Mémoire présenté le : Mardi 9 Mai 2023

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

Par : Enola BARLIOT

Titre : Apport de la Data Science pour la tarification en assurance MRH

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

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

Membre présents du jury de l'Institut des Actuaires — signature

Mme Gabrielle TERRE

Membres présents du jury de l'ISFA

Mme Esterina MASIELLO

Entreprise :

Nom : GENERALI

Signature :

Directeur de mémoire en entreprise :

Nom : Sébastien LEFEVRE

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

Mémoire d'Actuariat

Apport de la Data Science pour la tarification en assurance MRH

BARLIOT Enola

Résumé

La constante évolution des contextes économique, réglementaire, juridique et environnemental conduit les différents acteurs du marché de l'assurance à faire face à une forte concurrence. Ce contexte difficile contraint les assureurs à devoir trouver le bon équilibre entre proposer un niveau de prime acceptable et maintenir une rentabilité technique. Pour résoudre cette problématique, les compagnies se reposent notamment sur l'optimisation des processus de tarification, élément qui constitue un des cœurs de métier de l'actuariat.

Parmi tous les outils disponibles, les plus anciens, nommés GLM (modèles linéaires généralisés), restent les plus fréquemment utilisés dans le processus de tarification. Ce mémoire propose une étude comparative et concurrentielle de ce type de modèle avec des pratiques de *Data Science* innovantes sur cinq garanties d'une assurance multirisque habitation, où l'éventuel apport de ces méthodes non paramétriques (arbre de régression, *Random Forest* et *Gradient Boosting*) sera recherché.

Pour déterminer si l'utilisation des méthodes de *Data Science* représente ou non une avancée par rapport aux méthodes GLM, cette étude va s'appuyer sur l'analyse d'indicateurs permettant de mesurer les performances prédictives des différents modèles comparés : la MSE et sa racine carrée, la RMSE, moins sensible aux valeurs extrêmes, mais aussi l'indice de Gini.

L'étude comporte trois grands axes : le premier présente le marché et les modèles utilisés, le second met en œuvre ces modèles pour l'estimation de la fréquence et du coût et le dernier analyse les résultats obtenus en les comparant avec le tarif actuellement en vigueur au sein de l'entreprise initiatrice de ce projet, l'ÉQUITÉ, filiale dédiée à la gestion des partenariats de GENERALI France.

Mots clés : Tarification, assurance MRH, méthode fréquence-coût, modèles linéaires généralisés (GLM), Data Science, CART, arbres de régression, Random Forest, eXtreme Gradient Boosting, indice de Gini, MSE, RMSE.

Abstract

The constant evolution of the economic, regulatory, legal and environmental context leads the various players in the insurance market to face a strong competition. This difficult context forces insurers to find the right balance between offering an acceptable level of premium and maintaining a technical profitability. To overcome this issue, companies are relying on the optimization of the pricing process in particular, which is one of the core businesses of actuarial sciences.

Among all the methods available, one of the oldest, named GLM (generalized linear models), remains the most frequently used in the pricing process. This dissertation proposes a comparative and competitive study of this type of model with innovative Data Science practices on five guarantees of home insurance, where the possible benefits of these non-parametric methods (regression tree, Random Forest and Gradient Boosting) will be sought.

To determine whether or not the use of Data Science methods represents an advance over GLM methods, this study will be based on the analysis of indicators that allow to measure the predictive performance of the various compared models : the MSE and its square root, the RMSE, less sensitive to extremes values, and also the Gini index.

The study has three main axes : the first presents the market and the models used, the second implements these models for both the frequency and cost estimation and the last one analyzes the results obtained by comparing them with the rate currently in effect within the company initiating this project, l'ÉQUITÉ, a subsidiary dedicated to managing GENERALI France's partnerships.

Keywords : Pricing, Home insurance, frequency-cost method, generalized linear models (GLM), Data Science, CART, regression trees, Random Forest, eXtreme Gradient Boosting, Gini index, MSE, RMSE.

Synthèse

Contexte et Objectif du mémoire

La tarification dans un contrat d'assurance est le processus par lequel un assureur évalue le risque représenté par un bien ou une personne assurée et fixe un montant de prime qui doit être payé par le souscripteur du contrat ainsi construit, pour couvrir ce risque. Cependant, contrairement à la détermination du prix dans d'autres secteurs commerciaux, la prime d'assurance est déterminée avant d'en connaître son montant exact, caractérisant ainsi l'inversion du cycle de production propre à ce secteur.

L'apparition de nouveaux acteurs et la prise en compte de contraintes toujours plus nombreuses (juridiques et réglementaires par exemple), sur un marché de la MRH déjà tendu et saturé, ont exacerbé la concurrence. Il devient primordial pour les assureurs d'optimiser leurs processus de tarification pour atteindre un double objectif : conserver, voire augmenter leur part de marché et maintenir leur équilibre technique.

Cette optimisation est possible notamment par l'utilisation de modèles mathématiques prédictifs, comme les GLM, qui se sont imposés comme la norme dans le domaine de l'assurance, pour le développement de modèles de tarification. Toutefois, un certain enthousiasme, lié à l'apparition des modèles de type *Data Science*, autour des années 2010, a permis la mise en place de processus de tarification utilisant ces nouvelles méthodes. Ces modèles sont, malgré tout, encore peu mis en œuvre en raison de difficultés d'interprétabilité.

Au travers de l'étude de cinq garanties spécifiques des contrats MRH (Dégâts des eaux, Bris de Glace, Incendie, Vol et Responsabilité Civile), ce mémoire propose de comparer des modélisations de fréquences et de coûts des sinistres par les méthodes citées précédemment, afin d'obtenir, pour chaque garantie, une estimation de tarif la plus pertinente possible, dans une approche "fréquence-coût". Le tarif issu des calculs sera comparé au tarif actuellement proposé par l'ÉQUITÉ, filiale de GENERALI FRANCE. L'objectif final visé par l'entreprise étant le dégagement de pistes d'amélioration, pour chacune de ces garanties, permettant de moderniser le tarif en vigueur.

Les données, à l'origine de cette étude, reposent sur les bases sinistres des contrats de sept partenaires de l'ÉQUITÉ. La période d'observation retenue s'étend sur 7 années, du 01/01/2015 au 31/12/2021. Divers retraitements ont été réalisés pour obtenir les bases qui ont permis l'élaboration des différents modèles, comme ceux concernant les valeurs manquantes ou la gestion des corrélations entre les variables explicatives, par exemple.

Pour la modélisation des fréquences ainsi que celle des coûts, des calculs utilisant plusieurs GLM vont être mis en œuvre afin de déterminer le modèle le plus performant et de le retenir comme base de comparaison pour évaluer la pertinence d'autres modèles de type *Machine Learning* également testés. Les performances des différents modèles sont mesurées à partir du calcul d'indicateurs, notamment la MSE (ou RMSE) et l'indice de Gini.

Les modèles de *Data Science* mis en concurrence avec les GLM dans cette étude sont :

- les arbres de régression CART
- les forêts aléatoires ou *Random Forest*
- les *eXtreme Gradient Boosting* ou *XGBoost*

Modélisation des fréquences

Dans un premier temps, deux types de GLM "classiques" seront considérés pour modéliser la fréquence des sinistres : le GLM avec la loi de Poisson comme loi sous-jacente et le GLM reposant sur une loi Binomiale Négative (BN). Après avoir réalisé une sélection des variables et les regroupements nécessaires afin d'optimiser chaque modèle, les indicateurs de performance cités précédemment ont été mesurés. Les MSE et les indices de Gini obtenus étant très similaires pour les deux types de lois sous-jacentes testées, deux autres indicateurs ont été pris en compte (la déviance et la statistique de Pearson).

	DDE		Incendie		Vol		BDG		RC	
	Poisson	BN	Poisson	BN	Poisson	BN	Poisson	BN	Poisson	BN
Déviance	27 086	22 119	4 697.64	4 126.94	4 887.43	3 819	6 355.29	5 296.45	6 993.16	5 408.78
MSE	0.0442	0.04421	0.00526	0.00525	0.0071	0.0071	0.01077	0.01077	0.00774	0.00774
Pearson	500 892	491 479	944 845.27	980 529.13	365 955.56	363 419.42	342 688	340 092.38	589 882.6	585 348.59
Indice de Gini	0.417795	0.4197186	0.5575089	0.5577626	0.4439663	0.4451197	0.4628807	0.4629143	0.4178024	0.4185361

TABLE 1 – Comparaison GLM fréquence par garantie

D'après les différents résultats obtenus pour chaque garantie, conclusion est faite que, quelle que soit la garantie considérée, le GLM avec pour loi sous-jacente la loi Binomiale Négative est le plus pertinent pour modéliser la fréquence.

Les résultats obtenus par ces GLM sont ensuite comparés à ceux des autres modèles mis en œuvre dans l'étude. Avant de déterminer les performances des modèles de *Data Science*, ces derniers sont optimisés en testant différentes combinaisons de paramètres influençant le modèle. Dans un souci d'harmonisation et de cohérence globale des modèles, les variables retenues dans le cadre des modélisation de *Data Science* sont également celles qui ont été retenues pour les GLM.

Le tableau suivant regroupe l'ensemble des résultats des différents indicateurs de performance.

	DDE		Incendie		Vol		BDG		RC	
	MSE	Gini	MSE	Gini	MSE	Gini	MSE	Gini	MSE	Gini
CART	0.04413	0.3272846	0.00524	0.3823965	0.0071	0.2602603	0.01077	0.2660607	0.00774	0.2698749
Random Forest	0.04404	0.2647458	0.00525	0.4352728	0.00709	0.2802304	0.01076	0.2471913	0.00774	0.258314
XGBoost	0.0443	0.4433974	0.00526	0.5213907	0.0071	0.4528791	0.0108	0.4865415	0.00775	0.4470658
GLM	0.04421	0.4197186	0.00525	0.5577626	0.0071	0.4451197	0.01077	0.4629143	0.00774	0.4185361

TABLE 2 – Comparaison des modèles fréquences par garantie

La première observation pouvant être faite à partir de ce tableau est que la MSE varie faiblement d'un modèle à l'autre pour chaque garantie.

Cette tendance ne se confirme pas avec l'indice de Gini. En effet, pour les modèles CART et *Random Forest*, ces indices sont inférieurs à ceux obtenus dans les GLM les plus performants. En prenant en considération la quasi-stabilité des MSE, ces modèles ne sont donc pas retenus pour la modélisation des fréquences.

Dans le cas des modèles *XGBoost*, à l'exception de la garantie Incendie, tous les indices de Gini sont supérieurs à ceux des GLM. Cette hausse reste pourtant peu significative, représentant moins de 10% dans une comparaison relative avec le même indice obtenu pour les modèles GLM.

Toutefois, la détermination du modèle le plus performant ne se réduit pas aux seuls résultats des différents indices calculés. D'autres éléments doivent être pris en compte comme la vitesse d'apprentissage du modèle, sa facilité d'explication et son interprétabilité. Ainsi, pour les modèles *XGBoost* créés, dans le cas des fréquences, l'analyse des résultats permet de conclure que l'évolution favorable d'un seul indicateur et le gain de prédiction qui en découle ne suffisent pas à surpasser les inconvénients inhérents à ce type de modèles, et donc à les retenir comme les plus performants dans le cadre de cette étude.

Pour toutes les garanties, le modèle retenu pour l'estimation des fréquences est ainsi le modèle GLM avec comme loi sous-jacente la loi Binomiale Négative.

Modélisation des coûts des sinistres

Comme pour la modélisation de la fréquence, la première étape de la modélisation des coûts des sinistres par garantie repose sur la création de deux modèles GLM avec respectivement la loi Gamma et la loi Log-normale comme lois sous-jacentes.

Pour certaines garanties, le manque de données au niveau des coûts des sinistres conduit à ne capter l'influence que de certaines variables et pour celles retenues, peu de modalités apparaissent significatives. Pour tenter de pallier aux conséquences de ce manque de données, une approche de modélisation du coût "toutes garanties confondues" (modèle Global) est réalisée, autant pour les GLM que pour les approches de *Data Science*.

	DDE		Incendie		Vol		BDG		RC	
	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini
Gamma	1 537.423	0.2494755	4 508.29	0.1118854	2 402.897	0.314756	476.0823	0.2915579	1 636.005	0.2669111
Log-normale	1 547.791	0.2354648	4 594.252	0.09233819	2 385.457	0.2713674	479.8847	0.3775043	1 628.876	0.2059338
Global - Gamma	1 544.154	0.2017297	4 690.938	0.2059591	2 407.435	0.2038371	481.903	0.212794	1 627.014	0.1876072
Global - Log-normale	1 548.291	0.1857381	4 709.741	0.1746979	2 417.439	0.1732983	483.581	0.1895579	1 629.254	0.1681829

TABLE 3 – Comparaison GLM coût par garantie

La comparaison des modèles, par garantie, permet de déduire que pour les garanties DDE, Vol et RC, les GLM Gamma spécifiques aux garanties sont les plus pertinents et pour la garantie BDG, le modèle GLM Log-normale spécifique à la garantie est retenu. Pour la garantie Incendie, le modèle GLM Gamma "toutes garanties confondues" est conservé. Pour expliquer le résultat obtenu pour cette dernière garantie, il convient de remarquer que l'interaction entre les variables tarifantes est relativement similaire entre les garanties DDE et Incendie. Le modèle "toutes garanties confondues" étant fortement influencé par la présence importante des sinistres DDE, il améliore la compréhension de l'effet des variables sur le coût incendie, au lieu de générer des bruits.

Les GLM sélectionnés comme plus performants sont ensuite comparés aux modèles réalisés à partir d'approches de *Data Science*. Le même principe que pour la modélisation des fréquences est appliqué : les algorithmes sont entraînés à partir des variables explicatives choisies par les GLM. Ce principe correspond à l'"Approche 1" (notée A1 dans les tableaux). Comme expliqué précédemment, le manque de données a pu avoir pour conséquence le retrait de variables explicatives dont l'impact n'a pas pu être capté par les GLM. Pour tenter de remédier à ce constat, une seconde approche (Approche 2 notée A2 dans les tableaux) est mise en œuvre pour éventuellement permettre aux modèles de *Data Science* de saisir certains de ces effets, en allouant l'ensemble des variables disponibles pour l'apprentissage des modèles.

Le tableau suivant résume l'ensemble des résultats pour les modélisations des coûts.

		DDE		Incendie		Vol		BDG		RC	
		RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini
CART	A1	1 534.523	0.1736876	4 509.869	0.08684865	2 423.11	0.313368	477.824	0.2563355	1 634.765	0.2663519
	A2	1 516.723	0.1708604	4 534.496	0.1216667	2 402.455	0.2063629	477.7486	0.2324474	1 632.9	0.2182258
	Global	1 548.503	0.1896539	4 530.407	0.0000	2 417.811	0.2138844	503.7382	0.07471641	1 635.777	0.07092169
RF	A1	1 537.729	0.2081555	4 510.345	0.08309836	2 423.454	0.3096154	483.993	0.3244768	1 634.818	0.2681032
	A2	1 542.535	0.2246438	4 510.48	0.1580726	2 395.905	0.2873589	477.3401	0.2059568	1 632.622	0.3389661
	Global A1	1 546.163	0.2156593	4 564.004	0.1496013	2 572.604	0.1473126	493.758	0.1714064	1 627.813	0.1583464
	Global A2	1 543.662	0.1973317	4 598.818	0.1123101	2 597.546	0.1436079	477.4478	0.1472163	1 628.226	0.1793928
XGB	A1	1 556.399	0.2607366	4 507.757	0.1117732	2 407.209	0.2832557	483.456	0.3210083	1 634.336	0.2657947
	A2	1 550.5	0.2536284	4 590.634	0.2958713	2 436.144	0.3392619	487.6812	0.370297	1 647.323	0.4289614
	Global A1	1 555.304	0.254827	4 658.478	0.2859541	2 466.951	0.3621171	485.9179	0.3159959	1 669.346	0.4196129
	Global A2	1 554.252	0.2665662	4 719.267	0.3409151	2 516.372	0.3769019	490.8996	0.3535124	1 656.167	0.4629331
GLM		1 537.423	0.2494755	4 690.938	0.2059591	2 402.897	0.314756	479.8847	0.3775043	1 636.005	0.2669111

TABLE 4 – Comparaison des modèles de coût par garantie

Le choix du modèle le plus performant pour la modélisation du coût, pour chaque garantie, s'appuie sur l'analyse réalisée, combinée à l'arbitrage issu des résultats de deux indicateurs (RMSE et indice de Gini), mais aussi sur la prise en compte de la facilité d'explication et d'interprétation du modèle considéré, par rapport au GLM.

Afin d'illustrer les arbitrages, le cas de la garantie DDE est présenté à titre d'exemple. Pour les modèles CART "toutes garanties confondues" et les *Random Forest*, les RMSE augmentent et les indices de Gini diminuent par rapport aux GLM. Ces modèles ne sont donc pas retenus. Pour le modèle CART spécifique à la garantie, la RMSE s'améliore de moins de 1% pour une dégradation de plus de 30% de l'indice de Gini. L'utilisation de ce modèle est également écartée. Enfin, pour les *XGBoost*, l'augmentation des indices de Gini est de moins de 6% et celle des RMSE d'environ 1%. Cette amélioration constatée de l'indice de Gini n'est pas suffisante pour pallier les difficultés d'explication et d'interprétabilité des *XGBoost*. Le choix pour la modélisation des coûts DDE se porte ainsi sur le modèle GLM utilisant une loi Gamma.

Les indicateurs du modèle de coût sélectionné comme le plus performant apparaissent en rouge dans le tableau 4.

Comparaison avec le tarif actuel

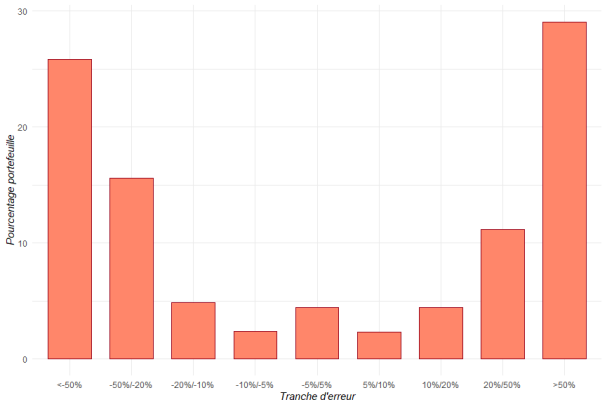

FIGURE 1 – Comparaison tarif estimé/tarif actuel pour la garantie BDG

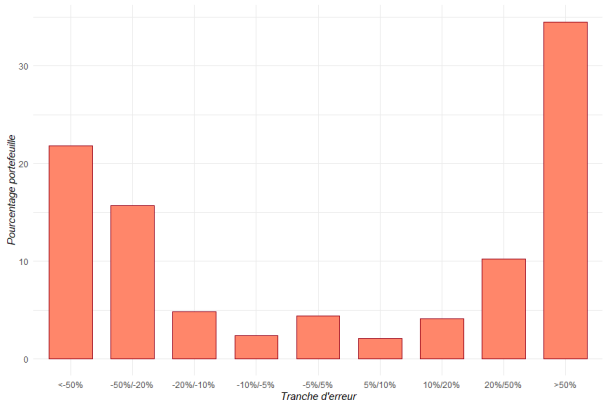

FIGURE 2 – Comparaison tarif estimé/tarif actuel pour la garantie DDE

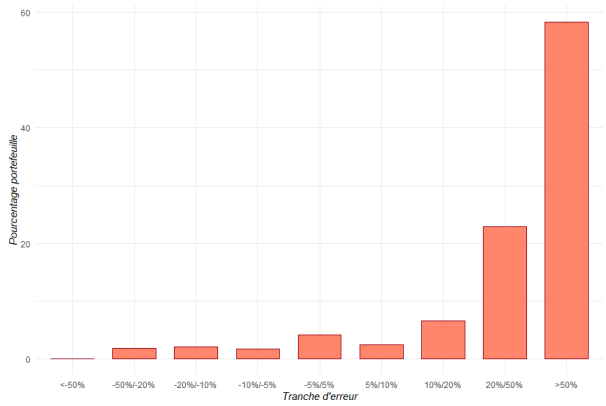

FIGURE 3 – Comparaison tarif estimé/tarif actuel pour la garantie Incendie

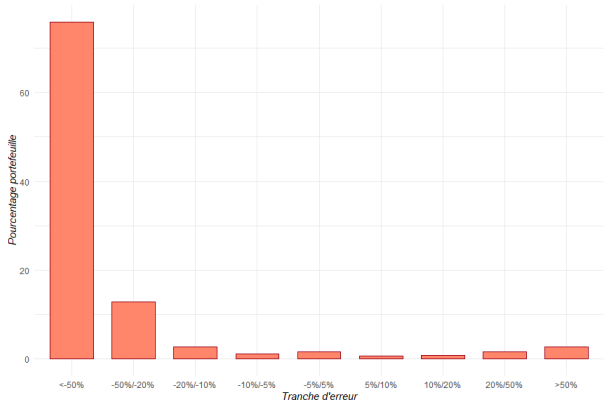

FIGURE 4 – Comparaison tarif estimé/tarif actuel pour la garantie Vol

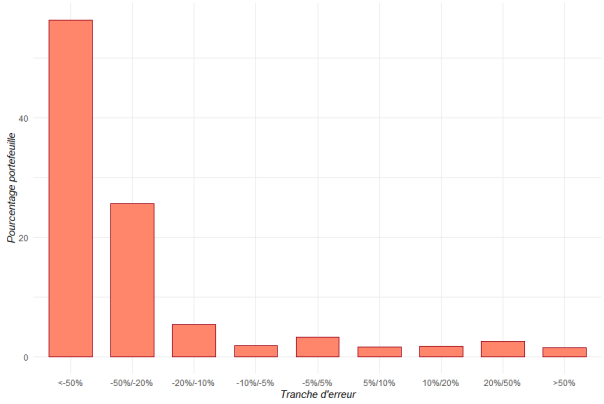

FIGURE 5 – Comparaison tarif estimé/tarif actuel pour la garantie RC

Pour les garanties BDG et DDE, une sous-estimation et une sur-estimation par rapport aux tarifs actuels sont constatées. Pour les trois autres garanties étudiées, une sur-tarification ou une sous-tarification par rapport au tarif actuel sont observées.

Une étude plus approfondie de ces surestimations et sous-estimations sur l'intégralité des garanties amène à la détection des profils les plus représentés dans les groupements d'écarts relatifs les plus extrêmes, permettant ainsi de mettre en avant les biais de tarification.

L'ensemble des conclusions concernant ces profils est consigné dans le tableau suivant.

	Sous-tarification	Sur-tarification
DDE	Maison-Locataire-plus de 2 pièces principales-sans franchise/200/400-sans OV	Maison-Propriétaire-plus de 4 pièces principales
		Appartement/Maison-PNO
		Appartement-Propriétaire
Incendie		Moins de 3 pièces principales
Vol	Plus de 3 pièces principales	
RC	Moins de 3 pièces principales	

Table 5 – Récapitulatif conclusion par garantie

Dans le cas de la garantie BDG, aucun effet significatif ne permet de réaliser une conclusion précise sur les profils dans les écarts de tarifs les plus importants.

Conclusion et ouverture de l'étude

Les modèles de *Data Science* restent actuellement peu utilisés pour la tarification par rapport aux modèles GLM. Toutefois, ce type de modèles peut s'avérer plus performant dans l'estimation de la prime. Ainsi, lors de la création ou de la révision des tarifs, il serait judicieux de les mettre en concurrence avec les modèles GLM.

L'analyse des profils sur les groupements d'écart de tarif permet de dégager une première conclusion en mettant en évidence les effets qui devront être modernisés dans le tarif actuel, afin d'obtenir une meilleure adaptation à la sinistralité observée sur le portefeuille de l'ÉQUITÉ.

Pour toutes les garanties, exception faite de la garantie DDE, les bases sinistres sont peu fournies. Ceci peut être une des raisons pour lesquelles les tarifications de ces garanties ne peuvent capter tous les impacts des variables sur les coûts des sinistres. Cette constatation pourrait aussi expliquer pourquoi les modèles *XGBoost* "toutes garanties confondues" et "toutes variables" se sont avérés les plus performants sur les garanties Incendie, Vol et RC. Le phénomène est encore plus flagrant pour la garantie Incendie dont le GLM le plus performant était aussi celui "toutes garanties confondues" utilisant la loi Gamma. Il serait intéressant de pouvoir évaluer les modèles spécifiques sur des bases ayant plus de sinistres par garantie et ainsi confirmer ou infirmer l'hypothèse émise ci-dessus, en constatant ou non une meilleure estimation des coûts.

Pour aller au-delà des travaux effectués dans ce mémoire, et pouvoir ainsi confirmer les conclusions faites sur les tarifs, des comparaisons avec les résultats techniques observés pourraient être réalisées par segment et par garantie.

Executive summary

Background and Purpose of the dissertation

Pricing in an insurance contract is the process through which an insurer assesses the risk represented by an insured asset or person and sets a premium amount to be paid by the contract's policyholder to cover that risk. However, unlike pricing in other commercial sectors, the insurance premium is determined before the exact amount is known, thus characterising the inversion of the production cycle in this sector.

The emergence of new players and the consideration of ever more constraints (particularly legal and regulatory), in an already tense and saturated home insurance market, have increased competition. It is becoming essential for insurers to optimise their pricing processes to achieve a dual objective : to maintain or even increase their market share and to maintain their technical balance.

This optimisation is possible in particular through the use of predictive mathematical models, such as GLMs, which have become the standard in the insurance field for the development of pricing models. However, a certain enthusiasm, linked to the appearance of data science models around 2010, has allowed the implementation of pricing processes using these new methods. Despite this, these models are still not widely used due to difficulties of interpretation.

Through the study of five specific guarantees of the home contracts (Water Damage, Glass Breakage, Fire, Theft and Civil Liability), this dissertation proposes to compare the frequency modelling and claims' costs modelling by the methods mentioned above, in order to obtain, for each guarantee, the most relevant estimated rate possible, in a "frequency-cost" approach. The rate resulting from the calculations will be compared to the rate currently proposed by l'ÉQUITÉ, a subsidiary of GENERALI FRANCE. The company's final aim is to identify areas for improvement, for each of these guarantees, enabling a modernisation of the current rate.

The data used for this study is based on the claims databases of seven partners of l'ÉQUITÉ. The observation period used is 7 years, from 01/01/2015 to 31/12/2021. Various adjustments were made to obtain the databases that allowed the development of the different models, such as those concerning missing values or the management of correlations between explanatory variables, for example.

For the frequencies and costs' modelling, calculations using several GLMs will be carried out in order to determine the best performing model and to use it as a basis for comparison to assess the relevance of Machine Learning models also tested. The performance of the different models is measured by calculating indicators, notably the MSE (or RMSE) and the Gini index.

The data science models competing with the GLM in this study are :

- regression trees CART
- Random Forest
- *eXtreme Gradient Boosting* or *XGBoost*

Frequencies Modelling

First, two types of "classical" GLM will be considered to model the frequency of claims : the GLM with the Poisson distribution as underlying distribution and the GLM based on a Negative Binomial distribution (BN). After selecting the variables and making the necessary groupings to optimize each model, the performance indicators mentioned above were measured. Since the MSE and Gini indexes obtained were very similar for the two types of underlying distributions tested, two other indicators were taken into account (deviance and Pearson's statistic).

	Water Damage		Fire		Theft		Glass Breakage		Civil Liability	
	Poisson	BN	Poisson	BN	Poisson	BN	Poisson	BN	Poisson	BN
Deviance	27 086	22 119	4 697.64	4 126.94	4 887.43	3 819	6 355.29	5 296.45	6 993.16	5 408.78
MSE	0.0442	0.04421	0.00526	0.00525	0.0071	0.0071	0.01077	0.01077	0.00774	0.00774
Pearson	500 892	491 479	944 845.27	980 529.13	365 955.56	363 419.42	342 688	340 092.38	589 882.6	585 348.59
Gini Index	0.417795	0.4197186	0.5575089	0.5577626	0.4439663	0.4451197	0.4628807	0.4629143	0.4178024	0.4185361

TABLE 6 – Frequency GLM comparison by guarantee

From the different results obtained for each guarantee, it is concluded that, regardless of the guarantee considered, the GLM with the underlying Negative Binomial distribution is the most relevant to model the frequency.

The results obtained by these GLM are then compared with those of the other models implemented in the study. Before determining the performance of the Data Science models, they are optimised by testing different combinations of parameters influencing the model. For the sake of uniformity and overall consistency of the models, the variables used for modelling in Data Science are also those selected for the GLM.

The following table summarises the results of the different performance indicators.

