bonnes que le modèle fréquence DOM). Ainsi nous avons décidé d'utiliser les modèles "hybrides" pour cette garantie.

Concernant le choix des distributions à utiliser pour un modèle coût-fréquence, nous avons conclu que, pour la garantie dégâts des eaux, la fréquence devait être modélisée par une loi de Sichel, et à défaut par une loi binomiale négative, là où la loi de Poisson est beaucoup moins adaptée. Pour le coût moyen, bien que les distributions Box-Cox et Beta généralisée ajustent mieux la distribution de nos variables réponses, la modélisation avec la loi gamma permet au final d'obtenir le meilleur pouvoir prédictif.

Plus loin, nous avons modélisé la sinistralité grave pour la garantie dégâts des eaux en réalisant une mutualisation légèrement segmentée groupe portant le plus de risque grâce à un arbre de régression. La

sinistralité grave du périmètre "DOM" était confondue avec celle de la métropole et nous avons ainsi dû corriger cet écart avec un coefficient permettant d'être accolé à la sinistralité observée. Toutes ces étapes ont ainsi permis de modéliser le dégâts des eaux sur le périmètre "DOM".

Après cela, nous avons étendu nos conclusions sur les autres garanties. Et c'est la modélisation hybride qui a été choisie, notamment car elle offre de bonnes performances mais permet aussi de réduire l'impact opérationnel.

Enfin, une analyse commerciale a permis de comparer l'ancienne prime commerciale à la nouvelle. Et après avoir corrigé cette dernière pour s'accoler aux ancienne primes commerciales, nous avons établi quelles étaient les catégories d'assurés impactées grâce à un arbre de régression. Le même processus a été réalisé sur la rentabilité technique avec l'indicateur ELR.

Bibliographie

[1] Kratz M. (2020) *Extreme Value Theory, Theory and Application to Risk Management, ISUP*

[2] Hunault G. (2022) *Découpage en classes et discrétisation, Université d'Angers*

[3] Hyunji K. (2015) *Data preprocessing, discretization for classification, package R*

[4] Charpentier A. (2018) *Actuariat Assurance Non-Vie, ENSAE*

[5] Frees E. & Meyers G. & Cummings A. (2012) *Insurance Ratemaking and a Gini Index, University of Wisconsin and ISO Innovative Analytics*

[6] Thomas M. (2021) *Econométrie de l'assurance non-vie, ISUP*

[7] Stasinopoulos & Rigby R. (2017) *Flexible Regression and Smoothing Using GAMLSS in R , Chapman & Hall Book*

[8] Ruoyan M. (2004) *Estimation of Dispersion Parameters in GLMs with and without Random Effects, Stockholm University*

[9] Perrin S. (2021) *Refonte des ELR sur la garantie Dégâts des eaux pour le produit Habitation, Mémoire d'actuariat*

[10] Toesca R. (2020) *Zonier Georev, AXA*

Table des figures

Liste des tableaux

Annexes

Annexe I. Moments de la loi Beta généralisée II

Soit Y suivant une loi Beta généralisée de type 2 de paramètre μ, σ, ν et τ.

Alors :

$$\mathbb{E}(\mathbf{Y}) = \begin{cases} \mu \frac{\mathbf{B}(\tau+\mathbf{1}/\sigma, \tau-\mathbf{1}/\sigma)}{\mathbf{B}(\nu,\tau)} & si\ \tau > 1/\sigma \\ \infty & si\ \tau \leq 1/\sigma \end{cases}$$

et :

$$\mathbf{Var}(\mathbf{Y}) = \begin{cases} \mu^{\mathbf{2}} \{ \frac{\mathbf{B}(\nu+\mathbf{2}\sigma, \tau-\mathbf{2}\sigma)\mathbf{B}(\nu,\tau) - [\mathbf{B}(\nu+\mathbf{1}/\sigma, \tau-\mathbf{1}/\sigma)^{\mathbf{2}}]}{[\mathbf{B}(\nu,\tau)^{\mathbf{2}}]} & si\ \tau > 2/\sigma \\ \infty & si\ \tau \leq 2/\sigma \end{cases}$$

avec :

$$\mathbf{B}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{\Gamma}(\mathbf{a})\mathbf{\Gamma}(\mathbf{b})}{\mathbf{\Gamma}(\mathbf{a}+\mathbf{b})}$$

et Γ la fonction Gamma.

Annexe II. Famille de distribution t

En section 3.2.2 a été introduite la distribution de Box-Cox t (pour le coût moyen) étant définie grâce à une variable aléatoire T suivant une loi de la famille de distribution t aussi appelée loi de Student t généralisée.

Soit Y suivant une loi de la famille de distribution t, de paramètre μ dans $\mathbb{R}$, $\sigma > 0$ et $\nu > 0$. Alors Y a pour densité sur $\mathbb{R}$:

$$\mathbf{f_Y}(\mathbf{y}|\mu,\sigma,\nu) = \frac{\mathbf{1}}{\sigma \mathbf{B}(\mathbf{1/2},\nu/\mathbf{2})\nu^{\mathbf{1/2}}}\left[\mathbf{1} + \frac{(\mathbf{y}-\mu)^{\mathbf{2}}}{\sigma^{\mathbf{2}}\nu}\right]^{-(\nu+1)/2}$$

de moments :

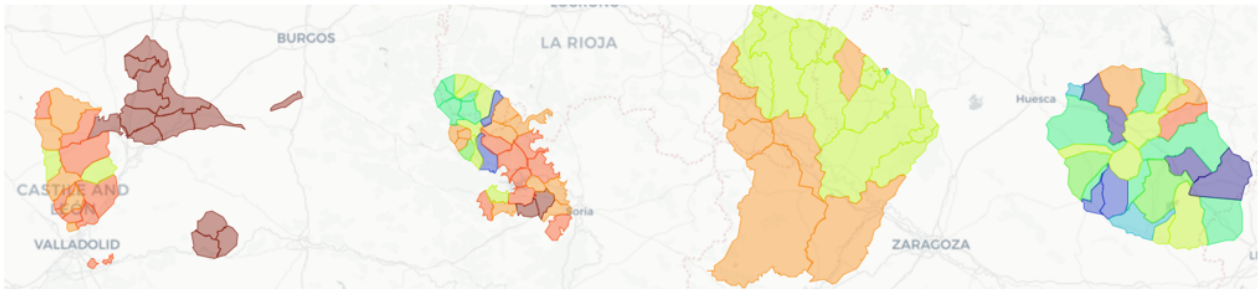
$$\mathbb{E}(\mathbf{Y}) = \begin{cases} \mu & si\ \nu > 1 \\ \text{indéfini} & si\ \nu \leq 1 \end{cases}$$

et :

$$\mathbf{Var}(\mathbf{Y}) = \begin{cases} \frac{\sigma^{\mathbf{2}}\nu}{\nu-\mathbf{2}} & si\ \nu > 2 \\ \infty & si\ \nu \leq 2 \end{cases}$$

Annexe III. Cartographie des risques coût moyen DOM

En section 4.3.1 a été décidé d'intégrer le risque géographique avec une variable "Département DOM". Ici sera présentée, la cartographie du risque coût moyen en utilisant un zonier INSEE :



Zones de risques coût moyen par INSEE

Et les spreads associés :

DOM	Modèle DOM	Exposition
Réunion	17,60%	56,85%
Martinique	27,60%	20,94%
Guadeloupe	27,60%	13,83%
Guyane	9,43%	8,38%
Tout DOM	29,56%	100,00%

Spread de la variable "Zonier" par DOM