	Water Damage		Fire		Theft		Glass Breakage		Civil Liability	
	MSE	Gini	MSE	Gini	MSE	Gini	MSE	Gini	MSE	Gini
CART	0.04413	0.3272846	0.00524	0.3823965	0.0071	0.2602603	0.01077	0.2660607	0.00774	0.2698749
Random Forest	0.04404	0.2647458	0.00525	0.4352728	0.00709	0.2802304	0.01076	0.2471913	0.00774	0.258314
XGBoost	0.0443	0.4433974	0.00526	0.5213907	0.0071	0.4528791	0.0108	0.4865415	0.00775	0.4470658
GLM	0.04421	0.4197186	0.00525	0.5577626	0.0071	0.4451197	0.01077	0.4629143	0.00774	0.4185361

TABLE 7 – Frequency models' comparison by guarantee

The first observation that can be made from this table is that the MSE varies slightly from one model to another for each guarantee.

This trend is not confirmed with the Gini index. Indeed, for the CART and Random Forest models, these indicators are lower than those obtained by the best performing GLM. Taking into consideration the near stability of the MSE, these models are therefore not retained for frequency modelling.

In the case of the XGBoost models, with the exception of the Fire guarantee, all the Gini indexes are higher than those of the GLM. However, this increase remains relatively insignificant, constituting less than 10% in a relative comparison with the same index obtained for the GLM models.

Nevertheless, the determination of the best performing model cannot be reduced to the results of the different indicators calculated. Other elements must be taken into account, such as the learning speed of the model, its ease of explanation and its interpretability. Thus, for the XGBoost models created, in the case of frequencies, the results' analysis leads to the conclusion that the beneficial evolution of a single indicator and the resulting gain in prediction are not sufficient to overcome the intrinsic disadvantages of this type of models, and therefore to retain them as the best performing models in the framework of this study.

For all guarantees, the chosen model for the frequencies' estimate is therefore the GLM model with the underlying distribution being the Negative Binomial distribution.

Claims Cost Modelling

Following the example of frequency modelling, the first step in modelling claims costs by guarantee is based on the creation of two GLM models with the Gamma distribution and the Log-normal distribution as underlying distributions, respectively.

For some guarantees, the lack of data regarding the claims costs leads to capturing the influence of only certain variables, and for those selected, few of the modalities appear significant. In an attempt to compensate the consequences of this lack of data, an "all guarantees" cost modelling approach (Global model) is carried out, both for GLM and for Data Science approaches.

	Water Damage		Fire		Theft		Glass Breakage		Civil Liability	
	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini
Gamma	1 537.423	0.2494755	4 508.29	0.1118854	2 402.897	0.314756	476.0823	0.2915579	1 636.005	0.2669111
Log-normal	1 547.791	0.2354648	4 594.252	0.09233819	2 385.457	0.2713674	479.8847	0.3775043	1 628.876	0.2059338
Global - Gamma	1 544.154	0.2017297	4 690.938	0.2059591	2 407.435	0.2038371	481.903	0.212794	1 627.014	0.1876072
Global - Log-normal	1 548.291	0.1857381	4 709.741	0.1746979	2 417.439	0.1732983	483.581	0.1895579	1 629.254	0.1681829

TABLE 8 – Cost GLM comparison by guarantee

The models' comparison, by guarantee, enables us to deduce that for the Water Damage, Theft and Civil Liability guarantees, the guarantee-specific Gamma GLM are the most relevant and for the Glass Breakage guarantee, the guarantee-specific Log-normal GLM model is retained. For the Fire guarantee, the "all guarantees" Gamma GLM model is retained. To explain the result obtained for this last guarantee, it should be noted that the interaction between the pricing variables is relatively similar between the Water Damage and Fire guarantees. As the "all guarantees" model is strongly influenced by the significant presence of water damage claims, it improves the understanding of the variables' effect on the fire cost, instead of generating noises.

The GLM selected as the best performing ones are then compared with models based on data science approaches. The same principle as for frequency modelling is applied : the algorithms are trained on the explanatory variables selected by the GLM. This principle corresponds to our "Approach 1" (noted A1 in the tables). As explained above, the lack of data may have resulted in the removal of explanatory variables whose impact could not be captured by the GLM. In an attempt to remedy this, a second approach (Approach 2, denoted A2 in the tables) is implemented to eventually allow data science models to capture some of these effects, by allocating all available variables for model training.

The following table summarises all the results for the cost modelling.

		Water Damage		Fire		Theft		Glass Breakage		Civil Liability	
		RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini	RMSE	Gini
CART	A1	1 534.523	0.1736876	4 509.869	0.08684865	2 423.11	0.313368	477.824	0.2563355	1 634.765	0.2663519
	A2	1 516.723	0.1708604	4 534.496	0.1216667	2 402.455	0.2063629	477.7486	0.2324474	1 632.9	0.2182258
	Global	1 548.503	0.1896539	4 530.407	0.0000	2 417.811	0.2138844	503.7382	0.07471641	1 635.777	0.07092169
RF	A1	1 537.729	0.2081555	4 510.345	0.08309836	2 423.454	0.3096154	483.993	0.3244768	1 634.818	0.2681032
	A2	1 542.535	0.2246438	4 510.48	0.1580726	2 395.905	0.2873589	477.3401	0.2059568	1 632.622	0.3389661
	Global A1	1 546.163	0.2156593	4 564.004	0.1496013	2 572.604	0.1473126	493.758	0.1714064	1 627.813	0.1583464
	Global A2	1 543.662	0.1973317	4 598.818	0.1123101	2 597.546	0.1436079	477.4478	0.1472163	1 628.226	0.1793928
XGB	A1	1 556.399	0.2607366	4 507.757	0.1117732	2 407.209	0.2832557	483.456	0.3210083	1 634.336	0.2657947
	A2	1 550.5	0.2536284	4 590.634	0.2958713	2 436.144	0.3392619	487.6812	0.370297	1 647.323	0.4289614
	Global A1	1 555.304	0.254827	4 658.478	0.2859541	2 466.951	0.3621171	485.9179	0.3159959	1 669.346	0.4196129
	Global A2	1 554.252	0.2665662	4 719.267	0.3409151	2 516.372	0.3769019	490.8996	0.3535124	1 656.167	0.4629331
GLM		1 537.423	0.2494755	4 690.938	0.2059591	2 402.897	0.314756	479.8847	0.3775043	1 636.005	0.2669111

TABLE 9 – Cost models' comparison by guarantee

The choice of the best performing model for cost modelling, for each guarantee, is based on the analysis carried out, combined with the ruling between the results of two indicators (RMSE and Gini index), but also on the consideration of the ease of explanation and interpretation of the model considered, compared to the GLM.

To illustrate the ruling, the example of the Water Damage guarantee is presented. For the CART models "all guarantees" and the Random Forest, the RMSE increase and the Gini indexes decrease compared to the GLM. These models are therefore not retained. For the CART model specific to the guarantee, the RMSE improves by less than 1% for a deterioration of more than 30% in the Gini index. The use of this model is also discarded. Finally, for the XGBoost, the increase in the Gini indexes is less than 6% and the increase in the RMSE is about 1%. This observed improvement in the Gini index is not sufficient to overcome the difficulties of explanation and interpretability of the XGBoost. The choice for the modelling of water damage costs is therefore the GLM model using a Gamma distribution.

The indicators of the cost model selected as the best performing appear in red in Table 9.

Comparison with the current rate

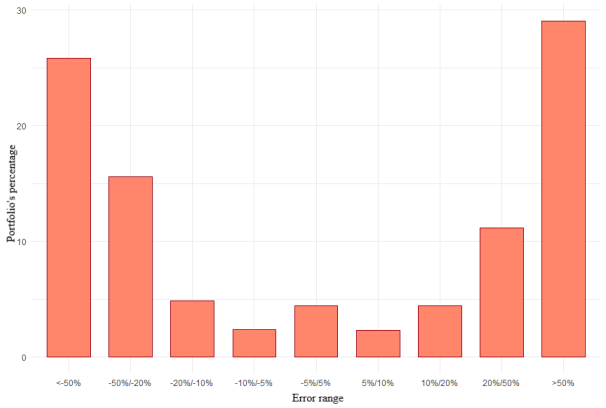

FIGURE 6 – Comparison of estimated rates with current rates for Glass Breakage guarantee

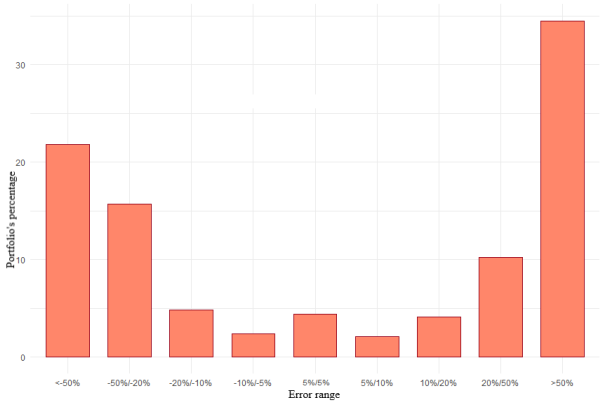

FIGURE 7 – Comparison of estimated rates with current rates for Water Damage guarantee

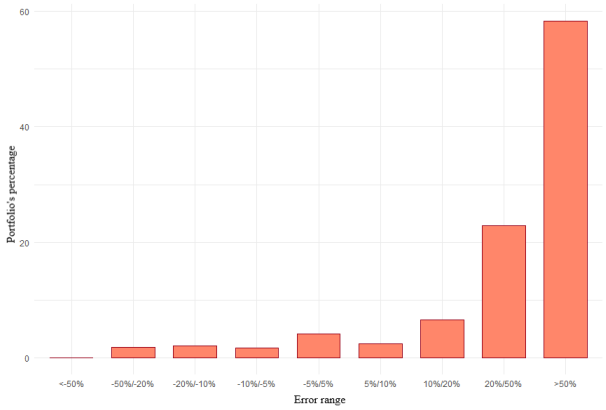

FIGURE 8 – Comparison of estimated rates with current rates for Fire guarantee

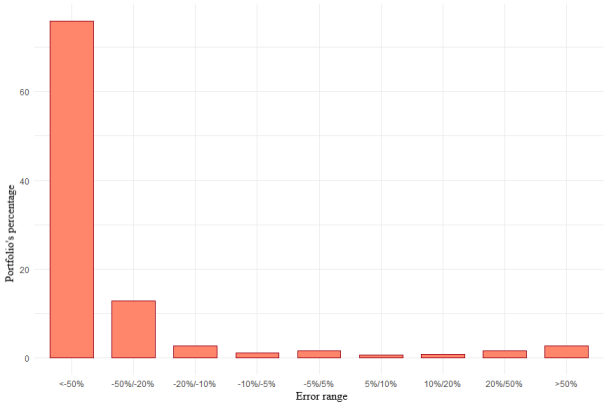

FIGURE 9 – Comparison of estimated rates with current rates for Theft guarantee

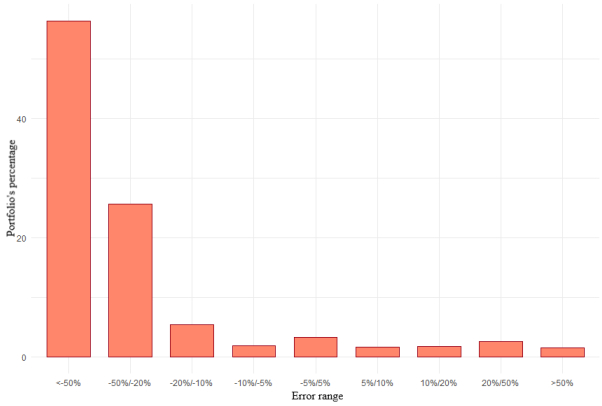

FIGURE 10 – Comparison of estimated rates with current rates for Civil Liability guarantee

For the Glass Breakage and Water Damage guarantees, an underestimation and an overestimation compared to the current rates is noted. For the other three guarantees studied, an overestimation or an underestimation compared to the current rate is observed.

A more detailed study of these overestimations and underestimations across all guarantees leads to the detection of the most represented profiles in the most extreme relative gap groupings, thus highlighting pricing bias.

The overall conclusions regarding these profiles are recorded in the following table.

	Underpricing	Overpricing
Water Damage	House-Tenant-more than 2 main rooms-without excess/200/400-without valuables	House-Owner-more than 4 main rooms
		Flat/House- non-occupying owner
		Flat-Owner
Fire		Less than 3 main rooms
Theft	More than 3 main rooms	
Civil Liability	Less than 3 main rooms	

TABLE 10 – Overview of conclusion by guarantee

In the case of the Glass Breakage guarantee, there are no significant effects that allow a precise conclusion to be drawn about the profiles in the largest price differences.

Conclusion and opening of the study

Data Science models are currently rarely used for pricing compared to GLM models. Nevertheless, this type of model may prove to be more efficient in estimating the premium. Thus, when creating or revising rates, it would be wise to put them in competition with GLM models.

The profiles' analysis on the rate gap groupings allows a first conclusion to be drawn by highlighting the effects which will have to be modernised in the current rate, in order to obtain a better adaptation to the claims observed on the l'ÉQUITÉ portfolio.

For all guarantees, with the exception of the Water Damage guarantee, the claims databases are poorly supplied. This may be one of the reasons why the pricing of these guarantees cannot capture all the variables' impacts on claims costs. This may also explain why the "all guarantees" and "all variables" XGBoost models performed best for the Fire, Theft and Civil Liability guarantees. The phenomenon is even more blatant for the Fire guarantee, for which the best performing GLM was also the "all guarantees" GLM using the Gamma distribution. It would be interesting to be able to evaluate the specific models with databases with more claims per guarantee and thus confirm or refute the hypothesis put forward above, by observing or not a better estimation of the costs.

To go beyond the work done in this dissertation, and thus be able to confirm the conclusions made on rates, comparisons with the observed technical results could be made by segment and by guarantee.

Table des matières

Synthèse **1**

Executive summary **6**

Remerciements **14**

Introduction **15**

1 Contexte de l'étude **17**
1.1 L'assurance MRH . . . 17
1.1.1 Les garanties . . . 17
1.1.2 Le marché français . . . 18
1.1.2.1 En globalité . . . 18
1.1.2.2 Le cas de la MRH . . . 20
1.2 La direction des Partenariats . . . 21

2 Aspects théoriques pour la tarification **23**
2.1 Les enjeux de la tarification . . . 23
2.1.1 Le principe de la mutualisation . . . 23
2.1.2 Le principe de la segmentation . . . 24
2.1.3 L'arbitrage entre mutualisation et segmentation . . . 25
2.2 Le modèle fréquence-coût . . . 25
2.3 Les modèles linéaires généralisés . . . 26
2.3.1 Les composantes d'un GLM . . . 26
2.3.2 Estimations des paramètres . . . 28
2.3.3 Les tests de significativité . . . 29
2.3.4 Avantages et Inconvénients . . . 30
2.4 Tree-based Models . . . 30
2.4.1 Arbre de régression . . . 30
2.4.2 Les forêts aléatoires . . . 32
2.4.3 *Gradient Boosting* . . . 35
2.4.4 Optimisation des paramètres . . . 37
2.5 Les indicateurs de performances . . . 37
2.5.1 Indicateurs pour la sélection du GLM . . . 37
2.5.2 Indicateurs pour la comparaison des modèles de l'étude . . . 38

3 Traitements des données **41**
3.1 Périmètre de l'étude . . . 41
3.2 Données de la base contrats . . . 42
3.2.1 Les variables conservées . . . 42
3.2.2 Construction et traitements sur cette base . . . 42
3.3 Données des bases sinistres . . . 44
3.3.1 Les variables de l'étude . . . 44
3.3.2 Construction de la base . . . 44
3.4 Bases sévérité et bases fréquence . . . 46
3.5 Statistiques Descriptives . . . 47

3.6 Analyses des variables . . . 50
3.6.1 Regroupement des variables . . . 50
3.6.2 Analyse des corrélations entre les variables . . . 50
3.6.3 Séparation de la base en base d'apprentissage et de validation . . . 52

4 Application des modèles pour la modélisation de la fréquence 54
4.1 Modèle Linéaire Généralisé . . . 54
4.1.1 Avant la modélisation . . . 54
4.1.2 Choix des variables et calibration de leurs effets . . . 55
4.1.3 Analyse des résidus . . . 56
4.1.4 Validation des modèles et sélection du GLM le plus efficace . . . 58
4.2 Arbre CART . . . 64
4.3 Les Forêts Aléatoires . . . 67
4.4 eXtreme Gradient Boosting : XGBoost . . . 69

5 Application des modèles pour la modélisation du coût 72
5.1 Modèle Linéaire Généralisé . . . 72
5.1.1 Avant la modélisation . . . 72
5.1.2 Choix des variables et calibration de leurs effets . . . 73
5.1.3 Analyse des résidus . . . 73
5.1.4 Sélection du GLM le plus efficace par garantie . . . 75
5.1.5 Modèles de coût global . . . 81
5.2 Arbre CART . . . 83
5.3 Les Forêts aléatoires . . . 86
5.4 eXtreme Gradient Boosting : XGBoost . . . 89

6 Comparaison avec le tarif actuel 93

Conclusion et Ouverture 104

Bibliographie 105

A Glossaire 107

B Algorithme de Newton-Raphson 108

C Provisionnement Non-Vie 109
C.1 Théorie de *Chain-Ladder* . . . 109
C.2 Application sur les données . . . 109

D Graphiques complémentaires pour l'étude des indices *As If* 111
D.1 Indice FFB . . . 111
D.2 Indice de l'inflation . . . 112

E Les tests de corrélation 114
E.1 Pearson . . . 114
E.2 Kendall . . . 114
E.3 Spearman . . . 114

F Les lois de fréquences 115
F.1 La loi de Poisson . . . 115
F.2 La loi Binomiale Négative . . . 115
F.3 Les lois "zéro-tronquée" . . . 116

G Les lois de coût 117
G.1 La loi Normale . . . 117
G.2 La loi Gamma . . . 117
G.3 La loi Log-normale . . . 118

H Tableaux complémentaires pour les GLM fréquence **119**
H.1 Tableaux pour adéquation des lois avant modélisation . . . 119
H.2 Tableaux complémentaires pour la significativité des variables sélectionnées . . . 120

I Illustrations complémentaires pour les GLM fréquence **123**
I.1 Garantie DDE . . . 123
I.2 Garantie BDG . . . 124
I.3 Garantie Incendie . . . 127
I.4 Garantie Vol . . . 130
I.5 Garantie RC . . . 132

J Eléments graphiques complémentaires pour les modèles CART fréquence **136**

K Eléments Graphiques *Random Forest* fréquence **138**

L Eléments Graphiques *XGBoost* fréquence **142**

M Tableaux complémentaires pour les GLM coût **144**
M.1 Tableaux pour adéquation des lois avant modélisation . . . 144
M.2 Significativité des variables sélectionnées après regroupement . . . 145

N Illustrations complémentaires pour les GLM coût **147**
N.1 Garantie BDG . . . 147
N.2 Garantie Incendie . . . 149
N.3 Garantie RC . . . 151
N.4 Garantie Vol . . . 153
N.5 Toutes Garanties (Global) . . . 155

O Eléments graphiques complémentaires pour les modèles CART cout **157**

P Eléments Graphiques *Random Forest* coût **161**

Q Eléments Graphiques *XGBoost* coût **168**

Table des figures **172**

Liste des tableaux **179**

Remerciements

Je tiens à remercier, en premier lieu, mon tuteur en entreprise, Sébastien LEFEVRE, responsable d'études d'actuariat au sein de l'ÉQUITÉ, pour son écoute, ses conseils, sa bienveillance et le temps passé à la transmission de ses compétences d'actuaire accompli durant mon stage de master 1 et mon année d'alternance. Toute ma gratitude pour sa patience, son encadrement régulier et avisé pour la réalisation de ce mémoire.

Merci aussi à l'ensemble du personnel du Département des Partenariats de GENERALI pour leur accueil, leur soutien et leur aide ainsi que leur bonne humeur. Ils m'ont permis d'effectuer mon alternance dans une ambiance dynamique et de développer mes connaissances dans le domaine des assurances.

Je veux aussi associer à ces remerciements l'intégralité de l'équipe pédagogique de l'Institut de Science Financière et d'Assurances (ISFA) qui a assuré ma formation. Plus particulièrement, un grand merci à mon tuteur académique, Frédéric PLANCHET, professeur au sein de l'ISFA, pour son suivi, ses conseils et ses remarques lors de l'élaboration de ce mémoire, mais aussi pour son enseignement de qualité lors de mes années passées à l'ISFA.

Pour finir, je remercie également toutes les personnes qui m'ont accompagnée et soutenue durant mes travaux et mon parcours universitaire et à qui je souhaite témoigner toute ma reconnaissance.

Introduction

Un contrat d'assurance est un contrat par lequel un assureur s'engage, auprès d'un souscripteur, à couvrir un ou plusieurs risques spécifiques moyennant le paiement d'une somme d'argent appelée cotisation. Cette définition du contrat d'assurance amène à se questionner sur l'élaboration de ces cotisations par la tarification.

L'objet de l'étape de tarification, quel que soit le risque assuré, est de déterminer le plus précisément possible ce que chaque souscripteur d'un contrat devrait, en moyenne, coûter à son assureur. S'il est plus facile, dans toute autre activité commerciale, de déterminer un prix de vente, l'activité d'assurance est confrontée à un cycle de production inversé, qui lui impose de commercialiser un produit avant de savoir ce qu'il va réellement lui coûter. La sous-évaluation des risques pouvant découler d'une tarification insuffisamment segmentée et donc mal calibrée, fait de la détermination la plus précise possible du tarif, un des sujets centraux pour les assureurs.

Or, le marché de l'assurance Habitation est tendu et saturé notamment avec l'apparition de nouveaux acteurs, les "bancassurances", mais aussi la montée en puissance des insurtechs présents sur les comparateurs. Cette tension a aussi été accentuée depuis l'entrée en vigueur de la loi Hamon en 2015. Cette loi donne le droit aux assurés de résilier leur contrat d'assurance, à tout moment, après un an de souscription et ce, sans pénalités. En facilitant le changement d'assureur, cette loi a exacerbé la concurrence entre compagnies et a eu comme conséquences, entre autres, une baisse des cotisations demandées aux assurés. Les assureurs font alors face à deux enjeux antagonistes : conserver, voire augmenter leur part de marché et assurer le maintien de leur équilibre technique.

Dans ce contexte de forte tension, optimiser le processus de tarification apparaît comme étant un défi vital pour assurer le bon équilibre entre ces deux enjeux. Un assureur qui optimisera efficacement sa tarification pourra élargir et fidéliser sa clientèle en proposant un prix de contrat attractif, tout en évitant la dégradation de son équilibre technique. Si cet équilibre n'est pas respecté, c'est-à-dire s'il y a une sur-tarification ou une sous-tarification, cela aura pour conséquences respectivement une perte de portefeuille (le principe de mutualisation, qui est une hypothèse majeure en assurance, ne serait donc plus applicable) ou un déficit (qui nécessitera au mieux une restructuration avec un changement de stratégie, voir dans les cas les plus graves, un arrêt complet des activités de l'entreprise).

Pour obtenir des tarifs compétitifs, il faut en premier lieu mettre en relation les risques avec les variables décrivant ces risques. Pour tenter d'y parvenir, les compagnies d'assurance ont mis en œuvre des modèles de régression. Depuis 1972, date à laquelle les GLM ont été introduits par Nelder et Wedderburn, ils se sont imposés comme la norme dans le domaine de l'assurance pour développer des modèles de tarification, remplaçant progressivement les modèles de régressions linéaires simples utilisés jusqu'alors.

Cette utilisation des GLM a permis la description de distributions non gaussiennes et de comportements non linéaires, apportant ainsi une solution pour les distributions de coûts et de fréquences de sinistres qui suivent rarement des distributions gaussiennes. Néanmoins, ils font face à la concurrence montante d'autres outils basés sur la *Data Science*, capables eux aussi, selon la littérature, de permettre l'élaboration de modèles de tarification efficaces.

Un engouement pour ce type de modèles a commencé au début des années 2010 dans des domaines divers et variés, de la biologie au monde actuariel. Cet enthousiasme a alors entraîné une vulgarisation des différentes méthodes. Des stratégies d'application et d'interprétation ont été développées.

Pourtant, cette popularité ne s'est pas traduite par un remplacement systématique des modèles classiques, car leur développement s'est vu limité par le phénomène de *Black Box* : les résultats obtenus sont difficilement exploitables et il est difficile d'expliquer simplement ce qu'il se passe entre l'entrée des données et la sortie des résultats. Ainsi, ces modèles restent encore peu déployés par les équipes opérationnelles de tarification.

Cette étude, produite au sein de l'ÉQUITÉ, filiale dédiée à la gestion des partenariats de GENERALI France, est consacrée à la comparaison de modèles pour la tarification en assurance MRH. Elle va s'appuyer à la fois sur des modèles couramment utilisés pour ce type d'analyse, les GLM (*Generalized Linear Models* ou Modèles Linéaires Généralisés) et sur des méthodes plus récentes de type *machine learning*.

Le corps de ce mémoire sera donc constitué d'une mise en concurrence de différents modèles de tarification afin de déterminer s'il existe une méthode alternative plus efficace pouvant challenger les GLM. Le but recherché par la direction de l'ÉQUITÉ est d'utiliser les résultats obtenus comme base de travail pour moderniser leur tarification existante et tenter d'en corriger les possibles biais de tarification.

Cette étude se concentrera sur les données de sept partenaires de l'ÉQUITÉ et sur les garanties principales des contrats MRH : Dégâts des Eaux, Bris de Glace, Incendie, Vol et Responsabilité Civile.

Le premier chapitre sera consacré à la présentation du contexte de l'étude, notamment le marché de l'assurance multirisque habitation et la particularité liée à l'exécution de ce mémoire dans l'environnement de la direction des Partenariats de l'ÉQUITÉ.

Le second chapitre abordera le contexte de la tarification ainsi que les différents modèles qui seront utilisés dans ce mémoire. Seront donc présentés les GLM ainsi que les modèles de *Data Science* : les arbres de régression, les forêts aléatoires et l'*Extreme Gradient Boosting*.

Le chapitre 3 présentera les données du portefeuille, les traitements effectués dans l'optique de la tarification et la corrélation entre les variables.

Les chapitres 4 et 5 se concentreront sur l'application et la comparaison des méthodes pré-citées, pour la modélisation de la fréquence et du coût moyen sur les garanties étudiées. L'ensemble de ces comparaisons est basé sur l'étude d'indicateurs de performance, principalement la MSE et l'indice de Gini (qui est lié à la courbe de Lorenz).

Le chapitre 6 sera dédié à la comparaison entre les nouveaux tarifs modélisés et les tarifs en vigueur, au global et par garantie. Ces comparaisons permettront l'analyse des écarts entre ces tarifs et la détection d'éventuels biais du tarif actuel, qui pourraient ainsi être corrigés.

Chapitre 1

Contexte de l'étude

Ce chapitre a pour but de se familiariser avec le contexte de l'étude en présentant le marché MRH, les garanties qui s'y rattachent et la particularité du *business model* des Partenariats.

1.1 L'assurance MRH

Les assurances IARD (Incendie, Accident et Risques Divers) regroupent les assurances de biens et de responsabilité. Ce qui s'oppose aux assurances de personnes qui sont composées des assurances santé et des assurances vie.

Le sujet de l'étude va porter sur l'assurance Multirisque Habitation qui est un des produits phares des offres d'assurance IARD. Ce type d'assurance a pour dessein de couvrir une habitation et son mobilier en cas de survenue d'un sinistre, que la responsabilité de l'assuré soit mise en cause ou pas, mais aussi la responsabilité civile des occupants envers un tiers.

1.1.1 Les garanties

Un contrat MRH regroupe des garanties qui peuvent être de deux types : des garanties dommages aux biens et des garanties de responsabilité civile.

Différentes garanties peuvent être présentes dans un contrat MRH, que l'assuré soit propriétaire ou locataire. Ainsi, les garanties suivantes peuvent être citées :

- La garantie **dégât des eaux (DDE)** : cette garantie couvre les dommages sur l'habitation assurée mais aussi les conséquences de ces dégâts sur les habitations voisines. Les dégâts des eaux comprennent notamment les fuites d'eau, les infiltrations d'eau, les dégâts dus au gel, les ruptures de conduite, etc... Attention, l'indemnisation du dégât des eaux ne prend pas en compte la réparation de la construction ou de l'appareil qui est à l'origine du dommage.
- La garantie **incendie** : cette garantie assure le logement et les biens personnels contre le feu et la fumée liés à une combustion, un embrasement ou une conflagration ainsi que les dommages corporels éventuels et rembourse aussi les dégâts occasionnés par l'intervention des pompiers. Attention, cela veut dire que les incendies liés à la foudre seront pris en compte dans cette garantie. De plus, pour que l'incendie soit pris en charge par l'assureur, il doit pouvoir être qualifié d'accidentel ou de criminel.
- La garantie **bris de glace (BDG)** : cette garantie concerne la protection de tout élément en verre qui forme une séparation avec l'extérieur.
- La garantie **vol** : la garantie vol couvre les conséquences d'un vol ou d'un cambriolage. Cette garantie s'applique aux biens de toutes les personnes vivant au domicile assuré et s'étend aux objets prêtés ou loués.

- La garantie **catastrophes naturelles** : cette garantie prend en charge les dégâts dus aux catastrophes naturelles (tremblement de terre, inondation, sécheresse, glissement de terrain, . . .).
- La garantie **évènements climatiques** : cette garantie couvre le bien contre les dommages résultant du vent, de la grêle et de la neige. Elle doit être distincte de la garantie catastrophes naturelles. En effet, cela correspond à des dégâts liés à de grandes intempéries mais dont le niveau de gravité ne permet pas de déclarer l'état de catastrophe naturelle. Elle est obligatoire depuis juin 1990.
- La garantie **attentats** : cette garantie couvre les biens face à des dommages qui pourraient être causés par un attentat, un sabotage, une émeute, un acte de terrorisme ou un mouvement populaire. Elle est obligatoire depuis septembre 1986.
- La garantie **responsabilité civile vie privée (RC)** : la garantie RC vie privée permet de protéger l'assuré et ses proches de dommages accidentels causés sur autrui dans la vie quotidienne.

Les garanties citées sont les garanties de base en MRH. Certaines garanties optionnelles peuvent être ajoutées dans les contrats. Elles peuvent être, par exemple, la garantie aménagements extérieurs, la garantie piscine, l'assurance voyage, la garantie perte et vols de clefs.

1.1.2 Le marché français

1.1.2.1 En globalité

Le dernier bilan 2021 de l'assurance française, présenté par France Assureurs, témoigne d'un vif rebond du secteur après le "trou d'air" de 2020 lié aux conséquences de la crise sanitaire. Malgré cette embellie, les assureurs continuent de surveiller de près les tendances de fond (toutes catégories assurantielles confondues), qui s'inscrivent dans un nouveau contexte macro-économique.

Ce contexte 2021, caractérisé par une croissance de 7% du PIB, une progression de près de 30% du CAC 40, mais aussi une forte inflexion sur le front des taux d'intérêt qui remontent après deux décennies de baisse, a mécaniquement boosté la performance du secteur assurantiel dans son ensemble.

Même si les pourcentages de progression de chiffre d'affaire 2021 bénéficient d'un effet d'optique favorable du fait de la crise de 2020, ils traduisent également des hausses observées dans toutes les branches par rapport à l'activité de 2019.(Figure 1.1)

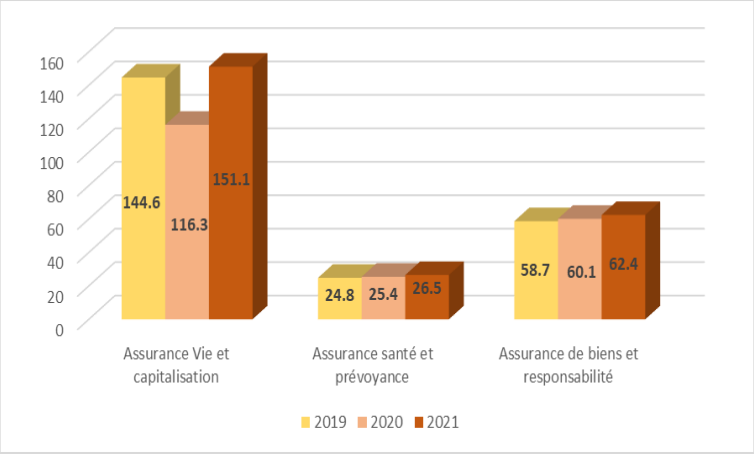

Figure 1.1 – Évolution du chiffre d'affaire annuel par branche, en Mds d'€

Cependant, les assureurs considèrent ce rebond comme fragile avec la nouvelle donne 2022 qui ajoute de nouvelles crises (guerre en Ukraine et inflation galopante) aux anciennes toujours présentes.

Malgré ces crises, la solvabilité des assureurs français reste solide. Les ratios de solvabilité sont en hausse en 2021 et se rapprochent de ceux de 2019 après une légère dégradation en 2020.(Figure 1.2)

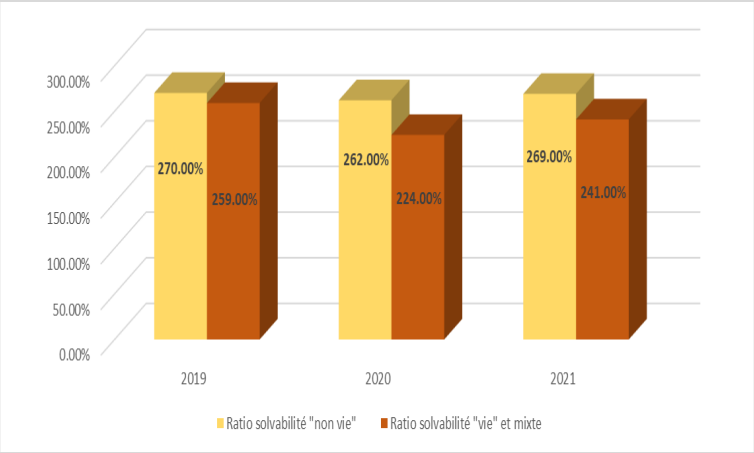

FIGURE 1.2 – Ratios de solvabilité des assureurs français

Dans le marché français de l'assurance, les 26 branches officiellement répertoriées dans l'article R321-1 du Code des assurances, peuvent être regroupées au sein de catégories plus vastes fondées sur deux critères précis : le mode de gestion des primes et le principe d'indemnisation des sinistres. Le mode de gestion des primes permet d'établir une distinction entre "assurances IARD" (aussi appelées "assurances non-vie") et "assurances vie".

En 2021, la part de chiffre d'affaire de l'assurance IARD s'établi à 37,04% du total contre 62,96% pour l'assurance vie, ratio en très légère hausse par rapport aux chiffres de 2019, ceux de 2020 ayant vu la part "vie " s'effondrer sous les effets de la crise sanitaire.(Figure 1.3)

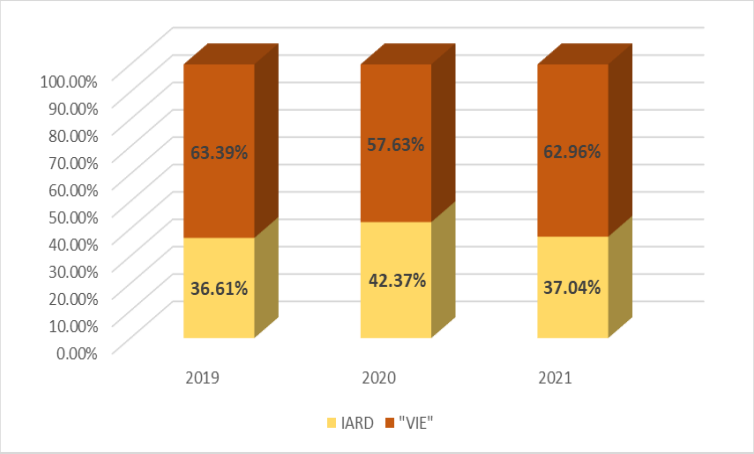

FIGURE 1.3 – Parts de chiffre d'affaire entre IARD et "Vie" en pourcentage

Concernant l'assurance de biens et de responsabilité dont fait partie la MRH, elle représente en chiffre d'affaire, pour 2021, 70,19% du chiffre d'affaire total de l'IARD. Ce chiffre reste quasiment le même en 2020 et 2019.(Figure 1.4 et 1.5)

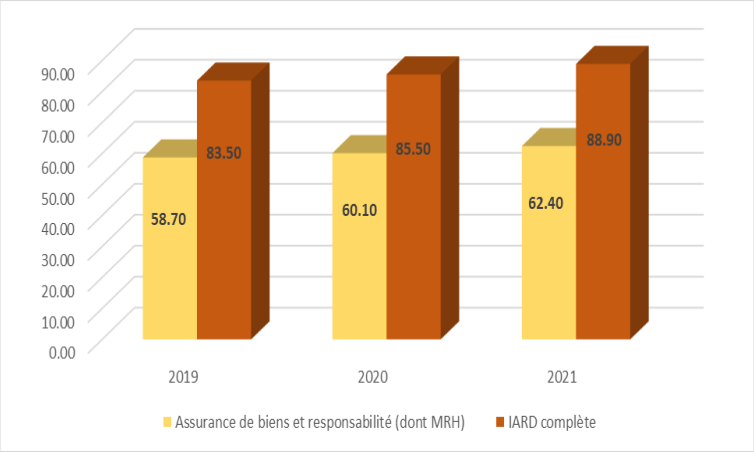

FIGURE 1.4 – Part de l'assurance de biens et responsabilité (dont MRH) et part IARD complète, en mds d'€

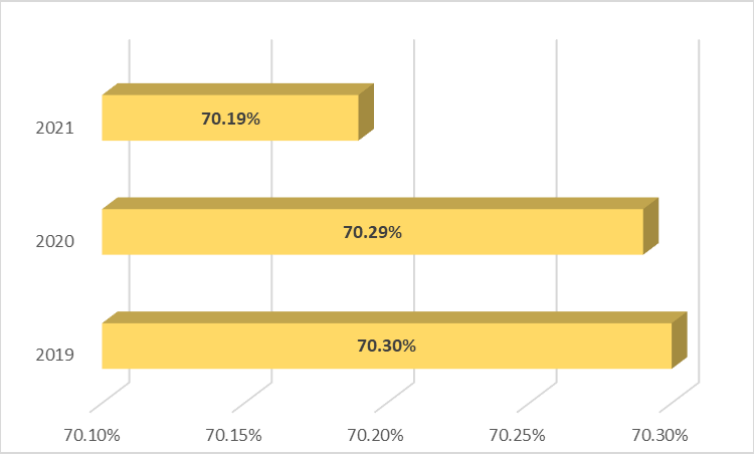

FIGURE 1.5 – Évolution annuelle du chiffre d'affaire de la MRH dans l'IARD, en pourcentage

1.1.2.2 Le cas de la MRH

En 2021, le marché de l'assurance habitation rebondit avec un nombre de contrats en hausse de 2,2% par rapport à 2020, soit 44,4 millions de contrats, avec une progression de 1,8% pour les contrats occupants et de 3,8% pour les contrats non occupants, pour un nombre de logements estimé à 37,6 millions fin 2021.

Une croissance de l'ensemble des cotisations des contrats MRH, toutes catégories confondues, de 3,6% pour atteindre 11 700 millions d'euros peut être observée. Cette croissance confirme celle de 2020 qui atteignait déjà 3,2% de hausse par rapport à 2019 et une somme de 11 276 millions d'euros (10 930 millions d'euros en 2019).

La prime moyenne est également en progression de 1,4% (263 euros HT) par rapport à celle de 2020 (260 euros HT) qui confirmait déjà une tendance haussière à 1,7% par rapport à 2019 (255 euros HT).

En parallèle, la sinistralité du marché de la MRH est elle aussi en forte hausse sur 2021, se rapprochant de son niveau d'avant la crise sanitaire. Par rapport à 2020, le nombre de sinistres progresse de 6% et le montant des charges de prestations de 13%. Cette forte hausse des charges de prestations provient plus de l'augmentation des coûts moyens (de 6,3%) que de la hausse de la fréquence des sinistres (4%). Si l'on compare à 2019, la fréquence est désormais légèrement inférieure (-1,5%), mais le coût moyen reste quasi-stable par rapport à 2019 (+0,8%). (Figure 1.6)

	2020	2021	Variation
Fréquence MRH *	**88,4 ‰**	**91,9 ‰**	**+4,0%**
Dont Incendie	3,98 ‰	4,01 ‰	+0,6%
Dont Tempête, Grêle, Neige	10,2 ‰	7,9 ‰	-22,6%
Dont Vol	7,4 ‰	7,3 ‰	-1,6%
Dont Dégâts des eaux	34,2 ‰	38,5 ‰	+12,6%
Dont RC	9,1 ‰	9,2 ‰	+1,1%
Dont Bris de glaces	7,1 ‰	7,2 ‰	+1,1%
Coût moyen MRH *	**1 574 €**	**1 672 €**	**+6,3%**
Dont Incendie	8 399 €	9 768 €	+16,3%
Dont Tempête, Grêle, Neige	2 788 €	3 524 €	+26,4%
Dont Vol	1 838 €	1 818 €	-1,1%
Dont Dégâts des eaux	1 128 €	1 138 €	+0,9%
Dont RC	1 176 €	1 254 €	+6,6%
Dont Bris de glaces	469 €	481 €	+2,4%

** Y compris les Catastrophes naturelles*

FIGURE 1.6 – Variations fréquence et coût moyen MRH par garantie

Concernant les taux de résiliation et d'affaires nouvelles des contrats MRH, l'année 2021 montre une hausse combinée liée à la reprise de l'activité des différents secteurs après le ralentissement de l'année 2020. Cette évolution est résumée dans les graphiques ci-dessous, sur les 9 dernières années. (Figure 1.7)

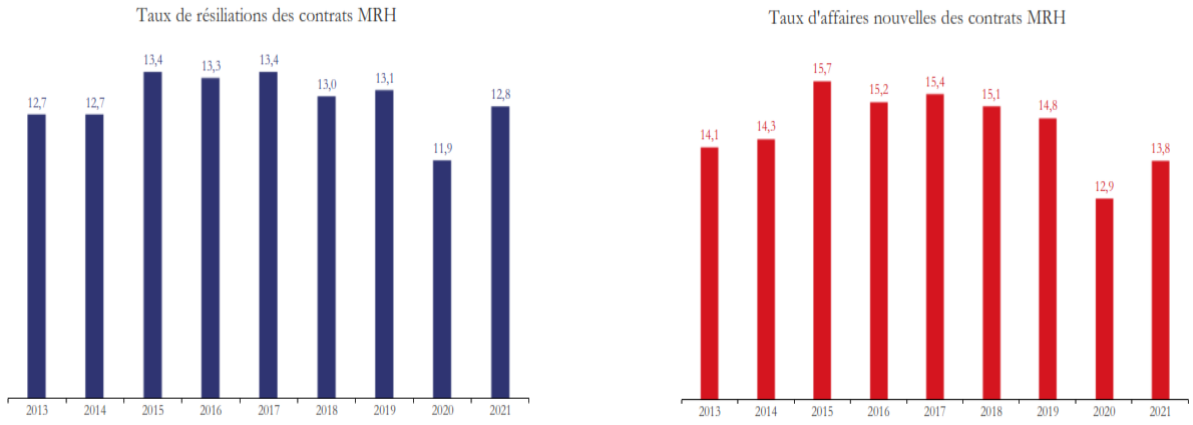

FIGURE 1.7 – Taux moyens de résiliations et d'affaires nouvelles des contrats MRH

Ainsi, le taux de couverture (ratio du taux d'affaires nouvelles sur taux de résiliations) s'établit à 107,4% en 2021, en recul de 0,4 points de pourcentage par rapport à 2020.

1.2 La direction des Partenariats

L'ÉQUITÉ est une filiale de GENERALI France spécialisée dans les partenariats. Les partenaires sont accompagnés dans la mise en place de programmes d'assurance spécifiques à destination de particuliers, pouvant aller de l'assurance dommages à l'assurance prévoyance accident et santé en passant par la protection juridique.

Cette direction est atypique sur le marché par l'accompagnement sur mesure pour chaque partenaire, mais aussi par la diversité des partenaires. En effet, ceux-ci peuvent être des professionnels de la distribution d'assurances

(courtiers, courtiers grossistes,...), des mutuelles ou des instituts de prévoyance mais aussi des grands comptes et enseignes agissant dans l'intérêt de leurs clients.

Le principe de délégation est ce qui caractérise la direction des Partenariats. Ainsi la gestion des contrats et des sinistres est à la charge des partenaires tandis que l'ÉQUITÉ a pour missions l'étude de nouveaux partenariats, l'élaboration de solutions techniques ainsi que la gestion du partenariat (qui peut comprendre la maintenance des documents contractuels ou la mise en place de suivi des partenariats).

Dans le cadre d'une refonte de cette filiale, une modernisation du segment MRH a été entreprise. Pour y parvenir, un nouveau zonier personnalisé a été créé spécifiquement pour l'ÉQUITÉ afin d'éviter l'utilisation d'un ancien zonier de GENERALI.

Le but de ce mémoire sera donc de comparer des modèles pour la tarification en assurance MRH à partir des données issues de ce nouveau zonier pour s'adapter à la clientèle spécifique de l'ÉQUITÉ.

Chapitre 2

Aspects théoriques pour la tarification

Ce chapitre va aborder les différents aspects théoriques qui seront nécessaires pour la compréhension de l'étude et, notamment, la présentation de plusieurs types de modèles permettant d'estimer la fréquence et les charges sinistres pour les sinistres attritionnels.

2.1 Les enjeux de la tarification

2.1.1 Le principe de la mutualisation

En assurance non-vie, une police d'assurance peut être définie comme un contrat entre deux parties, l'assuré qui est détenteur du contrat et l'assureur qui en est le pourvoyeur. L'assureur s'engage à indemniser tout ou partie des pertes potentielles réalisées sur une période précisée dans le contrat en échange du versement d'une prime par l'assuré.

Ainsi, un assuré peut se protéger contre les effets néfastes que pourrait avoir la survenance d'un sinistre élevé. Ce coût qui serait généralement trop important pour être supporté uniquement par l'assuré va être allégé par le montant remboursé par l'assureur. La prime d'assurance qui est payée par l'assuré est habituellement plus faible que le montant du préjudice causé par le sinistre. Le risque économique que portait initialement l'assuré a donc été transféré vers l'assureur.

L'élément clé qui permet le fonctionnement de l'assurance est la **mutualisation** induite par la couverture de nombreux risques similaires par la compagnie. Le coût des sinistres qui serait difficilement supportable par un assuré va être partagé par l'ensemble des souscripteurs à l'assurance. Comme les contrats sont à-priori indépendants les uns des autres et que la tarification se fait par garantie, cette mutualisation permet l'utilisation de la loi forte des grands nombres (LFGN).

Pour rappel, en supposant un portefeuille d'assurance contenant N polices avec une perte annuelle notée S_i pour le contrat i, alors la LFGN stipule :

$$\frac{1}{N}\sum_{i=1}^{N} S_i \underset{N\to+\infty}{\overset{p.s}{\longrightarrow}} \mathbb{E}[S_i] = \mu \tag{2.1}$$

Pour simplifier, la LFGN signifie que la moyenne empirique des pertes annuelles doit converger presque sûrement vers l'espérance de la loi de perte annuelle μ (qui est commune pour tous les contrats puisque les S_i sont indépendants, identiquement distribués (i.i.d)).

Ce résultat sera à la base du principe général de la tarification. En effet, la **prime pure** des contrats sera égale au μ obtenu par modélisation. Cette prime pure sera calculée par l'actuaire de manière à prévoir le montant de la charge future qui est aléatoire à cause de l'inversion du cycle de production de l'assurance. L'assureur pourra, en théorie, grâce à cette prime pure, exactement couvrir ses engagements vis-à-vis des assurés.

Il est important de noter que, théoriquement, la ruine est certaine à l'horizon infini dès lors que la prime pure estimée par l'assureur respecte cette condition.

Ainsi, plusieurs éléments seront ajoutés pour obtenir la **prime commerciale**, c'est-à-dire la prime qui sera demandée aux individus assurés. Ces éléments peuvent être notamment des taxes, des frais de l'assureur, des commissions ou des coûts liés à la réassurance.

Le calcul de ces charges à ajouter à la prime pure estimée varie selon les règles et consignes inhérentes à chaque compagnie d'assurance.

2.1.2 Le principe de la segmentation

Le principe de la segmentation est de permettre aux assureurs de grouper les assurés selon certains critères utiles pour la tarification et/ou de déterminer les garanties offertes. Cette segmentation sera spécifique au profil du portefeuille de l'assureur et donc des risques auxquels il est soumis. Ainsi, chaque groupe présente un risque similaire pour la compagnie et une mutualisation s'appliquera sur celui-ci.

La segmentation est un enjeu important lors de la tarification puisque cette étape permet d'attirer les "bons risques".

Étudions cet exemple pour constater ce phénomène : supposons qu'il y ai deux assureurs, notés A1 et A2. Le premier n'utilise pas de segmentation alors que le deuxième effectue cette étape lors de la tarification.

Les assurés de A1 paieront la même prime pure égale à l'espérance mathématique des charges annuelles futures. La prime commerciale sera donc la même pour tous les clients de cette assurance. Dans ce cas-là, les clients représentant un risque plus faible de sinistres permettront de faire un bénéfice alors que ceux avec le risque le plus fort entraîneront une perte pour l'assureur. Ainsi, les assurés de la première catégorie paieront plus que ce qu'ils représentent de risque tandis que les assurés à fort risque se retrouveront avantagés.

Dans le cas de l'assureur A2, les profils à risque faible paieront une prime plus faible que celle des profils plus risqués. Ainsi les assurés considérés comme des "bons risques" auront un avantage certain à prendre cet assureur et les "mauvais risques" iront vers l'assureur A1 où ils paieront moins. L'assureur A2 dégagera donc des bénéfices tandis qu'une perte sera probable pour l'assureur A1 qui se retrouvera avec l'intégralité des "mauvais risques".

En conclusion, la segmentation utilisée par l'assureur A2 a l'avantage d'attirer les profils avec des risques plus avantageux pour lui tout en tarifant une prime commerciale qui est plus adaptée au profil de l'assuré.

Ainsi un arbitrage doit être effectué lors de la segmentation. Cet arbitrage se fera entre :

- une segmentation grossière avec peu de tarifs différents
- une segmentation trop précise avec beaucoup de groupes de profil de risque et donc des tarifs plus personnalisés. Ce type de segmentation risquerait d'entraîner une discrimination importante des assurés ayant les risques les plus élevés.

Cet arbitrage est une question essentielle lors de la création d'un tarif pour les raisons suivantes :

- le principe de mutualisation pourrait être remis en cause en segmentant trop le portefeuille. En effet, la LFGN n'est applicable que sur une classe homogène suffisamment grande.
- Augmenter la segmentation ne permet pas forcement de diminuer le tarif des assurés car la prime d'incertitude augmente puisqu'elle est liée à l'incertitude des estimateurs. Le tarif augmenterait donc alors que l'un des objectifs de la segmentation est de diminuer le tarif pour certains risques.

Pour éviter de dépasser la segmentation et d'entrer dans la discrimination, les classes de risque doivent respecter certains principes :

- Légitime : obligation de respecter l'intérêt général
- Objective et pertinente : la mise en œuvre doit être justifiée scientifiquement
- Nécessaire : pouvoir justifier d'aucune autre alternative plus rentable
- Proportionnelle : stabilité entre l'intérêt pour l'assureur et le préjudice pour l'assuré

2.1.3 L'arbitrage entre mutualisation et segmentation

La mutualisation est applicable dès lors que le volume de données est suffisamment important et qu'il permet de répartir le coût des sinistres entre les éléments d'un groupe de risque. La segmentation, elle, permet d'établir des tarifs concurrentiels. Ainsi, un équilibre doit être trouvé entre ces deux principes pour garantir un tarif compétitif mais aussi un contrôle du risque.

Une analyse devra donc être effectuée pour établir des classes homogènes de risque (par la segmentation) puis obtenir un tarif pour chacune des classes. L'intérêt sera de pouvoir cibler les segments avec des risques plus élevés et mettre en place une gestion adaptée.

2.2 Le modèle fréquence-coût

Pour déterminer un tarif, selon le modèle classique de prime pure, il faut estimer la charge annuelle d'un contrat i, notée S_i.

Cette charge peut être définie de la manière suivante :

$$S_i = \begin{cases} \sum_{k=1}^{N_i} Y_{i,k} & \text{si } N_i \geq 1 \\ 0 & \text{sinon.} \end{cases} \tag{2.2}$$

où N_i est une variable aléatoire représentant le nombre de sinistres de l'individu i (qui se sont produits au cours de l'année) et les $Y_{i,k}$ sont des variables aléatoires symbolisant le montant du k^{ime} sinistre de l'individu i.

Dans ce modèle, plusieurs hypothèses sont constituées :

- Les suites $(Y_{i,k})_{1\leq k\leq N_i}$ sont supposées i.i.d.
- Le montant des sinistres est indépendant du nombre de sinistres N_i, c'est-à-dire

$$\forall k \in \{1, ..., N_i\} \ , \quad Y_{i,k} \perp\!\!\!\perp N_i$$

Ainsi, pour un individu, la formule de l'espérance sera :

$$\mathbb{E}[S_i] = \mathbb{E}\left[\mathbb{E}[S_i|N_i]\right] = \sum_{j=1}^{+\infty} \mathbb{P}(N_i = j) \times \mathbb{E}\left[\sum_{k=1}^{j} Y_{i,k}\right] \tag{2.3}$$

$$= \mathbb{E}[Y_{i,1}] \left(\sum_{j=1}^{+\infty} \mathbb{P}(N_i = j) \times j\right) \tag{2.4}$$

$$= \mathbb{E}[N_i] \times \mathbb{E}[Y_{i,1}] \tag{2.5}$$

Dans ce cas, la prime pure pour un individu correspondra au nombre moyen de sinistres multiplié par le coût moyen de ces sinistres.

En construisant un modèle pour la sévérité et un autre pour la fréquence, il y a un risque de propagation de l'erreur d'estimation. Cependant, utiliser un modèle qui ne s'intéresserait qu'à la charge annuelle du contrat engendrerait une perte d'information sur le nombre de sinistres et la masse en 0 qui viendrait des contrats non sinistrés entraînerait une difficulté de calibration des GLM. Ainsi, cette décomposition fréquence-coût est utilisée en pratique pour tarifer les contrats d'assurance non-vie.

2.3 Les modèles linéaires généralisés

Les modèles linéaires généralisés sont les outils les plus fréquemment utilisés en tarification non-vie pour estimer la prime pure. Ces modèles sont une généralisation des modèles de régressions linéaires simples souvent utilisés en économétrie.

Dans cette partie, un rappel de leur utilisation va être développé.

2.3.1 Les composantes d'un GLM

Il existe trois composantes dans un GLM (pour un individu i) :

- La **valeur réponse** notée Y_i qui est la composante aléatoire.
- Le **prédicteur** $\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j X_{i,j}$ qui est linéaire et déterministe.
  J représente le nombre de variables explicatives utilisées dans le modèle.
- La **fonction de lien** g qui est monotone, dérivable et inversible.

Composante aléatoire

Cette composante aléatoire correspond à la loi de la réponse aléatoire Y qui est à déterminer à partir des variables explicatives. Cette variable Y est de la même loi que tous les Y_i qui sont i.i.d (c'est-à-dire que tous les individus qui sont indépendants ont la même loi de distribution pour leur variable à expliquer).

Le but d'un GLM sera de spécifier cette loi en prenant en compte l'hypothèse, classique dans l'étude de GLM, qu'elle appartient à la famille exponentielle. Cette condition d'appartenir à une famille de loi spécifique est une amélioration par rapport aux modèles linéaires simples où la variable modélisée ne pouvait être que gaussienne.

Pour rappel, une variable est de la famille exponentielle si sa densité peut s'écrire de la manière suivante :

$$f_{\theta,\phi}(y) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{2.6}$$

avec :

- a une fonction définie sur $\mathbb{R}$ non nulle
- b une fonction définie sur $\mathbb{R}$ deux fois dérivable et de dérivé première injective
- c une fonction définie sur $\mathbb{R}^2$
- θ le paramètre canonique qui est inconnu
- ϕ le paramètre de dispersion qui est supposé connu

L'espérance et la variance de telles fonctions sont :

- $\mathbb{E}[Y] = b^{'}(\theta)$
- $\mathbb{V}[Y] = a(\phi) \times b^{''}(\theta)$

Composante déterministe

Le prédicteur qui est la composante déterministe, correspond au produit entre la matrice des variables explicatives X et le vecteur des coefficients linéaires β. C'est ce vecteur qui sera déterminé dans l'approche par GLM.

Il est important de noter que, dans la théorie des GLM, les variables explicatives contenues dans X doivent être non corrélées deux à deux.

Les différentes composantes du vecteur β peuvent s'interpréter de la manière suivante.

Tout d'abord, β_0, appelé l'intercept, représente la valeur de la variable à expliquer pour l'individu de référence. Cet individu de référence correspond à un profil d'assuré composé de toutes les modalités de référence des variables explicatives. Les modalités de référence ont été choisies en prenant la modalité la plus fréquente pour chacune des variables explicatives qualitatives.

Enfin, pour les autres β_j, un $\beta_j > 0$ aura tendance, pour la variable considérée, à entraîner une hausse de la valeur de la variable à expliquer par rapport à l'individu de référence et inversement, un $\beta_j < 0$ aurait pour conséquence une baisse de la valeur de la variable à expliquer, pour l'individu avec cette modalité, par rapport à l'individu de référence.

La fonction de lien

La fonction de lien vérifie la relation suivante :

$$g(\mathbb{E}[Y_i|X_i]) = \eta_i \tag{2.7}$$

Dans le modèle linéaire simple, la fonction de lien g sera la fonction identité. Il existe d'autres fonctions de lien qui pourront être utilisées en fonction de la loi que suit la variable à expliquer.

Loi usuelle de distribution	Fonction lien
Normale	$g(y) = y$
Poisson	$g(y) = log(y)$
Binomiale	$g(y) = log\left(\frac{y}{1-y}\right)$
Gamma	$g(y) = \frac{1}{y}$
Inverse Gaussienne	$g(y) = \frac{1}{y^2}$

TABLE 2.1 – Fonction Lien pour les lois usuelles

Cette fonction lien illustre le fait que l'espérance de la variable à expliquer est basée, à une transformation près, sur le prédicteur (qui dépend des variables explicatives).

Un autre élément, appelé offset, peut être ajouté au modèle. Cet offset est une constante qui vient modifier le risque de base (qui n'est pas lié au profil spécifique d'un assuré) et représente une sorte d'exposition. Concrètement, c'est un terme commun à tous les individus mais dont la valeur sera différente selon l'individu considéré.

L'équation du GLM devient donc, en ajoutant cet offset :

$$g\left(\mathbb{E}\left[Y_i|X_i\right]\right) = offset + \eta_i \tag{2.8}$$

Le coefficient de cet offset est contraint à être égal à 1. C'est pour cette raison que pour la calibration, la régression est $g\left(\mathbb{E}\left[Y_i|X_i\right]\right) - offset = \eta_i$.

Ainsi les étapes suivies lors de la mise en place du GLM sont :

- Le choix de la loi pour la variable à expliquer Y parmi les lois de la famille exponentielle en fonction de ces caractéristiques
- Le choix de la fonction lien. En tarification, la fonction log sera privilégiée puisqu'elle permet d'avoir un modèle multiplicatif et donc des coefficients correcteurs pour le calcul de la prime.
- L'estimation du vecteur β
- La validation et l'application du modèle

2.3.2 Estimations des paramètres

Pour rappel de la partie précédente, l'estimation du modèle vient de l'appréciation du vecteur β par maximum de vraisemblance.

Cette estimation sera faite de manière à maximiser la fonction de log-vraisemblance.

$$ln(\mathcal{L}) = \sum_{i=1}^{N} ln(f_{\theta,\phi}(y_i)) \tag{2.9}$$

$$= \sum_{i=1}^{N} \underbrace{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)}_{l_i(\theta_i)} \tag{2.10}$$

Pour obtenir l'estimateur du maximum de vraisemblance de β_j, il faut avoir $\forall j \in 1, ..., J$:

$$\frac{\partial ln(\mathcal{L})}{\partial \beta_j} = 0 \tag{2.11}$$

Or :

$$\frac{\partial ln(\mathcal{L})}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i(\theta_i)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} \tag{2.12}$$

D'une part, il est possible de calculer $\frac{\partial l_i(\theta_i)}{\partial \theta_i}$:

$$\frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{y_i - b^{'}(\theta_i)}{a(\phi)} \tag{2.13}$$

D'autre part, par la définition de η_i, la réécriture se traduit par :

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}X_{i,j} \tag{2.14}$$

Il est aussi important de remarquer que :

$$\frac{\partial \eta_i}{\partial \theta_i} = \left(\frac{\partial \theta_i}{\partial \eta_i}\right)^{-1} \tag{2.15}$$

En posant $\mu_i = g^{-1}(\eta_i)$, l'équation précédente permet de déduire :

$$\frac{\partial \eta_i}{\partial \theta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = (g'(\mu_i)b''(\theta_i))^{-1} \tag{2.16}$$

Ainsi, ces deux remarques permettent d'écrire :

$$\frac{\partial ln(\mathcal{L})}{\partial \beta_j} = \frac{1}{a(\phi)}\sum_{i=1}^{N} \frac{(y_i - b^{'}(\theta_i))X_{i,j}}{g'(\mu_i)b''(\theta_i)} \tag{2.17}$$

En posant D la matrice diagonale composée des éléments $\frac{1}{g'(\mu_i)b''(\theta_i)}$, alors (2.11) est vérifiée si :

$$X^{'}D(y - g^{-1}(X\beta)) = 0 \tag{2.18}$$

Ainsi, l'estimateur de maximum de vraisemblance (EMV) $\hat{\beta}$ confirme (2.18).

Remarque :

- En posant $g(x) = x$, l'estimateur dans le cadre du modèle linéaire simple est bien :

$$\hat{\beta} = \left(X^{'}X\right)^{-1} X^{'}Y \tag{2.19}$$

- En pratique, l'équation 2.18 ne permet pas d'obtenir une expression explicite de $\hat{\beta}$. Une méthode qui est alors utilisée pour estimer ces valeurs pourrait être l'algorithme de Newton-Raphson (cf. annexe B).

2.3.3 Les tests de significativité

Dans les GLM, les coefficients obtenus doivent être testés pour vérifier leur significativité, c'est-à-dire déterminer si la variable concernée explique correctement la variable Y. Pour cela, le test de significativité de Wald [1] est utilisé.

Les hypothèses testées sont :

$$H_0 : \ \beta_j = 0 \ \ vs \ \ H_1 : \beta_j \neq 0 \tag{2.20}$$

La statistique de test est $\left(\frac{\hat{\beta_j}}{\mathbb{V}[\hat{\beta_j}]}\right)^2$ et suit une loi du χ^2 de degré 1. Ainsi, l'hypothèse H_0 est rejetée dès que la statistique est supérieure au fractile d'ordre $1-\alpha$ de la loi du χ^2 à 1 degré de liberté.

La loi asymptotique vient du fait que l'estimateur de maximum de vraisemblance est asymptotiquement gaussien, c'est-à-dire que :

$$\hat{\beta_j} \ \sim \ \mathcal{N}\left(\beta_j, \frac{1}{I(\beta_j)}\right) \tag{2.21}$$

où I est l'information de Fisher qui est définie par :

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} log(f(Y,\theta))\right)^2 |\theta\right] \tag{2.22}$$

La variance de l'estimateur peut devenir trop grande si l'information de Fischer est petite.

La technique consiste ainsi à regrouper certaines des modalités de variables qualitatives afin d'obtenir un modèle plus satisfaisant (c'est-à-dire avec moins de variance de l'estimateur) :

- Calibrer un modèle avec toutes les variables pour la modélisation
- Effectuer le test de significativité pour toutes les variables explicatives.
- Prendre la variable avec la pire p-value au-dessus du seuil α
- Agréger la modalité avec une autre modalité de la variable
- Calibrer à nouveau le modèle et répéter jusqu'à obtenir un modèle satisfaisant

1. Ce test est un test dit de type III qui peut être effectué pour évaluer la significativité des coefficients du modèle. Ces tests sont basés sur une comparaison de modèles emboîtés.

2.3.4 Avantages et Inconvénients

Les GLM sont une forme généralisée des modèles linéaires simples et ils ont conservé la simplicité d'exécution de ces derniers.

Cependant, une des limites de ce type de modèles est que la procédure n'est efficace que si la loi conditionnelle utilisée fait partie de la famille exponentielle. De plus, le choix de la fonction lien est souvent imposé par rapport à la famille exponentielle choisie.

Enfin, il est important de noter que l'utilisation de la famille exponentielle impose qu'il n'y ai pas de valeurs extrêmes.

2.4 Tree-based Models

Les méthodes traditionnelles de GLM imposent des suppositions sur la forme de l'équation pour estimer la valeur de la variable réponse. Les modèles de *Data Science*, comme les forêts aléatoires (*Random Forest*) ou la méthode *Gradient Boosting*, ne nécessitent pas de telles suppositions. Ces modèles sont souvent considérés comme non-paramétriques.

De plus, les modèles GLM obligent à sélectionner les variables avant la création du modèle. Un tel problème n'existe pas avec les modèles de *Data Science* où les variables sont sélectionnées par le modèle lui-même.

Ainsi, bien que le modèle de référence (le GLM) soit utilisé pour sa simplicité de mise en œuvre, son temps de calcul rapide et son interprétabilité, récemment, les modèles de *Data Science* ont été mis en valeur dans la littérature actuarielle et dans des mémoires d'actuaires.

Pour répondre à l'objectif de ce mémoire, la confrontation de ces modèles plus innovants à ceux plus "classiques" liés au GLM doit être envisagée.

Les modèles de *Data Science*, basés sur la construction d'arbres, vont être explicités dans la partie suivante.

2.4.1 Arbre de régression

Le concept de ce type de modèles peut être introduit par la description des algorithmes *Classification And Regression Tree* (CART). Les arbres CART seront la base des autres modèles de *Data Science* introduits par la suite.

L'objectif de l'utilisation d'un arbre sera, comme pour le GLM, d'expliquer une variable réponse Y en fonction de variables explicatives X_i quantitatives ou qualitatives. Si la variable à expliquer est une variable quantitative alors l'arbre créé sera appelé arbre de régression et si cette variable est qualitative, l'arbre sera appelé arbre de classification. Dans le cas de l'étude, l'objectif est de créer des arbres de régression.

Le but d'un arbre de régression est de séparer l'espace initial (dans notre cas, le portefeuille) qui est hétérogène, en sous-espaces qui seront eux homogènes, en utilisant des règles de séparations qui ne peuvent être que binaires et ne sont basées que sur l'une des variables explicatives du modèle. A la fin du processus, se dégage ainsi un partitionnement binaire récursif.

Pour décrire un arbre, la terminologie utilisée est la suivante :

- La racine : cet élément correspond au portefeuille initial.
- Les branches : ces éléments représentent les règles qui sont utilisées pour la subdivision.
- Les feuilles : elles caractérisent les sous-groupes homogènes à obtenir.

Ainsi, l'arbre débute au niveau du noeud racine. Puis, à chaque étape du processus, les noeuds seront divisés en deux autres noeuds les plus homogènes possibles au sens de la variable à expliquer. Lorsqu'un noeud ne peut plus être divisé, le noeud final, la feuille, est atteint. L'estimation de la quantité recherchée se lit sur cette feuille. Une représentation simplifiée d'un arbre CART est illustrée ci-dessous.(Figure 2.1)

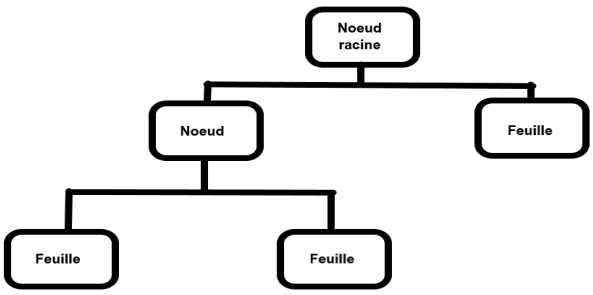

FIGURE 2.1 – Illustration simplifiée d'un arbre CART

Le choix des différentes règles pour la segmentation de l'espace s'effectue de la manière suivante.

Pour chacune des variables explicatives qualitatives des données, des tests pour les partitionnements seront effectués à l'aide d'un critère d'homogénéité. Le partitionnement le plus adapté pour la variable testée sera celui où les sous-espaces créés auront la plus grande homogénéité.

La règle et le partitionnement choisis lors de l'étape seront ceux qui maximisent l'homogénéité globale. Ce processus sera répété jusqu'à obtenir l'arbre maximal.

L'arbre maximal est l'arbre dont les feuilles ne contiennent chacune qu'un individu. Il y aura donc une segmentation maximale du portefeuille, c'est-à-dire un tarif spécifique à chaque individu.

Dans le cas où la variable à expliquer est une variable continue, la variable d'intérêt sera calculée de la manière suivante :

$$\pi_0(x) = \mathbb{E}[Y|X = x] \tag{2.23}$$

Le critère de division qui est utilisé lors des différentes séparations est celui qui entraîne la minimisation de l'erreur quadratique moyenne (*Mean Squared Error* ou MSE). Cette minimisation est mise en forme par l'équation suivante :

$$\pi_0(x) = \underset{\pi(x)}{argmin}\, \mathbb{E}[\Phi(Y, \pi(x))|X = x] \tag{2.24}$$

où $\Phi(Y, \pi(x)) = (Y - \pi(x))^2$ est la fonction de perte.

Ainsi, à chaque noeud, toutes les covariables sont testées avec les différentes règles possibles. Par exemple, si X_i est une variable qualitative, les différentes règles possibles sont $X_i = m$ où m est un ensemble des modalités de la variable X_i. La variable explicative qui sera choisie ainsi que la règle seront celles qui auront minimisé le critère de séparation qui caractérise l'hétérogénéité.

Pour résumer, la construction de l'arbre maximal peut se traduire par les étapes décrites ci-dessous :

1. En partant de la racine, la première segmentation en utilisant le critère de division 2.24 est effectuée.
2. Après la segmentation précédente, l'étape pour le choix de la division est réalisée sur chaque noeud fils.
3. La segmentation sur chaque noeud fils se poursuit.
4. Les deux dernières étapes sont réitérées jusqu'à obtenir l'arbre maximal.

Un risque de surapprentissage survient lorsque le modèle se calibre en prenant trop en compte les données d'apprentissage et donc les éventuelles données aberrantes ou les points exceptionnels. Ainsi, potentiellement, des caractéristiques exceptionnelles ou anormales peuvent être retenues comme normales. En conséquence, le modèle obtenu ne pourra pas être généralisé à une autre base qui sera indépendante de celle d'apprentissage.

Ce surapprentissage surviendrait si l'arbre maximal trouvé était utilisé tel quel. C'est pour cela que l'arbre optimal pour décrire la variable à expliquer pourra être trouvé en effectuant une procédure d'élagage afin de supprimer les feuilles qui n'apporteraient aucune valeur ajoutée à l'estimation du modèle. Ainsi, cet arbre optimal sera un sous-arbre de l'arbre maximal.

Cette étape essentielle pour la création de l'arbre est assez compliquée puisque supprimer un nombre trop important de noeuds pourrait impacter la valeur prédictive du modèle.

La procédure d'élagage consiste à construire une suite de sous-arbres emboîtés, et le sous-arbre optimal sera choisi à partir d'un critère d'optimisation se basant sur un compromis entre la taille de l'arbre et son coût de mauvaise estimation, qui est :

$$C_\alpha = c(T) + \alpha|T| \tag{2.25}$$

où $|T|$ correspond au nombre de feuilles de l'arbre T, α est le paramètre de complexité de l'arbre (c'est-à-dire le coût en terme d'erreur de l'addition d'un noeud dans le modèle) et c(T) l'erreur de régression.

La suite de sous-arbres est construite par le raisonnement suivant :

1. Le point de départ est l'arbre maximal
2. Une première valeur de α est considérée et ainsi un sous-arbre est obtenu.
3. A partir de ce sous-arbre, une valeur de α supérieure à la précédente est choisie et un sous-arbre du sous-arbre est récupéré.
4. La même procédure est réitérée jusqu'à la racine.

Une suite de α croissante est ainsi obtenue. Dans cette liste de sous-arbres optimaux est choisi celui dont le α associé minimise le compromis C_α.

Le problème majeur de l'utilisation d'un arbre CART est son **instabilité**. En effet, la construction d'un arbre optimal peut varier considérablement si la base d'apprentissage est modifiée. Les prédictions de la variable à expliquer pourraient ainsi être différentes.

De plus, la modélisation par un arbre de régression peut amener un biais à cause d'un potentiel **surapprentissage** même si la procédure d'élagage de l'arbre maximal tend à le limiter.

Afin de pallier au manque de robustesse de l'arbre CART, d'autres méthodes ont été mises en place en se basant sur son principe. Deux de ces méthodes sont décrites dans les parties suivantes (*Random Forest* 2.4.2 et *Gradient Boosting* 2.4.3).

2.4.2 Les forêts aléatoires

L'objectif des *Random Forest* (ou forêts aléatoires) est de conserver des avantages des arbres CART tout en éliminant certains des inconvénients notamment le risque de surapprentissage et l'instabilité. L'algorithme pour la création des forêts aléatoires reste robuste et flexible.

Le principe d'une forêt aléatoire est de proposer un estimateur "moyenné" afin d'améliorer la qualité de la quantité d'intérêt estimée. Ainsi, plusieurs estimateurs de qualité individuelle faible seront réunis pour fournir une vision globale plus performante.

Pour illustrer ce principe, en imaginant la construction de B arbres CART en modifiant l'échantillon d'apprentissage à chaque fois et en notant $\hat{Y}_{i,b}$ l'estimateur de la variable à expliquer pour l'individu i, l'estimateur forêt aléatoire obtenu serait (dans le cas où Y est continue) :

$$\hat{Y}_i = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_{i,b} \tag{2.26}$$

Construction de la forêt

Deux stratégies différentes existent pour introduire de l'aléatoire dans la construction des sous-échantillons d'apprentissage utilisés pour la créations des différents arbres.

- ***Bagging***

Les sous-échantillons de la base d'apprentissage sont tirés aléatoirement avec remise. A partir de ce sous-échantillon, un arbre classique CART maximal (sans élagage) est construit.

Attention, il est important de noter que dans le cadre du *Bagging*, le sous-échantillon peut avoir la taille de l'échantillon d'apprentissage de départ. Souvent, pour des raisons de temps de calcul, les sous-échantillons peuvent être moins volumineux.

Cette stratégie permet de limiter la variance de l'estimateur issue de certains individus bien spécifiques.

En ne retenant que cette stratégie pour introduire de l'aléatoire (c'est la méthode de *Bagging*), le principe pour la construction de l'estimateur est résumé par la figure 2.2 :

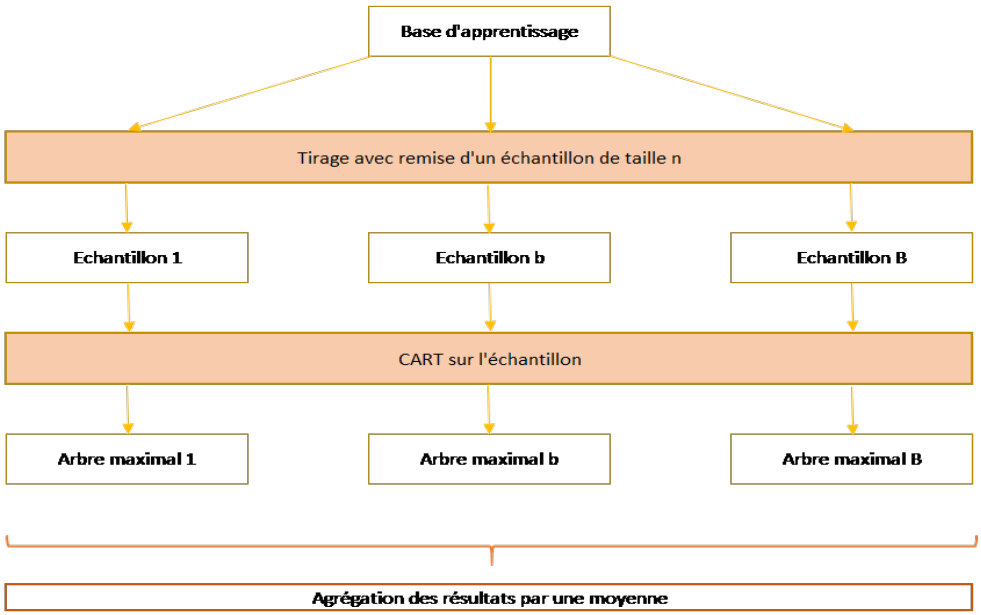

FIGURE 2.2 – Schéma Explicatif du *Bagging*

- **Randomisation des variables explicatives utilisées**

Une autre forme d'aléa est introduite, dans l'algorithme pour les forêts aléatoires (par rapport à la méthode du *Bagging*), par une sélection des variables explicatives qui seront utilisées lors de la segmentation au niveau des noeuds pour la modélisation CART.

Ainsi, dans l'algorithme CART, au moment de la création d'un nouveau noeud et donc de la recherche de la variable explicative qui permettra la division optimale de l'ensemble, seulement m variables sur les J possibles seront testées. Cette méthode aura l'avantage de créer des arbres les plus indépendants possibles.

Cette stratégie permet de limiter la variance de l'estimateur en cas de corrélation entre les covariables.

En effet, la variance d'une moyenne de B estimateurs indépendants de variance σ^2 vaut :

$$Var\left[\frac{1}{B}\sum_{b=1}^{B} Y_b\right] = \frac{1}{B^2} Var\left[\sum_{b=1}^{B} Y_b\right] = \frac{\sigma^2}{B} \tag{2.27}$$

En revanche, la variance d'estimateurs qui seraient corrélés 2 à 2, de coefficient de corrélation ρ, s'exprime de la façon suivante :

$$Var\left[\frac{1}{B}\sum_{b=1}^{B} Y_b\right] = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.28}$$

Si les corrélations entre les estimateurs sont faibles ($\rho \to 0$), alors la variance pourra converger vers 0 lorsque B augmentera. Dans le cas inverse où les corrélations seraient plus importantes, même en augmentant la valeur

du facteur B, la variance serait toujours au minimum égale à $\rho\sigma^2$. Il est rare que les estimateurs utilisées soient indépendants, c'est la raison pour laquelle la randomisation des variables explicatives assure une diminution de la variance de l'estimateur de Y.

Le principe de forêts aléatoires pourrait être illustré de la manière suivante :

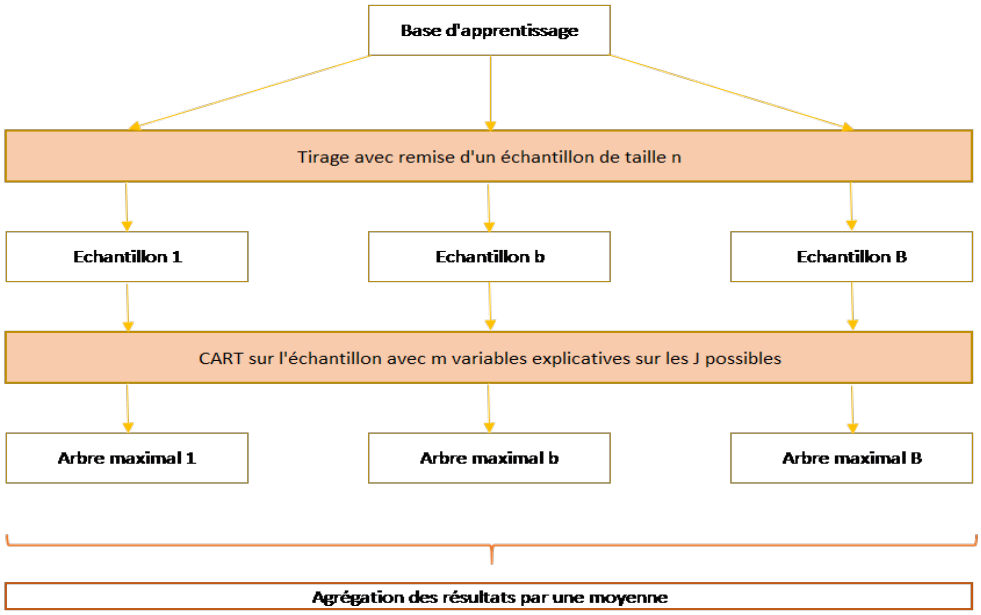

FIGURE 2.3 – Schéma Explicatif du *Random Forest*

L'erreur liée aux forêts aléatoires dépend de deux paramètres :

- la corrélation entre les variables (plus cette corrélation augmente, plus l'erreur est grande)
- la force des arbres de la forêt (capacité à estimer précisément la valeur de Y) (plus l'arbre est précis, moins l'erreur est grande)

Quand le facteur de la randomisation m est important, la corrélation et la force sont augmentées. Ainsi, un arbitrage sur ce facteur devra être effectué afin de pouvoir minimiser l'erreur.

Erreur *Out-of-Bag* et importance des variables

Lors de la réalisation d'un *Random Forest*, il est possible d'estimer aussi l'erreur *Out-of-Bag*. Effectivement, pour chaque arbre CART, seule une partie des données d'apprentissage est utilisée, le reste est considéré comme *Out-of-Bag*. Sur cet échantillon de données, une estimation non biaisée de l'erreur est calculée ainsi qu'une estimation de l'importance des facteurs de risques (les variables explicatives).

Ainsi, l'erreur *Out-of-Bag* permettra, pour chaque individu i, la prédiction moyenne de la variable Y uniquement prise en compte sur les arbres construits sur un sous-échantillon n'incluant pas cet individu.

De plus, en exploitant cette erreur, un ordre d'importance des variables explicatives dans la prédiction de la variable Y peut être estimé. L'importance d'une variable correspond à l'augmentation marginale de l'erreur *Out-of-Bag* due à une permutation des observations de la variable considérée. Si l'arrangement des observations n'impacte pas la prédiction, cela signifie qu'elle n'est pas importante lors de l'estimation de Y.

Inconvénients de ce modèle

Afin de diminuer la variance de l'estimateur de la variable à expliquer, il est nécessaire d'effectuer un nombre de modèles assez important afin de stabiliser l'erreur *Out-of-Bag*. Ceci entraîne souvent un temps de calcul qui est parfois considérable en plus d'une capacité mémoire suffisante pour stocker tous les modèles.

De plus, un inconvénient, absent lors de la création des modèles CART, consiste en une perte d'interprétabilité des résultats.

2.4.3 *Gradient Boosting*

Tandis que les forêts aléatoires permettent uniquement d'effectuer une diminution de la variance de l'estimateur de la variable à expliquer Y, les méthodes de *Boosting* ont un enjeu supplémentaire tout à fait différent puisqu'elles cherchent aussi à améliorer l'ajustement (le biais). Cette amélioration passe par une construction adaptative séquentielle d'estimateurs puis une combinaison de ces estimateurs pour éviter le surapprentissage. Ainsi, chaque modèle est une version adaptative du précédent. Le *Boosting* construit donc une famille de modèles récurrents.

Cette stratégie repose sur une adaptation de proche en proche où plus de poids est donné aux observations dont les prédictions sont moins précises dans le modèle précédent. Ainsi, intuitivement, la méthode va concentrer ses efforts sur les observations qui sont les plus difficiles à ajuster tout en limitant le surapprentissage par l'agrégation.

La procédure de *Boosting* peut être illustrée par les étapes suivantes :

- Estimation de Y par les variables X_j selon un modèle M_1. Il apparaît un vecteur d'erreur noté η_1.
- Estimation de η_1 par les variables X_j selon un modèle M_2. Il apparaît un vecteur d'erreur noté η_2.
- Les estimations sont réitérées pour avoir un estimateur agrégé qui s'exprime par l'équation suivante :

$$M^{(k)}(x) = M_1(x) + M_2(x) + ... + M_k(x) = M^{(k-1)}(x) + M_k(x) \tag{2.29}$$

Ainsi, en apprenant de son prédécesseur, le nouvel arbre créé pourra se concentrer sur les lacunes du premier qui ont été dévoilées par l'erreur.

Différents types d'algorithmes de *Boosting* existent avec des caractéristiques différentes :

- au niveau de la manière de pondérer les individus dont la valeur recherchée est mal estimée
- au niveau de la pondération des différents modèles lors de l'agrégation
- la fonction de perte qui mesure l'erreur d'ajustement

La suite présente l'une de ces méthodes de *Boosting*, appelée *Gradient Boosting* ou Descente de Gradient ainsi que son amélioration, l' *eXtreme Gradient Boosting* (ou ***XGBoost***).

Gradient Boosting

Le principe de l'algorithme de *Gradient Boosting* est de minimiser une fonction réelle différentiable f.

Cet algorithme itératif suit le principe suivant :

- Une valeur fixe x_0 est prise et un seuil de tolérance ϵ est aussi fixé.
- Calcul de $\bigtriangledown f(x_k)$
- Le critère d'arrêt $| \bigtriangledown f(x_k)| \leq \epsilon$ est pris en compte
- Calcul du pas α_k
- Calcul du nouveau point x_{k+1} selon la formule : $x_{k+1} = x_k - \alpha_k \bigtriangledown f(x_k)$

Les points calculés par l'algorithme convergeront et donneront le minimum recherché.

Cette approche est généralement utilisée avec des arbres de décision de taille fixe en tant que classificateurs faibles et, dans ce contexte, elle est qualifiée de gradient tree boosting. Si l'algorithme avec un arbre de décision devait être mis en forme, cela donnerait :

1. Initialisation du modèle :

$$F_0(x) = \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \rho) \tag{2.30}$$

2. Pour $m = 1, ..., M$

(a) Calcul des pseudo-résidus

$$\tilde{y}_{m,i} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}} \tag{2.31}$$

(b) Estimation du *weak-learner* $h(x, \alpha_m)$ sur les pseudo-résidus $(x_i, \tilde{y}_{m,i})_{i=1}^N$. Attention, le *weak-learner* est, dans le cas de l'étude, un arbre avec un nombre de feuilles L.

(c) Estimation du multiplicateur optimum sur les données initiales $(x_i, y_{m,i})_{i=1}^N$:

$$\rho_m = \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i, \alpha_m)) \tag{2.32}$$

(d) Incrémentation de l'estimateur : $F_m(x) = F_{m-1}(x) + \theta\rho_m h(x, \alpha_m)$ où $\theta < 1$ est un taux d'apprentissage qui est imposé pour limiter le surapprentissage.

Il est important de noter que, dans le cadre de l'utilisation de l'algorithme avec des arbres CART, le paramètre α représente un ensemble composé des régions de l'espace définies par les feuilles et de la valeur réponse associée.

Lors de la mise en œuvre de cette méthode, une étape importante sera le tuning des hyperparamètres :

- Le nombre de feuilles des arbres utilisés : L
- Le nombre d'arbres qui seront utilisés pour l'agrégation : M
- Le taux d'apprentissage qui sera utilisé : θ

XGBoost

Cette méthode de *Boosting* a permis, grâce à de nombreuses nouvelles options (et donc beaucoup de nouveaux paramètres), d'aborder avec plus de détails l'arbitrage biais-variance pour affiner la recherche du meilleur modèle.

En effet, les hyperparamètres de ce modèle sont au nombre de dix. Trois d'entre eux sont utilisés pour des tirages aléatoires sur les observations et les covariables afin de décorréler les arbres entre eux. Quatre autres sont nécessaires dans le but de « forcer manuellement » la taille des arbres, sans oublier le taux d'apprentissage du *Boosting*. Et enfin, les autres hyperparamètres sont utilisés afin d'effectuer une pénalisation qui aura un impact sur la fonction de perte objective.

Dans l'*XGBoost*, la fonction de perte est redéfinie comme une fonction objective avec :

$$obj = L + \Omega \tag{2.33}$$

Avec L la fonction de perte initiale et Ω une fonction de régularisation définie par :

$$\Omega(f) = \sum_{m=1}^{M} \left(\gamma L_m + \frac{1}{2}\lambda||\omega||_2^2 + \alpha||\omega||_1\right) \tag{2.34}$$

où :

- γ est un hyperparamètre pour la pénalisation sur le nombre de feuilles L de chaque arbre,
- λ est un hyperparamètre pour la pénalisation en norme L2 sur les valeurs réponses
- α est l'hyperparamètre pour la pénalisation en norme L1 sur les valeurs réponses

2.4.4 Optimisation des paramètres

Avant de développer la méthode qui sera utilisée pour l'optimisation des paramètres, un principe doit être introduit, celui de validation croisée (ou *k fold Cross Validation*). Lors de sa mise en œuvre, la base d'observations est séparée en k sous-échantillons de même taille. Un des sous-échantillons va être considéré comme base de test pour permettre le calcul de la mesure de performance, tandis que les k-1 sous-échantillons restants vont servir de base pour entraîner le modèle. L'opération est répétée k fois jusqu'à ce que chaque sous-échantillon ait été base de test. Les k mesures de performance ainsi obtenues sont ensuite combinées pour obtenir une mesure agrégée de la performance du modèle. Par exemple, la moyenne des k mesures de performance peut être utilisée pour évaluer les performances du modèle.

Une optimisation des paramètres des modèles *Random Forest* et *XGBoost* est nécessaire afin d'obtenir un modèle le plus performant possible. Pour cela, il faut ajuster l'exécution de l'algorithme et déterminer des hyperparamètres (ou paramètres d'ajustement) qui donneront les meilleurs résultats.

Pour réaliser cette optimisation, le choix est fait de réaliser un *gridsearch* (ou la recherche par quadrillage). Cette méthode de validation croisée, très répandue, consiste à choisir une liste de possibilités pour chacun des hyperparamètres et à entraîner ensuite le modèle pour chacune des combinaisons. Le but final est de déduire et conserver le meilleur paramétrage obtenu.

2.5 Les indicateurs de performances

2.5.1 Indicateurs pour la sélection du GLM

De nombreux indicateurs de sélection de modèles peuvent être utilisés. Les trois indicateurs qui sont le plus souvent utilisés sont la déviance, le *Akaike Information Criterion* (AIC) et le *Bayesian Information Criterion* (BIC).

La déviance est habituellement calculée pour évaluer la qualité de l'ajustement d'un GLM. Elle va comparer le modèle qui aura été estimé au modèle saturé. Le modèle saturé correspond au modèle qui comporterait autant de paramètres à estimer que d'observations dans la base de données.

En notant $\mathcal{D}$ la déviance, $\mathcal{L}$ la vraisemblance du modèle créé et $\mathcal{L}_{SAT}$ la vraisemblance du modèle saturé, il est possible de définir :

$$\mathcal{D} = -2\left(log\left(\mathcal{L}\right) - log\left(\underset{SAT}{\mathcal{L}}\right)\right) \tag{2.35}$$

Ainsi, l'idée est de comparer le modèle estimé à celui qui est considéré comme étant le mieux ajusté, le modèle saturé. Ce contexte fournira une description parfaite des données.

Plus la valeur de la déviance est réduite, plus les vraisemblances seront proches. Or plus les vraisemblances sont proches, mieux le modèle décrira les données. L'objectif dans le choix des modèles sera donc de minimiser la déviance.

Le critère AIC est défini de la manière suivante :

$$AIC = -2log(\mathcal{L}) + 2 \times J \tag{2.36}$$

où J est le nombre de variables dans le modèle et $\mathcal{L}$ la vraisemblance du modèle.
Le meilleur modèle selon le critère de l'AIC sera celui qui minimisera l'AIC.

Le critère du BIC pénalise plus les modèles complexes en introduisant dans la formule le nombre d'observations N.

$$BIC = -2log(\mathcal{L}) + J \times log(N) \tag{2.37}$$

Pour établir quel est le meilleur modèle selon le critère du BIC, il faut opter pour le modèle qui aura le plus petit BIC.

Ces deux critères aident à juger l'adéquation du modèle aux données en prenant en compte la complexité du modèle en question.

Dans le cas où une variable continue serait étudiée, des tests d'adéquation peuvent aussi être utilisés.

Avant d'effectuer le test d'adéquation, il faut au préalable faire une analyse graphique en traçant la distribution empirique pour la comparer avec différentes distributions théoriques. Ainsi, une pré-sélection des lois qui semblent être les plus pertinentes sera effectuée.

Plusieurs tests d'adéquation existent mais les plus fréquemment utilisés sont le test de Kolmogorov-Smirnov et le test de Cramer-Von-Mises.
Dans les deux tests, l'hypothèse testée H_0 sera la suivante :

$$H_0 \ : \ \hat{F} = F_0 \tag{2.38}$$

où $\hat{F}$ est la fonction de répartition empirique et F_0 la fonction de répartition à laquelle elle est comparée.

Ces deux tests se basent donc sur l'écart entre deux fonctions de répartition qui sont à comparer. La seule différence est que le test de Kolmogorov-Smirnov utilise la valeur maximale de l'écart alors que le deuxième test utilise la somme des différences. Ainsi, le test de Kolmogorov-Smirnov sera plus sensible aux valeurs extrêmes. Cela explique que le test de Cramer-Von-Mises est préféré pour vérifier l'adéquation des données à une loi.

2.5.2 Indicateurs pour la comparaison des modèles de l'étude

Les critères AIC et BIC ne pouvant être appliqués pour les modèles de *Data Science* utilisés dans cette étude, d'autres indicateurs devront être pris en compte pour comparer les modèles.

Un indicateur qui est habituellement utilisé est la MSE, qui est définie par :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.39}$$

Avec cet indicateur, la distance entre la valeur observée de la variable réponse y_i et la valeur prédite par le modèle considéré $\hat{y}_i$ sera calculée.

Ainsi, à partir de la MSE, il sera possible de regarder l'adéquation du modèle aux observations. Plus la MSE sera faible, plus le modèle créé sera représentatif des données initiales.

La MSE est donc une métrique qui facilite la détermination du modèle avec les meilleures prédictions des valeurs réelles. Cet indicateur de performance permet de mettre en évidence le pouvoir prédictif du modèle.

L'étude de cet indicateur sera parfois remplacée par celle de la RMSE (*Root Mean Square Error*) qui est la racine carrée de la MSE.

Afin de pouvoir approfondir la comparaison des modèles, la courbe de Lorenz sera aussi étudiée. Cette courbe permet la mesure de la qualité de la segmentation du modèle considéré.

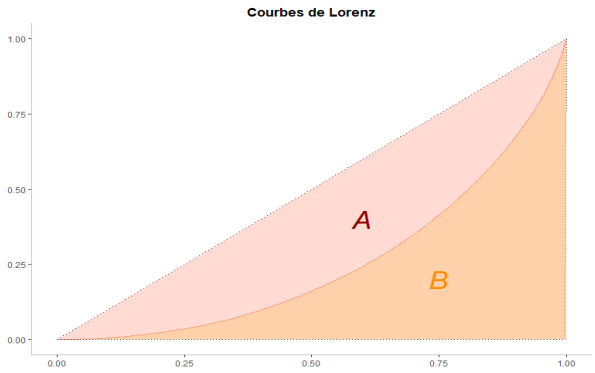

FIGURE 2.4 – Exemple de représentation d'une courbe de Lorenz

L'exemple ci-dessus (Figure 2.4) illustre la forme que les courbes de Lorenz peuvent prendre. En fonction de la variable étudiée, les axes auront des interprétations différentes. Par exemple, dans le cas où l'étude se porterait sur des coûts de sinistres, l'axe des abscisses représente la proportion des sinistres et celui des ordonnées illustre la proportion des coûts. Ainsi, à partir de la courbe de Lorenz, il sera possible d'exprimer la part du coût total correspondant aux $x\%$ des sinistres ayant les coûts les plus faibles.

Plus la courbe de Lorenz s'éloigne de la bissectrice, plus la distribution représentée sera discriminante.

L'indice de Gini est un indicateur lié à la courbe de Lorenz qui permet lui aussi de déterminer le pouvoir discriminant du modèle testé. Ce lien avec la courbe de Lorenz s'établi au travers la formule suivante :

$$Gini = \frac{A}{A+B} = 2A \tag{2.40}$$

Un indice de Gini proche de 0 indiquera une absence de discrimination de la distribution de la variable étudiée. La segmentation étant une part importante dans la réalisation de la tarification, l'indice de Gini doit donc être maximisé.

Dans l'équation de l'indice de Gini, hypothèse est faite que l'aire B est la zone du modèle optimal vers laquelle l'aire A devrait converger. Or, la courbe de Lorenz des observations, qui correspond à la discrimination à atteindre au travers des modélisations, peut ne pas atteindre l'intégralité de cette zone B. Cela correspond à la situation suivante :

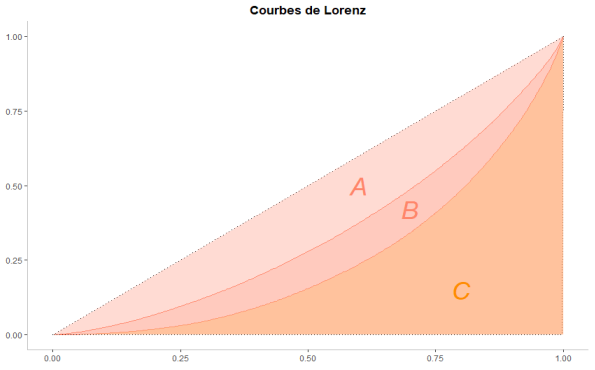

FIGURE 2.5 – Exemple de représentation d'une courbe de Lorenz

L'indice de Gini des observations sera ainsi $G_{obs} = \frac{A+B}{A+B+C}$ et celui du modèle sera $G = \frac{A}{A+B+C}$. En voulant comparer l'indice de Gini du modèle à celui observé, l'indice de Gini normalisé est introduit :

$$\tilde{G} = \frac{G}{G_{obs}} = \frac{A}{A+B} \tag{2.41}$$

Ainsi, le meilleur modèle selon l'indice de Gini normalisé sera celui dont l'indice sera le plus proche de 1.

Dans la suite de ce mémoire, l'indice de Gini correspondra à l'indice de Gini normalisé.

Chapitre 3

Traitements des données

Dans ce chapitre, les diverses étapes, qui ont permis d'obtenir des bases fréquences et sévérités pour chacune des garanties à tarifer lors de l'étude, seront abordées.

3.1 Périmètre de l'étude

Cette étude va porter sur la tarification de certaines des garanties les plus fréquentes de l'assurance MRH (Dégâts des eaux (DDE), Incendie, Vol, Bris de Glace (BDG) et Responsabilité Civile (RC)).

L'analyse se concentrera sur les données de sept partenaires. Ces partenaires seront notés A,B,C,D,E,F et G. Ils sont choisis parmi ceux qui ont le plus de contrats mais aussi parmi les partenariats les plus récents pour pouvoir caractériser le plus précisément possible le portefeuille de l'ÉQUITÉ.

Ces sept partenaires représentent un portefeuille de 279 450 contrats sur une période de 7 ans.

Ayant besoin d'idéalement aux alentours de 5 000 sinistres par garantie pour effectuer une tarification, il est décidé de réaliser l'étude sur une période de 7 ans. Ainsi, la période d'observation s'étend du 01/01/2015 au 31/12/2021. Les bases de données qui ont été utilisées sont en fait celles de mars 2022. Les trois mois de recul sont nécessaires pour garder une marge de prudence sur la déclaration de sinistres tardifs.

Pour chaque garantie à tarifer, le nombre de sinistres attritionnels survenus lors de cette période d'étude est disponible :

	DDE	**Incendie**	**Vol**	**BDG**	**RC**	**Total**
Nombre de sinistres	18 298	1 890	2 256	3 089	3 371	28 904

Table 3.1 – Sinistres Attritionnels par garantie

Il n'a pas été possible de récupérer plus de sinistres en raison de certaines contraintes au niveau du portefeuille. L'ensemble de ces 28 904 sinistres représente une charge totale de 37 501 513 €.

Le logiciel SAS Entreprise Guide est utilisé pour la création des différentes bases ainsi que pour leur analyse, avec l'aide d'Excel. De plus, Excel VBA permets de gérer certains problèmes de valeurs manquantes. Enfin, le logiciel R est choisi pour l'analyse des corrélations entre les variables.

3.2 Données de la base contrats

Dans cette partie, il sera décrit les différentes variables, caractéristiques des contrats, qui ont été conservées par rapport à leur utilité pour la suite de cette analyse, ainsi que les traitements qui ont du être effectués afin d'obtenir la base contrats agrégée des différents partenaires.

3.2.1 Les variables conservées

Une base contenant les contrats est disponible pour chaque partenaire. Ces bases regroupent toutes les informations que le partenaire conserve pour le contrat de ses assurés.

Les variables qui ont été conservées dans la base contrats sont les suivantes :

- Type d'habitation
- Type de résidence
- Qualité du souscripteur
- Nombres de pièces principales
- Variable définissant si l'habitation se trouve au rez-de-chaussée ou pas
- Montant du capital mobilier
- Pourcentage d'objets de valeur
- Le code postal
- Variable binaire caractérisant si des sinistres antérieurs ont eux lieu
- Numéro de police
- Nom du partenaire
- Franchise
- Variable binaire indiquant la présence d'options
- Nombres d'options souscrites dans le contrat
- Variables binaires indiquant si respectivement les garanties vol et bris de glace ont été prises en option
- Date d'effet et date de résiliation
- Valeurs des différentes primes des garanties à tarifer dans l'étude

3.2.2 Construction et traitements sur cette base

Pour obtenir la base contrats finale, il a été nécessaire d'effectuer des traitements sur les données présentes dans chacune des bases, pour pouvoir ensuite les réunir dans une base unique. Ces traitements allaient de la modification des formats des colonnes à la création de nouvelles variables pour uniformiser les sorties pour tous les partenaires en passant par le calcul de la durée d'exposition pour chaque individu.

Parmi les traitements effectués pour la création de cette base, les modifications détaillées dans la suite ont notamment été réalisées.

Dans le cas de trois des partenaires étudiés (E, F et G), l'information concernant le pourcentage d'objets de valeur n'était pas présente. L'information permettant de savoir si l'option objet de valeur a été prise au contrat était la seule disponible.

Ainsi pour obtenir approximativement le pourcentage en objet de valeur (OV) (qui ne pouvait être que 15% ou 30% [1]), les étapes suivantes sont effectuées :

1. D'après les Notes Techniques de ces partenaires

- Dans un premier temps, une base comprenant les contrats de la base des partenaires ayant l'option objets de valeur est créée.
- Ensuite, à partir du tarificateur à notre disposition et pour chacun des contrats disponibles, sont récupérés les montants de la prime vol dans les cas où le pourcentage en OV est de 15% et de 30%.
- Enfin, le choix est fait de conserver celui de ces deux montants de prime qui est le plus proche de la prime vol renseignée pour le contrat. Le cas qui a l'écart le plus petit est estimé comme étant le pourcentage OV choisi par le client.

Un autre traitement réalisé est la création de la variable RDC (rez-de-chaussée). Il a été nécessaire de créer une variable indiquant si le bien est au rez-de-chaussée ou non, puisque les partenaires représentant plus de 60% du portefeuille ne possédaient pas l'information sur le nombre d'étages dans l'immeuble mais uniquement si l'appartement est au rez-de-chaussée ou non.

Ne pouvant pas créer un modèle pour simuler la position du logement dans l'immeuble et avoir 3 modalités pour la variable illustrant l'étage de l'appartement (rez-de-chaussée, intermédiaire et dernier étage), le choix a été fait de conserver uniquement l'information sur la présence du bien au rez-de-chaussée.

De plus, pour ces mêmes partenaires, l'information sur l'étage pour les contrats les plus anciens n'était pas disponible puisque la variable n'était pas demandée avant 2015 et était manquante pour certaines des années suivantes. La modalité de cette variable est donc estimée afin d'éviter de perdre des contrats et leurs sinistres associés.

Une estimation sur R à partir d'une fonction se basant sur un *Random Forest*, la fonction *missForest*, est calculée.

Cet algorithme va prendre, dans un premier temps, les valeurs manquantes et les estimer en utilisant les valeurs médianes, puis elles seront marquées comme étant prédites. Un *dataset* transformé est obtenu. C'est ce *dataset* qui sera ensuite modifié à chacune des itérations demandées par l'utilisateur à l'aide d'un *Random Forest*. Ainsi, les valeurs manquantes du *dataset* seront estimées, à chaque itération, de manière plus précise.

Cette méthode de prédiction sera préférée à, par exemple, *K-nearest neighbors* (KNN), notamment parce qu'elle peut être utilisée sur des variables numériques ou qualitatives sans modification (parce qu'il n'y a pas d'hypothèse sur les relations entre les variables qui est faite en amont). De plus, elle peut être mise en pratique sur des variables colinéaires et avec des bruits. Enfin, le *Random Forest* est une méthode non paramétrique qui ne nécessite aucune approximation du paramétrage.

Avec cet algorithme et sur une base test, où des valeurs manquantes ont été aléatoirement simulées, une *accuracy* de 81% environ est obtenue. Pour rappel, l'*accuracy* mesure le taux de prédiction correcte à partir de la matrice de confusion et elle est calculée en prenant le nombre de valeurs bien prédites sur la longueur du *dataset*.

La décision est prise de conserver une *accuracy* dans les 80% puisque les valeurs manquantes pour cette variable ne représentaient qu'un faible pourcentage du portefeuille du partenaire (11.77%) et donc du portefeuille global (9.79%) [2].

Une telle approximation est aussi faite pour estimer les valeurs manquantes pour la variable sur la qualité du souscripteur, pour les mêmes partenaires. Dans ce cas-là, une *accuracy* d'environ 93% est obtenue. Cette estimation était suffisante par rapport à la représentation de ces valeurs manquantes, non seulement sur le portefeuille du partenaire (6.05%), mais aussi sur le portefeuille global (5.04%) [3].

Pour prendre en compte un certain profil d'individus, il a été nécessaire de considérer, dans la tarification, la présence ou non d'options. Ainsi, une variable pour la présence d'options est créée. En l'analysant, il est apparu que pour la plupart des partenaires, un pourcentage important d'individus de leur portefeuille avait fait le choix de

2. Les deux pourcentages qui ont été notés dans ce paragraphe ont été calculés alors que la date minimale pour l'étude n'était pas encore définie. Ils sont donc représentatifs d'une vision débutant en 2011, date minimale envisagée au départ.

3. Même remarque que la précédente, bien que l'effet soit moins important pour cette variable, puisque la répartition des valeurs manquantes, par rapport aux années, est moins flagrante.

prendre une option. Ainsi pour segmenter plus efficacement ce portefeuille, il est décidé de prendre en compte le nombre d'options choisies par un individu plutôt que leur seule présence.

Pour approfondir l'analyse de l'effet de la présence d'options sur la tarification, deux variables, permettant de déterminer respectivement si la garantie Vol et la garantie Bris de Glace avaient été souscrites en option, sont créées.

En effet, pour certaines des formules présentes chez les partenaires de l'ÉQUITÉ, ces garanties pouvaient être non obligatoires. Ainsi, utiliser ces variables lors de la tarification de ces garanties était susceptible d'apporter une information supplémentaire permettant d'améliorer les modèles.

Ayant plusieurs partenaires avec des modes de distribution variés, le choix de créer une variable sur le type de distribution s'est imposé. Ainsi, pour chaque police de chaque partenaire, une des modalités suivantes est assignée : "internet", "agence", "grossiste" et "réseau d'agents".

Pour finir, l'exposition de chaque ligne de la base combinée est calculée, pour avoir uniquement les images des contrats sur un an.

3.3 Données des bases sinistres

Cette partie est consacrée à la description des différentes variables liées aux sinistres qui ont été conservées, ainsi que les traitements qui ont dû être effectués afin d'obtenir une base sinistre agrégée ainsi qu'une base dédiée aux sinistres orphelins.

3.3.1 Les variables de l'étude

Une base contenant les sinistres pour chaque partenaire constitue le point de départ. Ces différentes bases, propres à chaque partenaire, sont regroupées en une base commune contenant 59 colonnes.

Les variables de la base sont notamment :

- Le numéro du sinistre
- Le numéro de police
- Les montants du règlement du sinistre par garantie
- La date de survenance
- Le code pour définir le type de sinistre

La présence de sinistres survenus et non rattachés à un contrat conduit à la création d'une base commune de sinistre "orphelins" contenant notamment les colonnes :

- Le numéro du sinistre
- Le numéro de l'assurance à laquelle il est fait référence
- La date de survenance
- Le coût du sinistre par garantie

3.3.2 Construction de la base

Les sinistres présents dans les bases des différents partenaires sont récupérés [4]. L'étape suivante est la fusion des informations de ces bases afin de récupérer les valeurs actualisées des montants nécessaires pour obtenir la charge sinistre à étudier.

4. Cela comprend les sinistres survenus durant l'intervalle de l'étude. Certains sont en cours et d'autres clôturés.

Dans la base ainsi obtenue, le montant des sinistres est calculé, en prenant en compte les dommages matériels et corporels.

Il est important d'ajouter les provisions pour les sinistres qui n'ont pas été payés intégralement. En effet, lorsqu'un sinistre est déclaré, une évaluation est associée à ce sinistre. Cependant, la charge associée à ce sinistre peut varier au cours du temps. D'où la nécessité d'affecter ces augmentations ou diminutions de la charge effective future à la charge utilisée lors de la tarification.

Un provisionnement est effectué à partir du triangle ci-dessous et de la méthode de *Chain-Ladder* (développée dans l'annexe C) :

	0	1	2	3	4	5	6	7
2015	24 774,85	1 559 314,78	1 979 875,88	2 022 542,38	2 098 195,40	2 116 536,00	2 121 863,55	2 141 871,51
2016	63 173,28	1 642 675,33	2 250 322,03	2 364 490,16	2 443 193,32	2 450 921,40	2 484 085,17	
2017	45 553,98	1 817 441,84	2 394 865,13	2 580 113,87	2 646 604,40	2 663 387,59		
2018	99 050,55	2 387 383,02	3 288 087,38	3 511 151,18	3 618 970,85			
2019	66 777,72	2 593 303,34	4 028 190,97	4 354 158,29				
2020	77 361,48	3 798 425,44	5 571 353,06					
2021	68 991,51	4 546 341,75						
2022								

FIGURE 3.1 – Triangle de règlement pour le provisionnement

La méthode de *Chain-Ladder* peut être appliquée puisque les variations des facteurs sont stables par année de développement (cf. le tableau C.2). De plus, l'alignement des *C-C plots* par année de développement est assuré (cf. partie annexe C).

A partir des facteurs trouvés, des coefficients pour chaque année de survenance sont estimés, ce qui permet de rajouter les provisions aux montants de règlement de nos sinistres.

Année survenance	2021	2020	2019	2018	2017	2016
Facteur CL	1,414112591	1,063918849	1,031366391	1,00596159	1,008427297	1,009429428
Coefficient	1,588939716	1,123630273	1,05612404	1,024004708	1,01793619	1,009429428

FIGURE 3.2 – Coefficients pour le provisionnement des sinistres

Ensuite, il existe quelques sinistres nuls en 2021, qui n'ont pas encore eu de règlement (le montant du sinistre est donc de 0). Afin de pouvoir prendre en compte les provisionnements de ces sinistres et le fait qu'ils auront un règlement futur, la décision est prise de considérer le règlement moyen lors de la première année, en fonction de certaines caractéristiques des contrats auxquels les sinistres correspondaient (en particulier le type de résidence, d'habitation, la qualité du souscripteur et la franchise).

Le provisionnement est aussi effectué sur ces montants de règlement de sinistres.

Avant de poursuivre, il apparaît impératif, en considérant l'intervalle d'étude défini, de passer les sinistres en *As If*. En effet, la valeur en euros d'un sinistre de 2011 ne sera pas la même que celle d'un sinistre de 2021. Cela est dû, par exemple, à l'évolution du prix des matériaux ou de la main d'œuvre. Ainsi pour revaloriser ces sinistres, un indice cohérent avec le portefeuille doit être trouvé.

Pour cela, deux indices susceptibles d'être retenus, sont étudiés, l'indice FFB (Fédération Française du Bâtiment) et l'indice de l'inflation. L'évolution des facteurs des moyennes de montants par année et l'évolution des facteurs *As If* sont comparées pour l'ensemble des garanties mais aussi pour chaque garantie.

Par les graphiques situés dans l'annexe D, le constat est fait que l'évolution des facteurs empiriques suit plutôt l'évolution de l'indice FFB que celle de l'indice de l'inflation (en globalité). L'indice FFB est donc choisi pour mettre en *As If* les montants de charges sinistres.

Enfin, les sinistres qui n'ont pas pu être liés à un des contrats sont mis dans une base séparée.

La première approche considérée pour la gestion de ces sinistres "orphelins" a été de répartir équitablement leur montant entre tous les contrats sinistrés de la garantie correspondante.

Cependant, en utilisant cette approche, il s'est avéré difficile d'évaluer l'effet de ces sinistres "orphelins" sur la fréquence de sinistres qui aurait été modélisée. C'est pour cette raison qu'il a été décidé d'abandonner cette première approche. Il a été préféré d'ajouter, à la prime pure attritionnelle estimée par modélisation, une prime qui sera appelée prime "orphelins". Cette prime "orphelins" sera calculée par un modèle du type coût moyen observé multiplié par fréquence observée, pour chaque garantie.

Une analyse plus poussée de ces sinistres est effectuée.

	Orphelins					
	DDE	BDG	RC	vol	inc	Total
Charge*	3,9%	1,3%	12,5%	3,3%	131,1%	25,1%
Coût moyen	2955,71	522,35	7103,71	3401,45	70572,58	17372,30
Fréquence	0,00063	0,00011315	0,0001094	0,00012163	0,00024025	0,00114762
Nombre**	1,5%	1,2%	1,5%	1,8%	5,8%	1,8%

* Par rapport à la charge attritionnelle du portefeuille

** Par rapport aux sinistres déclarés sur le portefeuille

FIGURE 3.3 – Tableau d'analyse des sinistres "orphelins"

Ce qui donne comme prime pure pour les sinistres "orphelins" :

	Orphelins					
	DDE	BDG	RC	vol	inc	Total
Prime pure	1,86402441	0,05910406	0,77713909	0,41370677	16,954972	19,9367453

FIGURE 3.4 – Prime pure pour les sinistres "orphelins" par garantie

De plus, pour la base sinistres et pour la base "orphelins", il est créé, en plus de la variable charge totale, une charge en fonction du type de sinistres et des variables pour chaque type de sinistre, qui attribuent une valeur 1 si le sinistre est de ce type. Ainsi, si le sinistre est par exemple un vol, la charge sera mise sur la charge vol et les autres charges seront mises à 0 et uniquement la variable nombre de vol sera mise à 1 et les autres à 0.

3.4 Bases sévérité et bases fréquence

Les bases contrats et sinistres sont ensuite réunies en une seule base grâce au numéro du contrat mais aussi à la date du sinistre. C'est cette base qui sera utilisée pour la suite de l'étude.

La sinistralité se décompose en trois typologies de sinistres : attritionnels, graves et CAT. Il est important d'effectuer la séparation des charges sinistres avant la modélisation car les modèles classiques ne peuvent être appliqués que sur les montants attritionnels à cause des queues des lois de distribution utilisées.

C'est pour cette raison que le choix est fait d'enlever les valeurs des sinistres supérieures à 30 000 euros de la base. Ce seuil de 30 000 euros a été fixé par l'entreprise en raison du nombre insuffisant de données avec des montants de sinistres dépassant cette valeur. En effet, cela aurait entraîné un biais trop important dans une analyse de seuil de sinistralité.

En enlevant ces sinistres graves, il a été estimé que le montant et le nombre de sinistres représentaient, dans la globalité, 26.8% du montant total des sinistres attritionnels et environ 0.3% de ces sinistres.

Pour plus de précision sur la survenue de ce type de sinistres, des statistiques par type de sinistre sont analysées.

	Graves					
	DDE	BDG	RC	vol	inc	Total
Charge*	2,4%	0,0%	11,9%	1,9%	148,9%	26,8%
Coût moyen	59656,36	0	49184,13	39514,21	110849,14	100393,99
Fréquence	0,00002	0	0,000015	0,000006	0,000174	0,000212
Nombre**	0,0%	0,0%	0,2%	0,1%	4,2%	0,3%

* Par rapport à la charge attritionnelle du portefeuille

** Par rapport aux sinistres déclarés sur le portefeuille

FIGURE 3.5 – Tableau d'analyse des sinistres graves

Comme pour les sinistres "orphelins", les sinistres graves ne représentent qu'une part minime des sinistres attritionnels. Deux méthodes sont possibles. Une première approche serait de rajouter des primes pures observées pour les sinistres graves à la prime pure attritionnelle. Les résultats pour ces primes figurent dans le tableau suivant :

	Graves					
	DDE	BDG	RC	vol	inc	Total
Prime pure	1,15170779	0	0,73852643	0,23443772	19,2601824	21,3199387

FIGURE 3.6 – Prime pure pour les sinistres graves, par garantie

Une autre méthode pour gérer ce type de sinistres pourrait être d'utiliser un provisionnement pour grave global et exprimé en pourcentage de la prime finale.

Après avoir fusionné la base contrats et la base sinistres, une variable zone est créée à partir du code postal. Cette variable zone est définie à partir du nouveau zonier créé et propre à l'ÉQUITÉ, pour sa nouvelle tarification.

3.5 Statistiques Descriptives

Ces études statistiques sont effectuées avec les montants attritionnels (après passage en *As If*).

Dans cette partie seront présentées quelques unes des variables disponibles dans la base obtenue et qui sont usuellement celles les plus influentes pour tarifer les garanties MRH.

La variable sur le type d'habitation

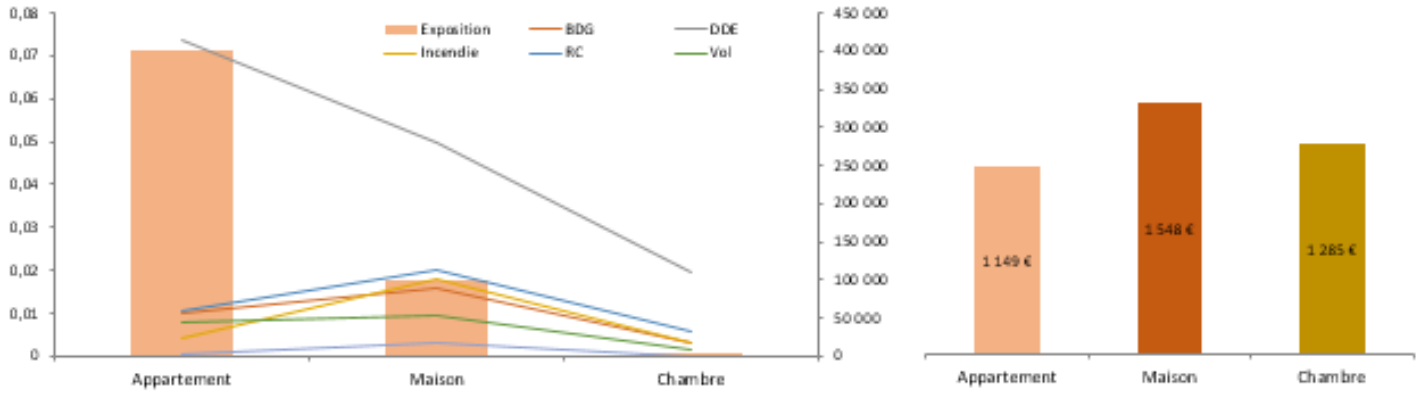

FIGURE 3.7 – Étude de la fréquence et du coût moyen suivant le type d'habitation

La figure 3.7 montre que le type d'habitation n'a pas une influence importante sur le coût moyen, bien qu'elle indique aussi que les montants moyens des sinistres avec la modalité "Maison" sont plus élevés que les deux autres.

En ce qui concerne les fréquences, le constat est que, pour toutes les garanties hors DDE, la fréquence des sinistres est plus élevée pour les maisons et moins pour les autres catégories d'habitation. Les fréquences sont différentes pour la garantie DDE et restent plus hautes que pour le reste des garanties. Dans ce cas particulier de la DDE, les appartements ont une occurrence de sinistres plus élevée que les deux autres modalités. Les modalités ont donc un impact sur le niveau des fréquences.

La variable sur le type de résidence

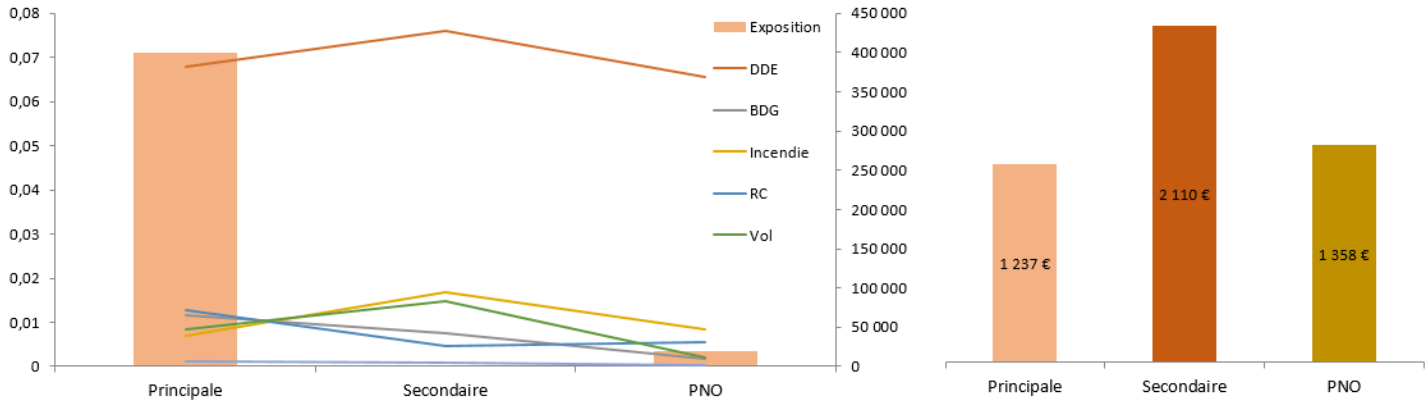

FIGURE 3.8 – Étude de la fréquence et du coût moyen suivant le type de Résidence

Contrairement au type d'habitation, les coûts moyens des sinistres semblent varier entre les modalités. Effectivement, les montants des sinistres moyens sont beaucoup plus importants lorsque l'habitation assurée est une résidence secondaire (près du double que dans les autres cas).

Comme dans le cas du type d'habitation, les fréquences pour la garantie DDE sont beaucoup plus hautes que celles des autres garanties. Deux types de forme de courbes de fréquences en fonction des garanties sont mises en évidence. En ce qui concerne le DDE, l'Incendie et le Vol, la fréquence est plus importante pour les résidences secondaires. Pour les autres garanties, la forme de la courbe de fréquence est inversée et la fréquence est plus faible pour les résidences secondaires. Les fréquences sont donc impactées différemment par la variable en fonction de la garantie. (figure 3.8)

La variable nombre de pièces principales

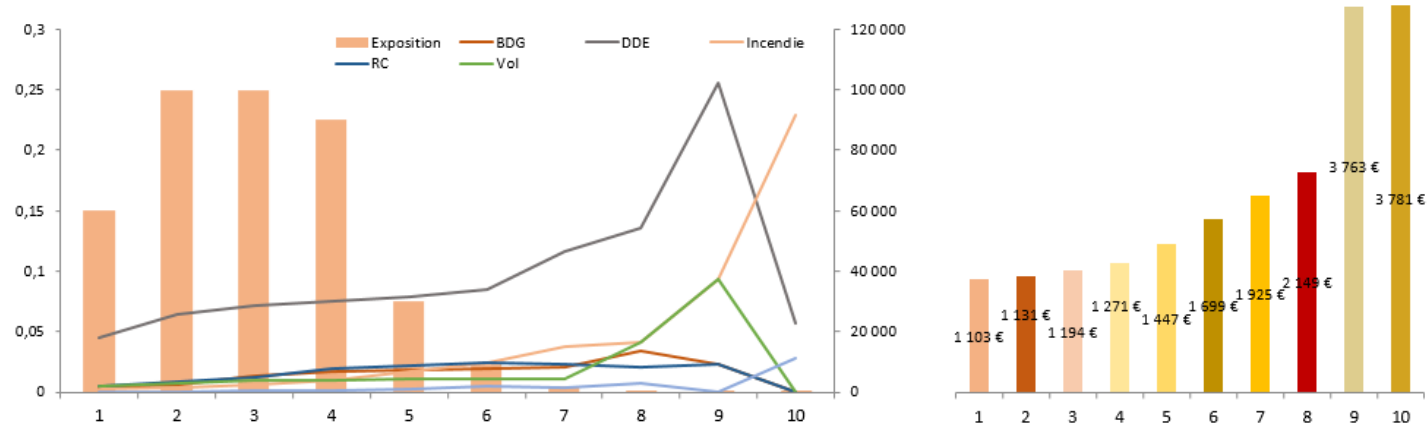

FIGURE 3.9 – Étude de la fréquence et du coût moyen suivant le nombre de pièces principales

Dans le cas du nombre de pièces principales, le coût moyen des sinistres augmente en fonction de leur nombre avec une stabilité qui s'effectue au niveau des habitations composées de 9 ou 10 pièces principales. Comme les coûts moyens, les fréquences par garantie s'amplifient lorsque le nombre de pièces principales augmente. Tout comme les autres variables, la fréquence DDE reste celle la plus élevée sauf pour 10 pièces principales où la fréquence Incendie est la plus élevée. Le nombre de pièces principales influence ainsi le coût moyen et les fréquences.

En comparant avec la fréquence maximale observée sur les autres variables, les valeurs des fréquences DDE (et pour une part des fréquences Incendie) sont beaucoup plus importantes en fonction du nombre de pièces principales. (figure 3.9)

La variable qualité du souscripteur

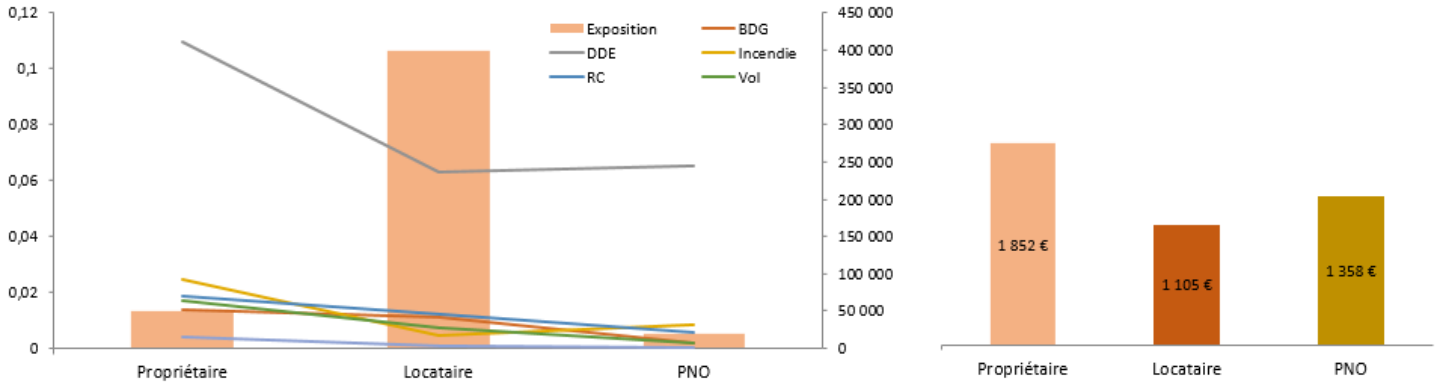

FIGURE 3.10 – Étude de la fréquence et du coût moyen suivant la qualité du souscripteur

A première vue, les coûts moyens ne varient pas significativement en fonction de la qualité du souscripteur. Une différence peut être constatée pour la garantie DDE concernant la fréquence mais pas pour les autres garanties qui restent assez stables en fonction des modalités. La fréquence est plus faible pour la garantie DDE quand la qualité du souscripteur est "Locataire" ce qui est d'autant plus intéressant que l'exposition est la plus importante pour cette modalité. (figure 3.10)

La variable partenaire

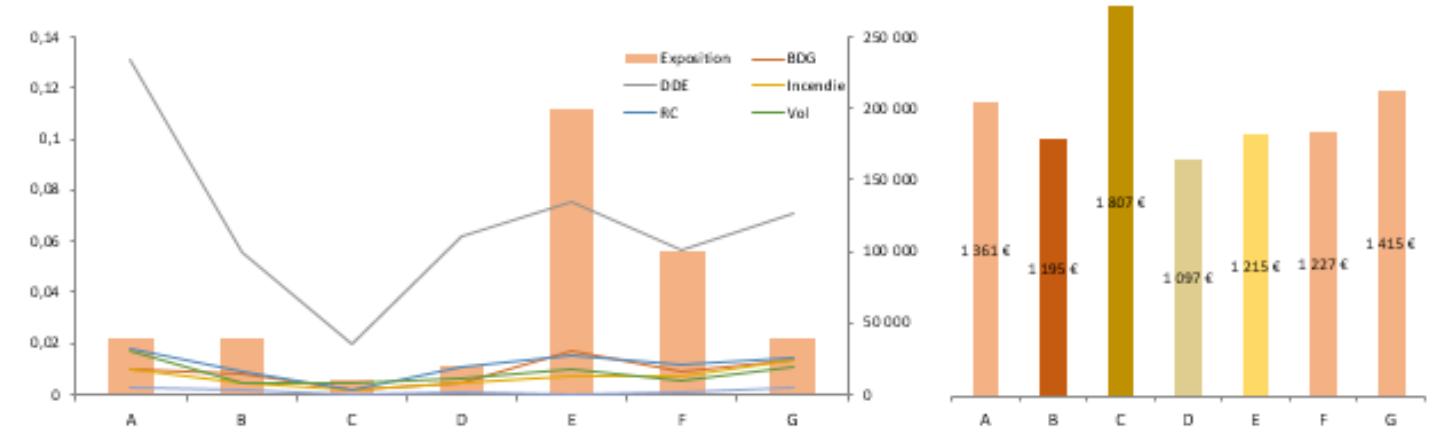

FIGURE 3.11 – Étude de la fréquence et du coût moyen suivant le partenaire

En visualisant les coûts moyens par partenaire, ces derniers ne varient pas significativement à l'exception du partenaire C qui a un coût moyen plus élevé que les autres.

De plus, la forme des courbes de fréquence est la même pour toutes les garanties sauf pour la fréquence DDE qui reste tout de même plus haute et moins stable que celle des autres garanties. (figure 3.11)

3.6 Analyses des variables

3.6.1 Regroupement des variables

Avant de réaliser les modèles, il faut regarder si les variables continues obtenues ne peuvent pas être transformées en variables qualitatives.

Les deux variables continues de la base (franchise et montant de capital mobilier) seront regroupées en modalités. En effet, pour certains des partenaires, ces variables sont de type qualitatives et non pas continues.

Pour la variable franchise, il est aisé de trouver un regroupement en fonction des modalités qui avaient été préalablement fixées par les partenaires, tout en prenant en compte le nombre de contrats ayant leur franchise autour de certaines valeurs, non déjà présentes en tant que modalités.

Ce regroupement conduit à une variable possédant six modalités (0, 75, 150, 200, 300 et 400) [5].

Plusieurs regroupements sont effectués en procédant par pas [6], plus ou moins importants, afin de réaliser un regroupement concernant le capital mobilier qui permet d'avoir, pour chaque catégorie, une exposition d'au moins 10 000.

Cette même procédure est aussi appliquée sur une autre variable créée, le montant de capital mobilier par pièce. Cette création s'est avérée nécessaire en raison de la forte dépendance de la variable montant capital mobilier à la variable du nombre de pièces principales (cf. section 3.6.2).

3.6.2 Analyse des corrélations entre les variables

Dans le cas des modèles linéaires généralisés, il est impératif de faire un choix dans les variables explicatives en prenant en compte qu'il doit y avoir une absence de corrélation entre ces variables.

Pour s'assurer de cette absence de corrélation, plusieurs indicateurs existent en fonction du type des variables qui sont testées.

Pour étudier la dépendance entre des **variables qualitatives**, il est possible d'utiliser le V de Cramer qui mesure la dépendance en se basant sur la statistique du χ^2.

La statistique du χ^2 (cf. équation 3.2) permet aussi d'étayer l'existence d'une relation entre deux variables. Or la valeur du χ^2 varie entre 0 et $+\infty$. Il n'est donc pas possible d'évaluer la dépendance entre les deux variables testées.

Le calcul du V de Cramer permet de normaliser cette statistique entre 0 et 1. Ainsi, l'existence de la dépendance ne sera plus seulement justifiée, mais cette dépendance pourra aussi être quantifiée. Plus la valeur de cet indicateur sera proche de 1, plus la dépendance entre les variables considérées sera importante.

Pour deux variables catégorielles comprenant respectivement r et s modalités, le V de Cramer sera défini par :

$$V = \sqrt{\frac{\chi^2}{N\ inf(r-1, s-1)}} \tag{3.1}$$

Avec

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{i,j} - \frac{n_{i.}n_{.j}}{N}\right)^2}{\frac{n_{i.}n_{.j}}{N}} \tag{3.2}$$

Dans l'équation 3.2, $\frac{n_{i.}n_{.j}}{N}$ correspond à l'effectif théorique sous l'hypothèse d'indépendance.

5. Les franchises sont ici en euros.
6. Ces pas sont fixés grâce à des tests afin d'obtenir l'exposition minimale recherchée dans le groupe créé.

La corrélation entre les variables qualitatives est illustrée par le graphique suivant :

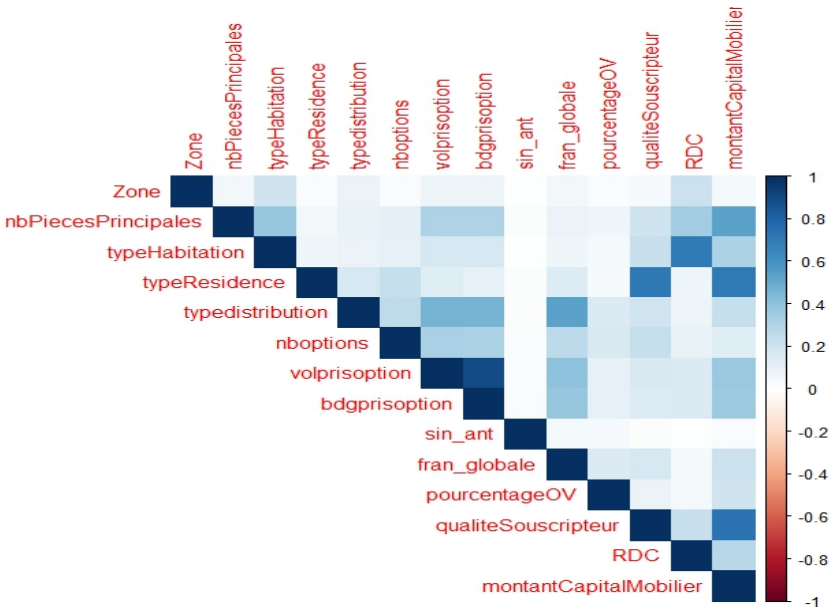

FIGURE 3.12 – Corrélation entre les variables pré-sélectionnées pour la tarification

Une forte corrélation entre certaines variables qui devaient être utilisées dans la tarification est constatée. C'est notamment le cas de la variable typeHabitation et de la variable pour caractériser l'étage. Cette liaison entre les variables peut découler du fait que les maisons avaient automatiquement la modalité "Inexistante" pour l'étage.

Un autre lien entre variables peut survenir entre le typeResidence et la qualiteSouscripteur. En effet, les propriétaires non occupants (qualité du souscripteur) ont une catégorie spécifique pour le type de résidence.

D'autres corrélations peuvent être citées, comme la variable montantCapitalMobilier qui peut interagir légèrement avec la variable pour le nombre de pièces principales ou encore la variable typedistribution qui peut être un peu dépendante de la franchise (ce qui est logique puisqu'en fonction des apporteurs, les seuils de franchise peuvent être différents et donc entraîner une certaine corrélation avec le type de distribution). (figure 3.12)

Pour remédier à ces problèmes de corrélation et ne pas perdre d'informations qui pourraient être utiles dans les modélisations de fréquence et de coût moyen, les informations des variables typeResidence et qualiteSouscripteur ainsi que typeHabitation et RDC [7] sont réunies. Il a aussi été tenté d'utiliser un montant de capital mobilier par pièce et de créer une variable de distribution différente n'ayant que deux modalités.

Quatre nouvelles variables sont donc créées.

7. Par exemple, dans le cas typeResidence/qualiteSouscripteur, les modalités présentes dans la variable obtenue seront PNO, Propriétaire - Principale, Propriétaire - Secondaire et Locataire - Principale

L'analyse sur les corrélations entre les variables est, de nouveau, effectuée.

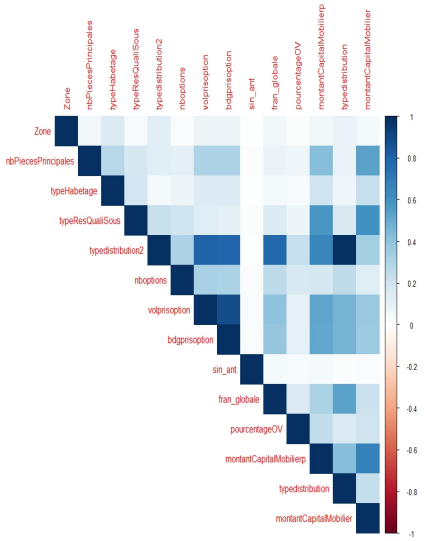

FIGURE 3.13 – Corrélation entre les variables pré-sélectionnées (avec variables créées)

Ce nouveau graphique montre qu'il n'est pas pertinent d'utiliser la variable typedistribution2 puisqu'elle est encore plus corrélée avec les autres variables que ne l'était la variable de départ. (figure 3.13)

Concernant la corrélation entre la variable nbPiecesPrincipales et montantCapitalMobilierp, le choix est fait de conserver cette variable pour les GLM plutôt que la variable montantCapitalMobilier.

De plus, les variables créées en combinant les informations de deux variables ne sont pas trop corrélées avec les autres. La seule exception se situe entre typeResQualiSous et montantCapitalMobilierp (corrélation inférieure à 0.5). Des tests devront donc être fait avant de commencer les GLM pour vérifier si cette liaison n'aura pas un impact lors de la modélisation.

3.6.3 Séparation de la base en base d'apprentissage et de validation

La base fréquence ainsi que les différentes bases sinistres (une par garantie à tarifer) sont séparées en échantillon d'apprentissage et de validation.

Le premier échantillon qui correspond à 80% de la base sera utilisé afin de modéliser la variable réponse.

Le reste sera utilisé pour une validation des modèles créés afin d'avoir une idée de leur efficacité sur les données qui n'auront pas servi à la calibration.

Les bases étant relativement volumineuses, le choix est fait de prendre aléatoirement des lignes dans chacune des bases afin de créer la base de validation.

Avant de commencer les chapitres sur la modélisation des fréquences et des montants pour les sinistres attritionnels, un test sur les bases de données, pour savoir si le modèle de fréquence-coût peut être appliqué pour les différentes garanties, est effectué.

Pour cela, les tests de corrélation de Pearson, Kendall et Spearman (cf. annexe E) sont notamment réalisés.
Les résultats de ces différents tests de corrélation se trouvent dans le tableau suivant :

	DDE	**Incendie**	**Vol**	**BDG**	**RC**
Pearson	-0.0043	0.0641	-0.0143	-0.0026	-0.0304
Kendall	-0.0213	0.0218	0.0008	0.0133	0.0105
Spearman	-0.0299	0.0296	0.0013	0.0188	0.0138

TABLE 3.2 – Tableau des corrélations entre le nombre et la charge des sinistres par garantie

A partir de ces résultats, il est possible de déduire que, pour l'ensemble des garanties, la méthode fréquence-coût peut être appliquée.

Chapitre 4

Application des modèles pour la modélisation de la fréquence

Dans cette partie, les différents résultats de l'application des modèles pour estimer la fréquence seront exposés (et comparés aux modèles "classiques" GLM). Toutes les analyses et modélisations présentées dans cette partie sont mises en place à partir du logiciel R.

4.1 Modèle Linéaire Généralisé

4.1.1 Avant la modélisation

Dans un premier temps, les deux lois les plus fréquemment utilisées pour la modélisation de la fréquence sont testées [1] afin de définir la loi sous-jacente pour la modélisation :

- La loi de Poisson
- La loi Binomiale Négative (BN)

Cependant, en regardant la distribution de la fréquence des sinistres, une masse d'observations avec 0 sinistres est constatée. C'est pour cette raison qu'il est décidé de tester un autre type de loi pour modéliser les fréquences : les lois "zéro-tronquées". Les résultats obtenus lors des différents tests, pour chaque garantie, ne se sont pas avérés concluants pour que l'utilisation de ces lois "zéro-tronquées" soit retenue pour la modélisation des fréquences. En effet, pour toutes les lois sous-jacentes utilisées pour estimer la partie non-nulle, les données ne permettent pas d'obtenir des valeurs satisfaisantes pour le test de significativité. Ainsi, les résultats ne sont pas présentés dans cette partie.
L'expression de l'ensemble de ces lois est développée dans l'annexe F.

Il est important de rappeler que la loi de Poisson est plus appropriée lorsque les distributions ne sont pas sur-dispersées.
Le tableau suivant représente les sur-dispersions calculées pour chaque garantie :

	DDE	**Incendie**	**Vol**	**BDG**	**RC**
Coefficient de sur-dispersion	1.058739	1.09731	1.07093	1.065048	1.080676

TABLE 4.1 – Tableau des coefficients de sur-dispersion par garantie

A partir de ces coefficients, l'hypothèse de l'utilisation du modèle de Poisson pour la modélisation des fréquences ne peut pas être rejetée.
Pour déterminer quelle loi est la plus logique entre la loi de Poisson et la loi Binomiale Négative, il faut regarder leur adéquation au travers de différents critères. Cette analyse est présentée dans les tableaux de l'annexe H.

1. Pour rappel, la fonction lien utilisée sera la fonction logarithmique.

D'après ces tableaux (avant modélisation et amélioration de chaque modèle), les modèles suivants ont été choisis comme les mieux adaptés :

	DDE	Incendie	Vol	BDG	RC
Modèle pré-sélectionné	BN	BN	BN	BN	BN

TABLE 4.2 – Tableau des GLM les plus adaptés par garantie (avant modélisation)

4.1.2 Choix des variables et calibration de leurs effets

Pour déterminer les variables explicatives du modèle, plusieurs méthodes de sélection automatique ainsi que des tests dits de "type III" sont effectués.

Les méthodes de sélection automatiques utilisées sont :

- Méthode ascendante (*Forward*) : cette méthode part du modèle le plus simple qui ne comporte que le terme constant. Les variables explicatives sont ensuite rajoutées, pas à pas, en fonction de celle qui améliorera significativement le critère de l'AIC. Le processus s'arrête à partir du moment où plus aucune variable n'améliore le modèle.
- Méthode descendante (*Backward*) : cette méthode part, à l'inverse, du modèle le plus complet (celui avec toutes les variables) et va enlever, pas à pas, les variables explicatives en fonction de celles dont la suppression entraîne la détérioration la moins significative de l'AIC. Le processus se poursuit jusqu'à ce que le modèle optimal, selon le critère de l'AIC, soit obtenu.
- Méthode progressive (*Stepwise*) : cette méthode consiste en un mélange des méthodes *Forward* et *Backward*.

Pour chaque GLM et pour toutes les garanties, l'obtention de la significativité des variables sélectionnées (avant tout regroupement des modalités des variables qualitatives) s'est avérée effective. Un exemple d'analyse de "type III" pour la loi de Poisson et la garantie DDE est illustré dans le tableau ci-dessous :

Variables	Df	Statistique	p-value
nbPiecesPrincipales	9	365.573	$< 2.2e-16$
Zone	9	531.009	$< 2.2e-16$
typeHabetage	4	532.241	$< 2.2e-16$
typeResQualiSous	3	287.504	$< 2.2e-16$
typedistribution	3	32.250	$4.636e-07$
nboptions	11	50.245	$5.655e-07$
sin_ant	1	562.036	$< 2.2e-16$
fran_globale	5	175.692	$< 2.2e-16$
pourcentageOV	4	52.332	$1.176e-10$
montantCapitalMobilierp	10	79.354	$6.720e-13$

TABLE 4.3 – Analyse de type III pour la loi de Poisson et la garantie DDE

L'ensemble des tableaux (avec loi de Poisson et Binomiale Négative) pour les autres garanties sont présents en annexe H.

Il faudra ensuite réaliser la calibration du modèle en effectuant des regroupements de modalités des variables sélectionnées dans l'étape précédente. Pour cela, les modalités qui n'étaient pas significatives ont été regroupées avec d'autres modalités de la même variable jusqu'à ce que tous les coefficients du GLM soient significatifs. Durant ces étapes, il a été vérifié que l'AIC et le BIC baissaient ou restaient similaires.

4.1.3 Analyse des résidus

Dans le cadre des GLM, bien que la normalité des résidus n'est plus une hypothèse, il est nécessaire d'analyser si l'erreur reste tout de même aléatoire. Pour vérifier cela, deux types de résidus sont principalement utilisés : ceux de déviance et ceux de Pearson.

Les résidus de déviance sont définis par la formule suivante :

$$r_i^D = signe(y_i - \hat{y}_i)\sqrt{d_i} \tag{4.1}$$

avec $d_i = 2\left(log\,\mathcal{L}(y_i, y_i) - log\,\mathcal{L}(y_i, \hat{y}_i)\right)$

Quant aux résidus de Pearson, ils sont définis par la formule suivante :

$$r_i^P = \frac{y_i - \hat{y}_i}{\sqrt{\mathbb{V}[\hat{y}_i]}} \tag{4.2}$$

où $\mathbb{V}[\hat{y}_i]$ est la variance de la prédiction de l'observation i.

Ces résidus ont été analysés pour les modèles utilisant respectivement la loi de Poisson et la loi Binomiale Négative.

Dans cette partie seront uniquement décrits les résultats obtenus pour l'une des études ainsi que pour l'une des garanties. L'exemple choisi est le GLM avec loi de Poisson dans le cadre de la garantie DDE (comme pour l'exemple du tableau de l'analyse de type III dans la partie 4.1.2). Les graphiques obtenus pour les autres modèles et garanties sont regroupés dans l'annexe I.

L'étude commence par l'interprétation des résidus groupés de Pearson. Pour réaliser cette analyse, des groupements sont faits à l'aide des valeurs prédites triées par ordre croissant, afin qu'ils soient homogènes en termes de risque prédit. L'objectif est de rassembler en groupe d'expositions homogènes les résidus du modèle. Les figures 4.1 et 4.2 ci-dessous illustrent l'exemple.

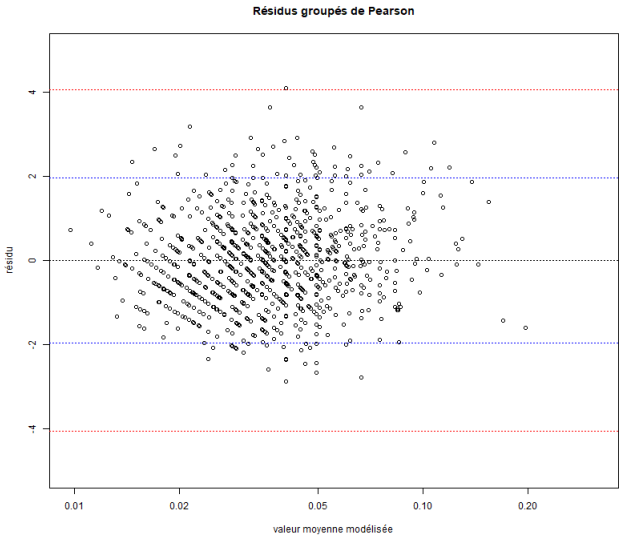

FIGURE 4.1 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - DDE

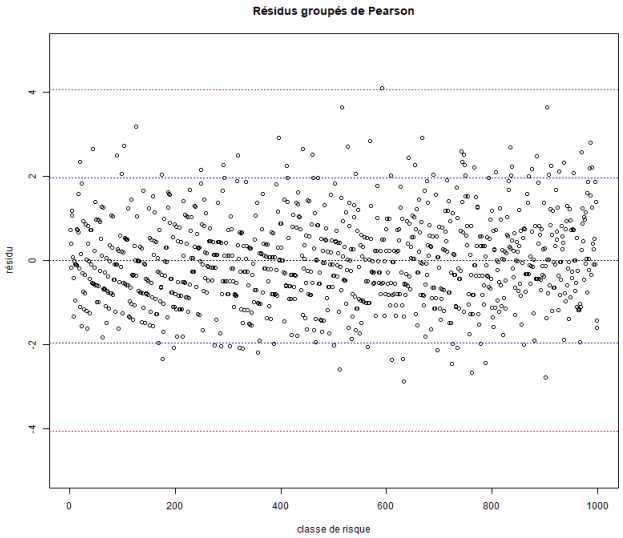

FIGURE 4.2 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- Poisson - DDE

Les deux graphiques précédents ne mettent en évidence aucune tendance particulière. Cependant quelques points s'écartent du nuage de points pour certaines valeurs moyennes modélisées sur le graphique 4.1 et pour les classes de risque supérieures (qui correspondent aux fréquences prédites fortes) sur le graphique 4.2. Ce constat ne remet pas en cause l'adéquation du modèle.

La plupart des valeurs sont comprises entre -2 et $+2$, avec quasiment aucune valeur au-delà de $+4$ ou en dessous de -4. Ceci permet de supposer la normalité des résidus. Mais pour justifier cette normalité, il faut comparer la distribution et la fonction de répartition des résidus avec celles de la loi normale. (Figures 4.3 et 4.4)

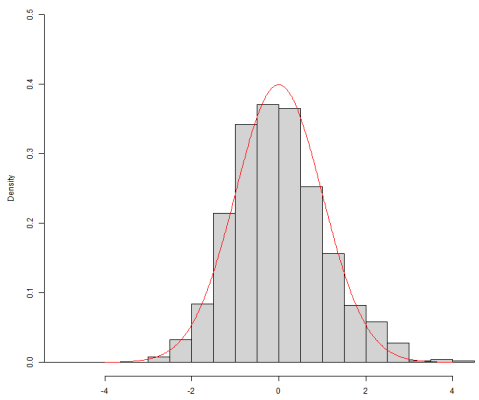

FIGURE 4.3 – Distribution empirique des résidus groupés de Pearson - Poisson - DDE

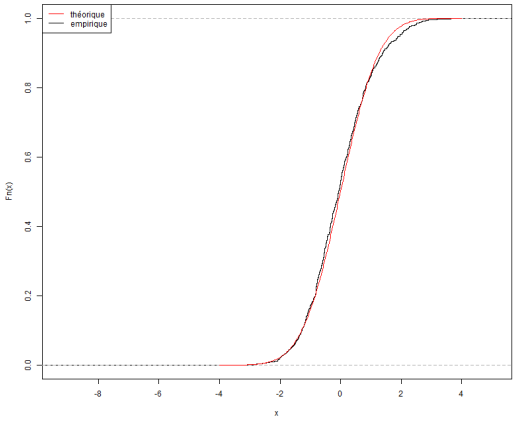

FIGURE 4.4 – Fonction de répartition empirique des résidus de Pearson - Poisson - DDE

L'étude des graphiques précédents montre que l'approximation par une loi normale de ces résidus est possible.

Pour parfaire l'étude des résidus, il faut aussi s'intéresser aux résidus quantiles randomisés normalisés. Ces résidus ont pu être déterminés en appliquant la fonction de répartition inverse d'une loi normale standard à la fonction de répartition empirique des valeurs prédites par le modèle.

Si le modèle est adapté alors ces résidus quantiles doivent suivre une loi normale standard.

Ces résidus quantiles randomisés normalisés sont représentés dans les deux graphiques suivants :

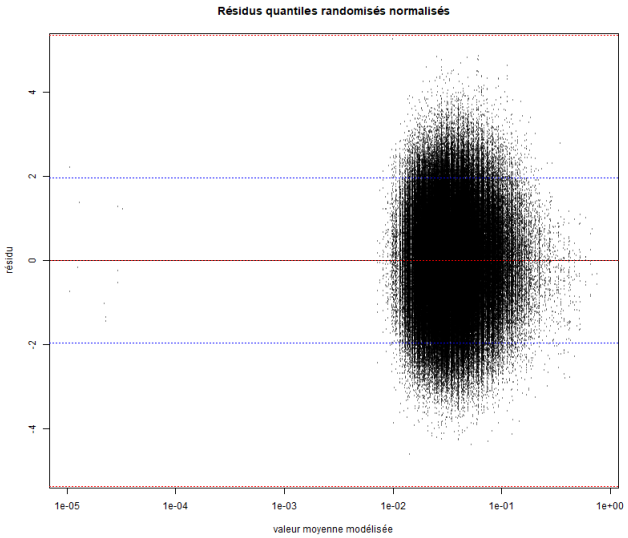

FIGURE 4.5 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - DDE

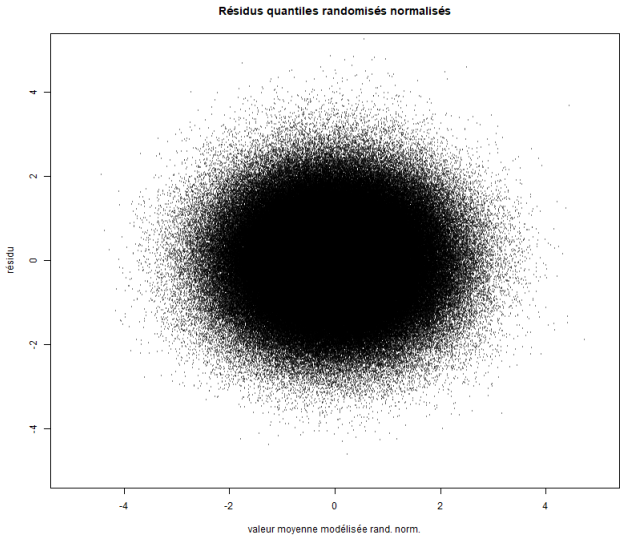

FIGURE 4.6 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisés normalisés) - Poisson - DDE

La construction de deux graphiques comparant la distribution et la fonction de répartition des résidus avec celles de la loi normale, permet de valider l'hypothèse que les résidus suivent une loi normale standard. (Figures 4.7 et 4.8)

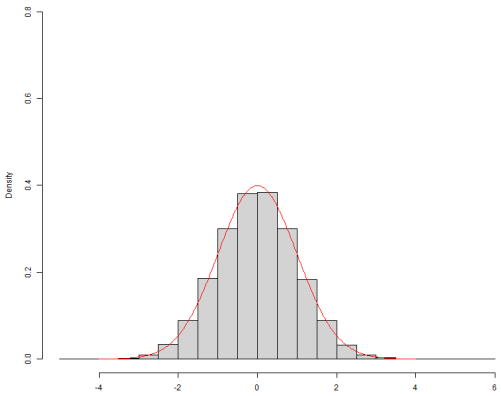

FIGURE 4.7 – Distribution empirique des résidus quantiles randomisés normalisés - Poisson - DDE

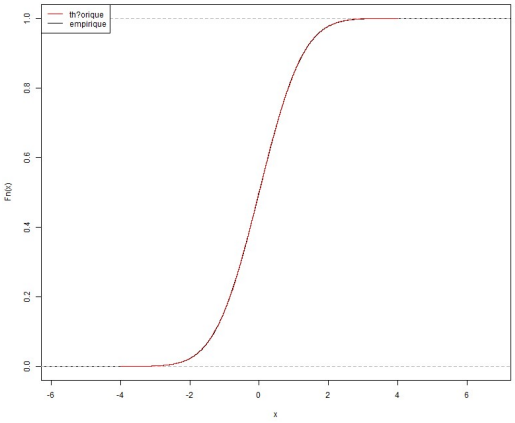

FIGURE 4.8 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - DDE

Au travers de l'étude des graphiques précédents, aucune tendance spécifique ne se dégage et l'approximation de la loi de distribution des résidus quantiles par la loi normale standard semble être logique. Ceci conclu l'analyse des résidus pour ce modèle de fréquence.

4.1.4 Validation des modèles et sélection du GLM le plus efficace

Afin de pouvoir faire la sélection des modèles adaptés à nos garanties, il faut vérifier l'adéquation des valeurs prédites et des valeurs observées sur l'échantillon d'apprentissage mais aussi la capacité prédictive du modèle sur une base de validation dont les valeurs n'auront pas servi à sa calibration.

Pour chacune des garanties sont regardées les mesures de performances sur les échantillons d'apprentissage et de validation pour ensuite analyser les courbes de Lorenz (cf. partie 2.5.2 pour rappel sur la courbe de Lorenz) et enfin, émettre une conclusion sur le GLM le plus efficace pour chaque garantie.

Pour la fréquence BDG :

Les mesures de performances établies pour ce modèle sont regroupées dans le tableau suivant :

	Poisson		Binomiale Négative	
	Apprentissage	Validation	Apprentissage	Validation
Déviance	24 147.46	6 355.29	20 150.71	5 296.45
MSE	0.01019	0.01077	0.01019	0.01077
Statistique de Pearson	1 483 789	342 688	1 474 333.11	340 092.38

TABLE 4.4 – Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie BDG

L'analyse des résultats montre que les MSE des différents modèles sont presque égales. Mais en comparant les deux autres indicateurs, il est à noter que la déviance et la statistique de Pearson sont meilleures dans le cas du modèle Binomial Négatif. Il semble donc que ce modèle soit le mieux adapté.

Pour approfondir cette étude, il faut maintenant analyser les courbes de Lorenz des différents GLM qui sont regroupées sur le graphique ci-dessous :

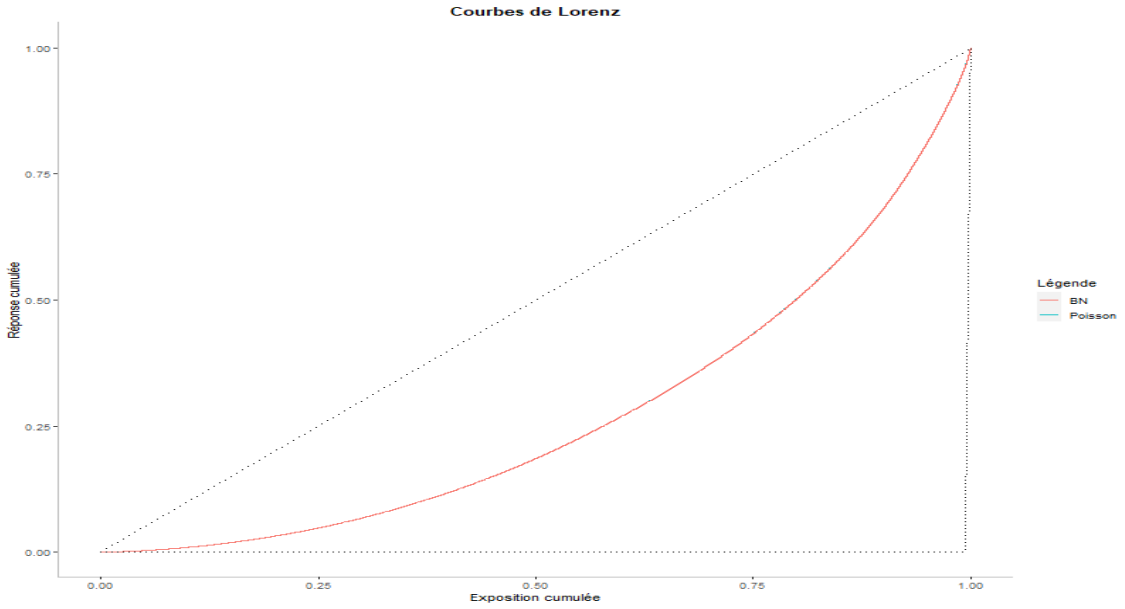

FIGURE 4.9 – Comparaison des courbes de Lorenz - GLM - BDG

Ces courbes permettent d'illustrer une bonne efficacité des modèles dans la segmentation de la modélisation de la fréquence. Cependant, elles sont indifférentiables sur le graphique 4.9. Il faut donc étudier les indices de Gini afin de détecter si l'un des modèles est un peu plus performant que l'autre.

	Poisson	**Binomiale Négative**
Indice de Gini	0.4628807	0.4629143

TABLE 4.5 – Tableau des indices de Gini - Comparaison GLM - BDG

Il apparaît, à partir du tableau précédent, que le GLM utilisant la loi Binomiale Négative a un indice de Gini légèrement meilleur. Cet indice seul ne permet pas de conclure sur le modèle le plus performant. Mais en prenant en compte les résultats des statistiques précédentes, il est possible de porter le choix du modèle le plus performant sur le modèle Binomial Négatif.

Pour la fréquence DDE :

Comme pour la fréquence précédente, un tableau contenant les mesures de performances des différents GLM testés est établi.

	Poisson		Binomiale Négative	
	Apprentissage	Validation	Apprentissage	Validation
Déviance	108 089	27 086	88 211	22 119
MSE	0.04493	0.0442	0.04493	0.04421
Statistique de Pearson	2 027 447	500 892	1 988 075	491 479

TABLE 4.6 – Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie DDE

L'analyse du tableau ci-dessus montre que le modèle utilisant la loi de Poisson est imperceptiblement plus performant que celui utilisant la loi Binomiale Négative, au niveau de la MSE. Cependant, la déviance ainsi que la statistique de Pearson conduisent de manière plus flagrante à la conclusion que le GLM utilisant la distribution Binomiale Négative est plus pertinent dans la modélisation.

Les courbes de Lorenz de ces différents modèles sont comparées, de la même façon que pour la garantie BDG.

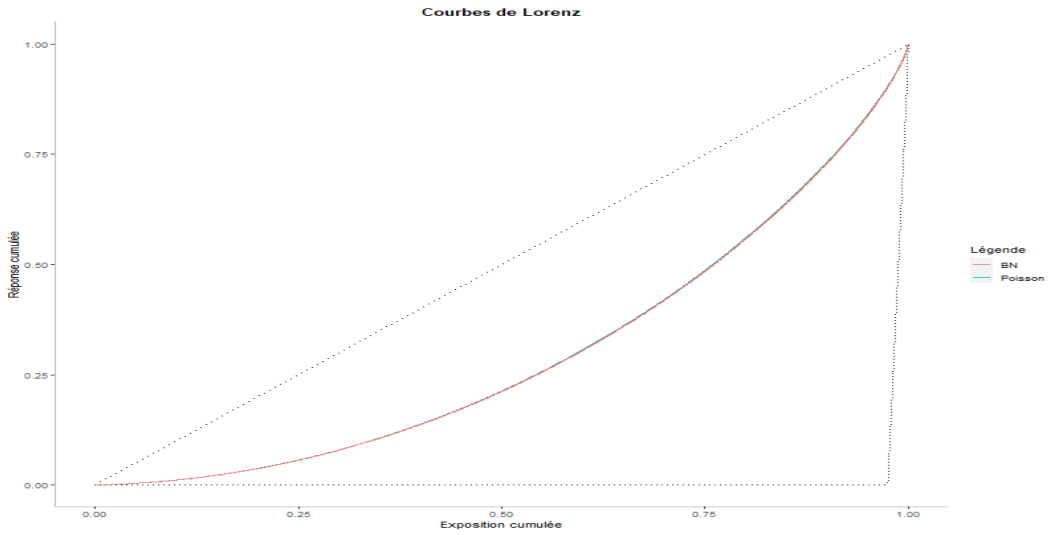

FIGURE 4.10 – Comparaison des courbes de Lorenz - GLM - DDE

Tout comme dans le cas étudié précédemment, il est impossible de comparer directement les courbes de Lorenz, sur le graphique. Néanmoins, il est possible de constater que les modèles permettent de s'éloigner de la mutualisation égale de la fréquence sur le portefeuille et de dégager une discrimination en fonction des individus.

Un nouveau tableau regroupant les différents indices de Gini est donc réalisé afin de voir si ces derniers donnent une information analogue à ce que le tableau 4.6 a fourni.

	Poisson	**Binomiale Négative**
Indice de Gini	0.417795	0.4197186

TABLE 4.7 – Tableau des indices de Gini - Comparaison GLM - DDE

Il apparaît que le modèle utilisant la loi Binomiale Négative est plus performant que celui avec la loi Poisson. Il est possible de remarquer que, pour cette garantie, la différence, au niveau de l'indice de Gini, est plus significative que dans le cas de la garantie BDG.

Par les deux analyses effectuées (mesures de performances et indice de Gini), le choix du modèle GLM Binomial Négatif est retenu dans le cadre de l'étude pour cette garantie.

Pour la fréquence incendie :

Le tableau des mesures de performances sur les échantillons d'apprentissage et de validation est réalisé et consigné ci-dessous :

	Poisson		Binomiale Négative	
	Apprentissage	Validation	Apprentissage	Validation
Déviance	16 405.16	4 697.64	14 467.63	4 126.94
MSE	0.00427	0.00526	0.00427	0.00525
Statistique de Pearson	2 591 394.82	944 845.27	2 409 494	980 529.13

TABLE 4.8 – Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie Incendie

Les mesures de performances des GLM sont comparativement assez similaires. Cependant, sur le seul critère de cette analyse, le modèle supposant une loi de distribution Binomiale Négative serait retenu.

Afin d'approfondir l'étude pour cette fréquence, les différentes courbes de Lorenz des modèles doivent être analysées.

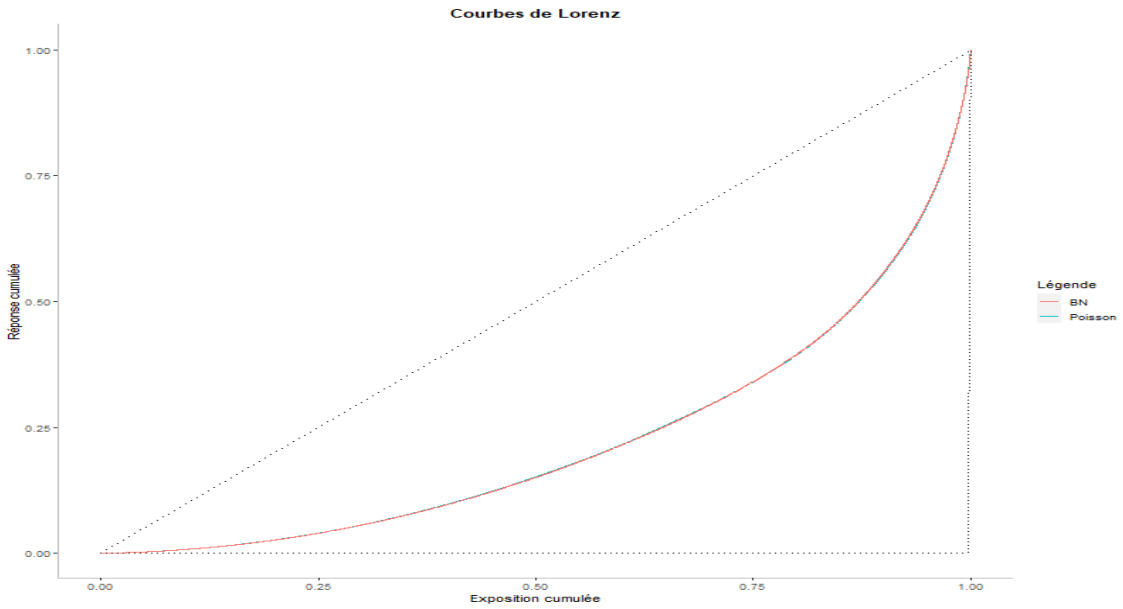

FIGURE 4.11 – Comparaison des courbes de Lorenz - GLM - Incendie

Les courbes étant très proches l'une de l'autre, il faut tenir compte des indices de Gini consignés dans le tableau ci-dessous :

	Poisson	**Binomiale Négative**
Indice de Gini	0.5575089	0.5577626

TABLE 4.9 – Tableau des indices de Gini - Comparaison GLM - Incendie

Par ce tableau, l'hypothèse formulée par l'analyse des indicateurs, concluant que le GLM utilisant la loi Binomiale Négative est plus performant, se confirme. Il est aussi possible d'établir que les deux modèles ont une segmentation efficace qui permet de se rapprocher de la situation présente sur le portefeuille (et non pas de la situation où tous les individus auraient la même fréquence estimée).

En conclusion, le choix est fait de sélectionner le GLM supposant que la loi de distribution est la loi Binomiale Négative.

Pour la fréquence Vol :

Le tableau des performances des GLM pour la garantie Vol est illustré ci-dessous :

	Poisson		Binomiale Négative	
	Apprentissage	Validation	Apprentissage	Validation
Déviance	19 049.36	4 887.43	14 852.14	3 819
MSE	0.00713	0.0071	0.00713	0.0071
Statistique de Pearson	1 382 473	365 955.56	1 371 285	363 419.42

TABLE 4.10 – Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie Vol

La différence entre les MSE des deux modèles n'est absolument pas significative. Les deux autres indicateurs de performance vont permettre de déduire que le modèle Binomiale Négatif est plus performant que l'autre modèle testé.

Comme pour les autres garanties, l'étude se poursuit en regardant les courbes de Lorenz, regroupées sur le graphique suivant :

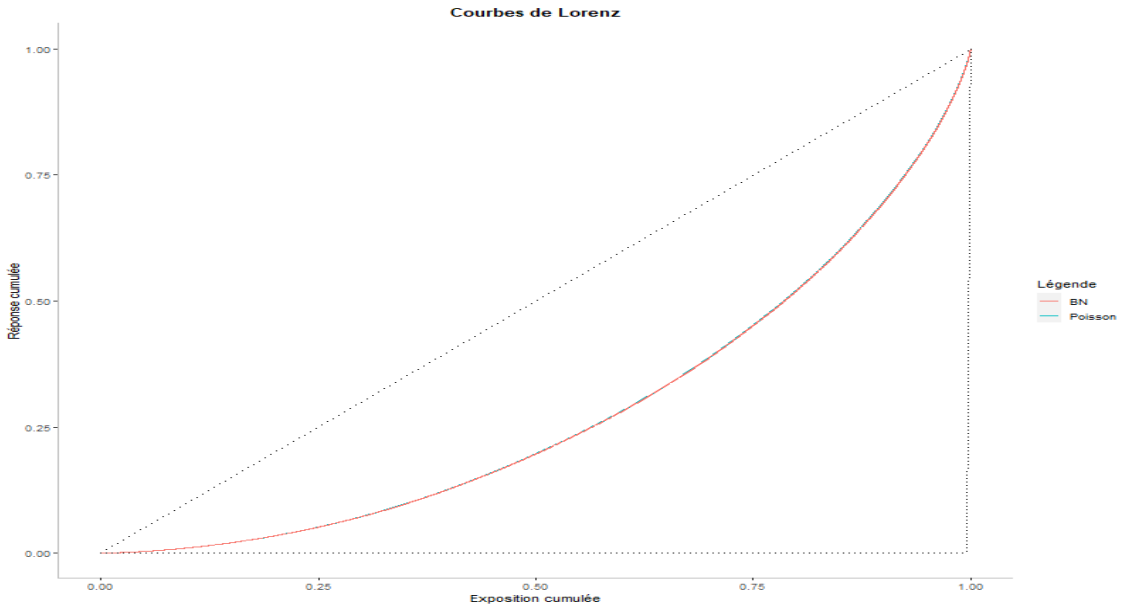

FIGURE 4.12 – Comparaison des courbes de Lorenz - GLM - Vol

Au travers des courbes, la capacité discriminante des modèles de fréquence concernant la garantie Vol est validée. Cependant, il faut remarquer que la meilleure performance ne peut pas être déduite uniquement du graphique. Ainsi, les indices de Gini sont réunis dans le tableau suivant.

	Poisson	**Binomiale Négative**
Indice de Gini	0.4439663	0.4451197

TABLE 4.11 – Tableau des indices de Gini - Comparaison GLM - Vol

Les deux indices de Gini obtenus sont proches. Toutefois, si un modèle devait être choisi en utilisant ce seul indicateur, ce serait le GLM Binomial Négatif. Cette conclusion vient corroborer celle des indicateurs de performance faite plus haut.

Ainsi, pour la garantie Vol, le choix se porterait sur le modèle utilisant la distribution sous-jacente Binomiale Négative.

Pour la fréquence RC :

Le tableau des mesures de performances est le suivant pour cette fréquence RC :

	Poisson		Binomiale Négative	
	Apprentissage	Validation	Apprentissage	Validation
Déviance	29 643.72	6 993.16	22 811.09	5 408.78
MSE	0.00854	0.00774	0.00854	0.00774
Statistique de Pearson	2 213 107	589 882.6	2 191 810	585 348.59

TABLE 4.12 – Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie RC

En comparant les MSE des différents modèles simulés, il apparaît qu'aucun d'entre eux ne se démarque. Les autres indicateurs calculés permettent de conclure que le modèle à conserver est celui avec la loi Binomiale Négative.

Pour pousser l'analyse effectuée, les courbes de Lorenz pour les deux modèles sont tracées.

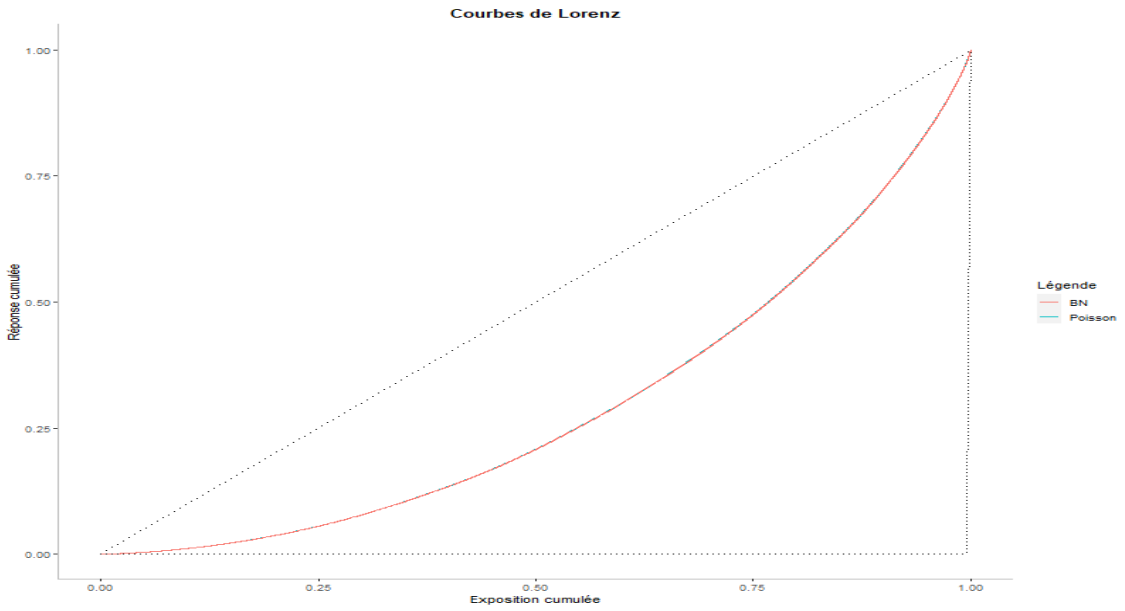

FIGURE 4.13 – Comparaison des courbes de Lorenz - GLM - RC

Le constat est que, bien que la performance des modèles ne soit pas maximale, la segmentation des individus est pertinente puisque les courbes sont éloignées de la courbe de mutualisation de la fréquence. Aucune conclusion satisfaisante (concernant la supériorité de performance de l'un des modèles) n'est possible en regardant ces courbes puisqu'elles restent très proches l'une de l'autre. Il faut donc procéder à l'analyse des indices de Gini.

	Poisson	Binomiale Négative
Indice de Gini	0.4178024	0.4185361

TABLE 4.13 – Tableau des indices de Gini - Comparaison GLM - RC

Ces derniers, bien que presque similaires, montrent, tout comme les indicateurs déjà calculés, que le modèle utilisant la loi Binomiale Négative est un peu plus performant que celui utilisant la distribution de Poisson.

En conclusion générale de l'ensemble de ces différentes analyses, les choix pour la validation des modèles les plus efficaces en fonction de chaque garantie sont regroupés dans le tableau récapitulatif suivant :

	DDE	Incendie	Vol	BDG	RC
Modèle Sélectionné	BN	BN	BN	BN	BN

TABLE 4.14 – Tableau des GLM les mieux adaptés, par garantie

Dans un objectif de recherche d'optimisation de la performance de la modélisation de chaque garantie, il est décidé d'estimer ces fréquences grâce à des modèles de Data Science se basant sur les arbres. Pour commencer, les modèles CART vont être appliqués à chacune des garanties.

4.2 Arbre CART

Pour rappel de la partie théorique des modèles CART, les étapes sont la création de l'arbre maximal (avec une fréquence spécifique à un risque de notre base), l'élagage de l'arbre et les analyses nécessaires pour la comparaison des modèles.

Ces étapes sont effectuées pour les cinq garanties tarifées par la modélisation.
Les différents arbres maximaux obtenus étant rendus illisibles de par leur taille et par le format de ce mémoire, le choix est fait de ne pas mettre d'illustrations complètes de ces arbres et de ne montrer que les arbres finaux.
Pour chacune des garanties, le tableau suivant regroupe les différents cp (paramètres de complexité - *complexity parameters*) utilisés pour l'élagage.

	DDE	Incendie	Vol	BDG	RC
cp	8.341868e-06	0.0001899871	4.765841e-05	5.84139e-05	3.255545e-05

TABLE 4.15 – Tableau regroupant les différents cp utilisés pour l'élagage - CART

Des arbres sont obtenus pour chacune des garanties. Dans un souci de simplicité, seul l'arbre final concernant la garantie DDE est présenté. Les représentations graphiques des autres arbres finaux se trouvent en annexe J.

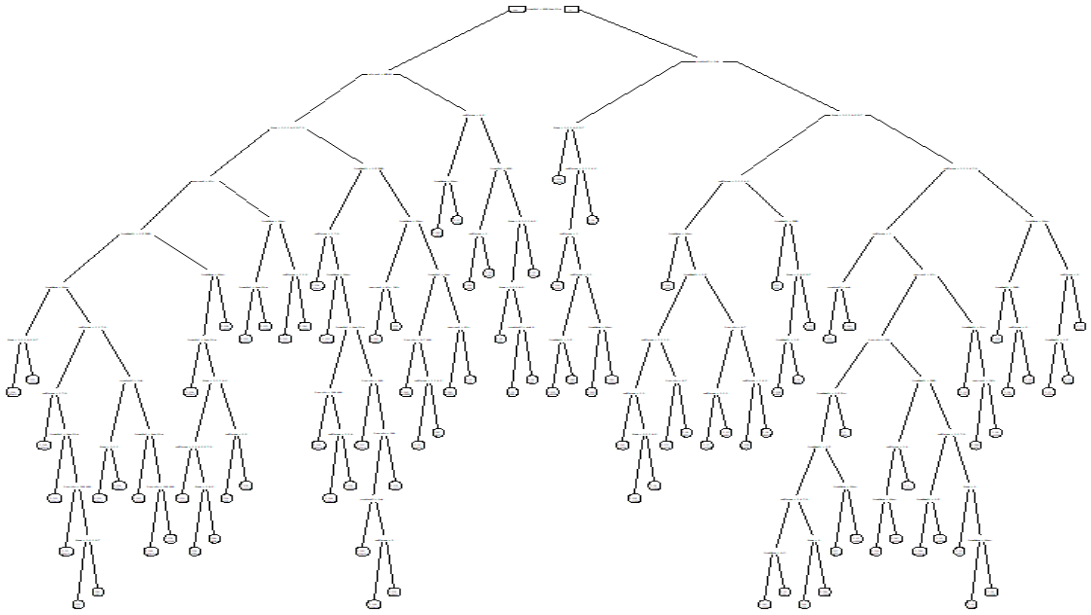

FIGURE 4.14 – Arbre final fréquence pour la garantie DDE

Pour permettre une meilleure visualisation des arbres, la lecture de ceux-ci est explicitée par le schéma ci-dessous. (Figure 4.15)

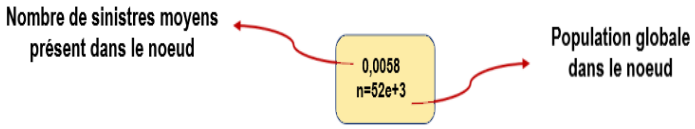

FIGURE 4.15 – Lecture de l'arbre

La pertinence des modèles s'établit, d'une part par le calcul de la MSE et d'autre part, par la création des courbes de Lorenz pour chacune des cinq garanties.

Tout d'abord, les MSE sur la base d'apprentissage et sur la base de validation se décomposent selon le tableaux suivant :

	DDE	Incendie	Vol	BDG	RC
Apprentissage	0.04476	0.00427	0.00712	0.01018	0.00854
Validation	0.04413	0.00524	0.0071	0.01077	0.00774

TABLE 4.16 – Tableau récapitulatif des MSE sur les bases - CART

Ce tableau montre que les modèles obtenus en utilisant les arbres de régression CART sont autant efficaces pour minimiser la MSE que la modélisation GLM (et dans le cas des garanties DDE et Incendie légèrement plus efficace).

Pour finir l'analyse des modèles CART réalisés, leur comparaison avec les GLM, au travers des courbes de Lorenz, est présentée ci-dessous :

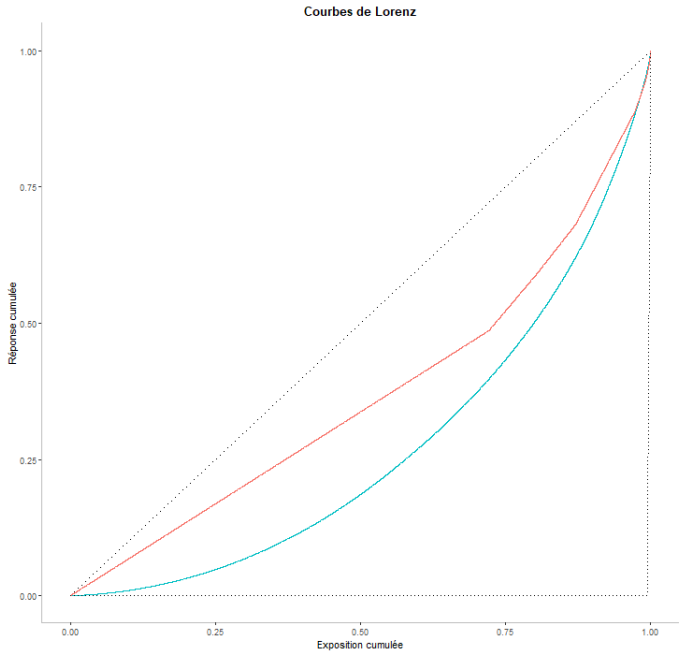

FIGURE 4.16 – Comparaison des courbes de Lorenz GLM/CART pour la garantie BDG

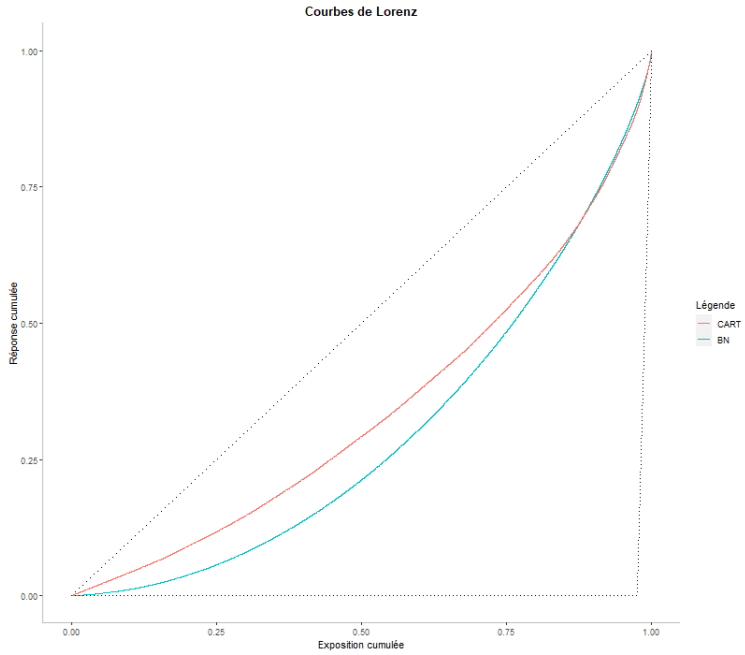

FIGURE 4.17 – Comparaison des courbes de Lorenz GLM/CART pour la garantie DDE

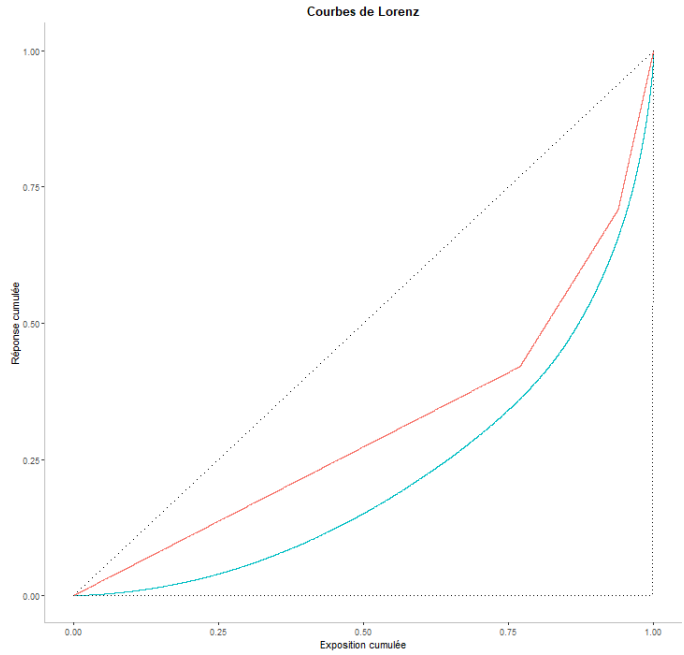

FIGURE 4.18 – Comparaison des courbes de Lorenz GLM/CART pour la garantie Incendie

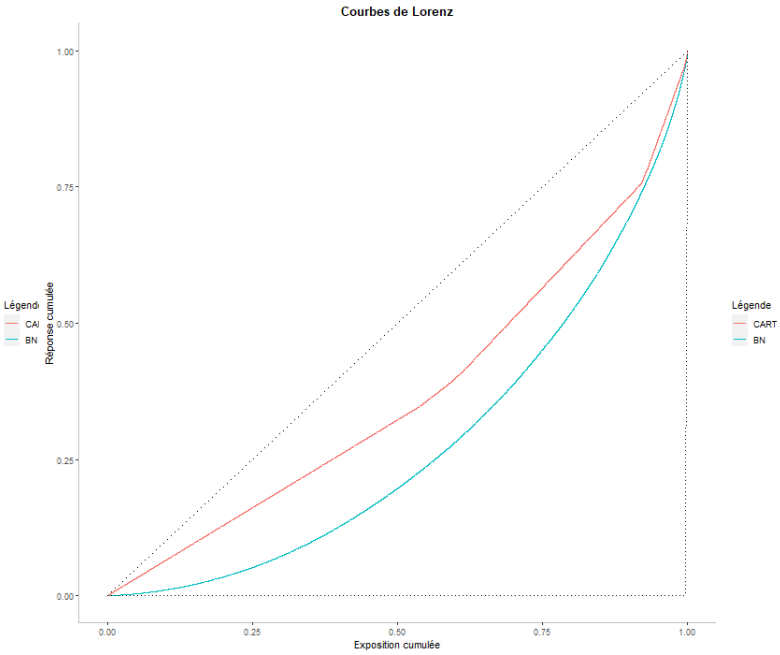

FIGURE 4.19 – Comparaison des courbes de Lorenz GLM/CART pour la garantie Vol

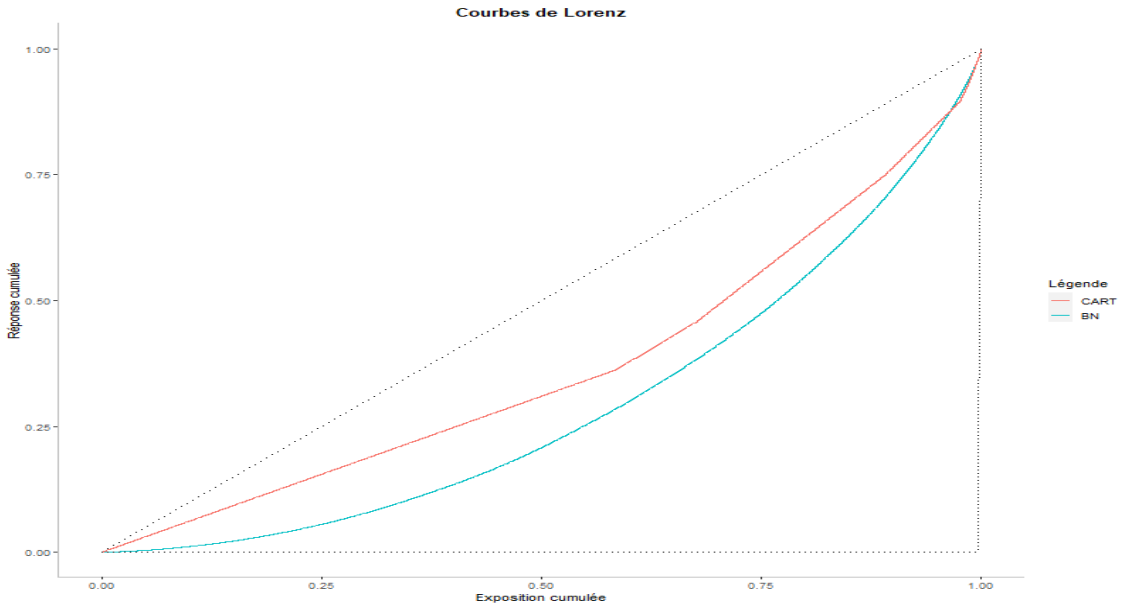

FIGURE 4.20 – Comparaison des courbes de Lorenz GLM/CART pour la garantie RC

Pour toutes les garanties, les courbes de Lorenz des modèles CART restent plus proches de la courbe représentant la mutualisation égale sur le portefeuille. Ce résultat, qui vient s'ajouter au constat fait plus haut sur les MSE , permet de déduire qu'il est préférable de conserver les GLM trouvés dans la partie précédente quelque soit la garantie sélectionnée.

Pour conclure cette partie, il apparaît que modéliser par CART la fréquence de nos garanties ne permet pas d'améliorer l'estimation, même si la MSE moyenne reste proche de celle obtenue par le GLM. Les modèles CART n'étant pas les modèles les plus performants utilisant des arbres, il faudra envisager de tester d'autres modèles plus poussés afin d'améliorer les résultats des modèles de *Data Science* et peut être surpasser l'efficacité des GLM.

4.3 Les Forêts Aléatoires

Pour rappel, les forêts aléatoires sont un type de méthode d'apprentissage d'ensemble pour la classification et la régression. L'idée derrière *Random Forest* est de combiner plusieurs arbres de décision, qui sont formés sur des sous-ensembles de données sélectionnés au hasard, pour créer un modèle plus robuste.

Chaque arbre de décision dans une forêt aléatoire fait une prédiction, et la prédiction finale est faite en faisant la moyenne des prédictions de tous les arbres. Cette approche permet de réduire le surajustement, car les erreurs commises par chaque arbre sont moyennées, et augmente également la stabilité du modèle en réduisant la variance des prédictions.

Méthodologie

Afin d'éviter un surapprentissage lors de la création de la forêt aléatoire et dans le but d'obtenir des résultats optimaux, la première étape à effectuer est l'hyperparamétrage des paramètres de la fonction.

La méthode utilisée pour cette étape d'obtention des paramètres optimaux est un *Grid Search*. Les paramètres suivant sont testés :

- Le paramètre du nombre d'arbres : un nombre d'arbres entre 20 et 700 par pas de 20 est testé
- Le paramètre nombre de variables à tester à chaque noeud : il est choisi de tester les valeurs 3, 5 et le nombre maximum de variables explicatives pour la garantie (attention : ces valeurs ont été optimisées pour certaines des modélisations coûts)
- Un paramètre influant sur la taille d'un arbre : les valeurs 20, 50 et 100 sont testées

Cette étape achevée, l'importance des variables est recherchée afin de palier à l'effet "boîte noire" de la méthode, en permettant une première visualisation des variables les plus influentes du modèle ainsi créé. Pour chaque arbre et par variable explicative, les individus *Out-Of-Bag* vont avoir leur modalité qui va être permutée avec une autre et ensuite, l'algorithme va déterminer l'erreur quadratique moyenne (MSE) pour ces individus désorganisés. L'importance de la variable sur l'ensemble de la forêt est trouvée en faisant la moyenne des différences entre MSE avant et après permutation, qui est ensuite divisée par l'écart-type des différences s'il n'est pas nul. Plus cette différence est grande, plus la variable est considérée influente dans le *Random Forest*.

Enfin, la pertinence de l'estimation est étudiée et comparée à celle obtenue par le GLM.

Application aux données

En appliquant la méthode *Grid Search* sur les différents modèles de fréquence, les paramètres suivants ont été choisis parmi les 315 modèles testés :

	DDE	Incendie	Vol	BDG	RC
Nombres d'arbres	420	500	480	60	480
Nombre de variables	3	3	3	3	3
Taille de l'arbre	100	50	20	20	20

TABLE 4.17 – Récapitulatif des paramètres optimaux - Random Forest - Fréquence

Les évolutions de la RMSE en fonction du nombre d'arbres pour les forêts optimales conservées sont présentes dans l'annexe K. L'ensemble de ces graphiques permet de déduire que la décision, concernant le paramètre du nombres d'arbres, justifie l'efficacité du modèle. En effet, il peut être constaté une stabilisation avant le nombre d'arbres optimal sélectionné.

Les graphiques permettant de visualiser l'importance des variables dans la modélisation sont également consignés dans l'annexe K, pour chaque garantie. Dans le cas de la garantie Incendie, il est constaté sur le *Random Forest* créé que la variable liée au capital mobilier par pièce avait une significativité négative. Ainsi la permutation qui a été faite dans les modalités permet une meilleure précision que celle utilisant la variable sans permutation. Il a donc été décidé de réinitialiser le modèle pour cette garantie en enlevant la variable avec l'importance négative. Le tableau ci-dessous résume les trois variables ayant le plus d'importance pour chacune des garanties.

	Variable 1	Variable 2	Variable 3
DDE	sin_ant	Zone	typedistribution
Incendie	nbPiecesPrincipales	typeHabetage	nboptions
Vol	Zone	typedistribution	typeHabetage
BDG	fran_globale	bdgprisoption	sin_ant
RC	nbPiecesPrincipales	sin_ant	typedistribution

TABLE 4.18 – Variables les plus importantes - RF - Fréquence

Pour finir l'étude, la pertinence du modèle, pour chaque garantie, est étudiée dans les tableaux suivants, regroupant d'une part les MSE calculées sur la base d'apprentissage et de validation et d'autre part les indices de Gini des modèles *Random Forest* et GLM.

	DDE	Incendie	Vol	BDG	RC
Apprentissage	0.04466	0.00426	0.00712	0.01017	0.00853
Validation	0.04404	0.00525	0.00709	0.01076	0.00774

TABLE 4.19 – Tableau récapitulatif des MSE sur les bases - RF

Les MSE obtenues par les modèles *Random Forest* sont équivalentes à celles des modèles GLM.

	DDE	Incendie	Vol	BDG	RC
Modèle RF	0.2647458	0.4352728	0.2802304	0.2471913	0.258314
Modèle GLM	0.4197186	0.5577626	0.4451197	0.4629143	0.4185361

TABLE 4.20 – Tableau de comparaison des indices de Gini (modèle GLM/modèle RF)

Les indices de Gini présents dans le tableau précédent viennent en appui pour étayer le constat que les méthodes de *Random Forest* ne sont pas aussi efficaces que les modèles GLM lors de la modélisation des fréquences pour les garanties considérées. En effet, l'ensemble des indices de Gini obtenus par les modélisations de cette partie est significativement inférieur aux indices obtenus par la modélisation GLM. Ainsi, la capacité discriminante des modèles *Random Forest* est moins prononcée par rapport aux modèles GLM.

Pour clôturer cette partie, il est possible de déduire que le modèle *Random Forest* n'est pas aussi efficace que le GLM pour la modélisation de la fréquence des garanties étudiées. Un autre modèle de *Data Science* utilisant les arbres va être appliqué pour estimer les fréquences : le XGBoost.

4.4 eXtreme Gradient Boosting : XGBoost

Méthodologie

Lors du lancement de l'algorithme *XGBoost*, il est tout d'abord impératif de rendre binaires les variables de type catégoriel, afin de permettre l'exécution de la fonction dans le logiciel R.

Pour optimiser les résultats obtenus par l'algorithme, un hyperparamétrage est nécessaire. Comme pour la méthode *Random Forest*, l'obtention des paramètres optimaux passe par un *Grid Search* avec les hypothèses suivantes :

- Le paramètre fixant le taux d'apprentissage : valeurs testées {0.01, 0.001, 0.05, 0.1}
- Le paramètre fixant la profondeur des arbres : valeurs testées {10 , 30, 50, 70, 100}
- Le paramètre fixant le pourcentage de variables utilisées lors de la création des arbres : valeurs testées {0.3, 0.5, 1}

Le nombre d'itérations (ou d'arbres) utilisés est fixé à 1 000. Afin de contrer un éventuel surapprentissage du modèle, il est décidé de prendre 50 comme *early stopping round*, c'est-à-dire qu'au bout de 50 itérations s'il n'y a pas d'amélioration de la qualité des arbres, l'algorithme s'arrête.

Dans le but de réaliser une première visualisation des effets des variables tarifantes sur le modèle créé, l'importance des variables est étudiée. A chaque noeud, un gain est calculé, représentant la contribution de la variable sélectionnée. Pour obtenir les effets des variables, les contributions de tous les *split* pour tous les arbres sont agrégées et sommées, par variable.

Enfin, la pertinence du modèle obtenue est comparée à celle des GLM.

Application aux données

Parmi les 90 modèles testés, la méthode *Grid Search* a permis de sélectionner les paramètres suivant comme étant optimaux :

	DDE	Incendie	Vol	BDG	RC
Taux d'apprentissage	0.05	0.05	0.05	0.05	0.1
Profondeur de l'arbre	30	10	10	30	30
Pourcentage de variables	1	1	1	1	1

TABLE 4.21 – Récapitulatif des paramètres optimaux - XGBoost - Fréquence

Afin de visualiser, pour ces cinq modèles, l'influence des variables explicatives pour le modèle optimal, le tableau suivant réuni les trois variables considérées comme étant les plus significatives pour chaque modèle. Le graphique des importances de toutes les variables significatives se trouve en annexe L.

	Variable 1	Variable 2	Variable 3
DDE	Distribution Internet	Zone 9	Franchise 200-300
Incendie	Maison	Propriétaire - Principal	Capital Mobilier inf 6
Vol	Capital Mobilier sup 4	Zone 9	Distribution hors Internet
BDG	Franchise 0-75	Locataire - Principal	Zone 8-9-10
RC	Appartement/ Chambre	Franchise 200-300-400	Moins de 3 options

TABLE 4.22 – Variables les plus importantes - XGBoost - Fréquence

Pour commencer l'analyse de la pertinence des modèles *XGBoost*, le tableau contenant les MSE sur les bases d'apprentissage et de validation est créé :

	DDE	Incendie	Vol	BDG	RC
Apprentissage	0.04437	0.00422	0.00711	0.01007	0.00852
Validation	0.0443	0.00526	0.0071	0.0108	0.00775

TABLE 4.23 – Tableau récapitulatif des MSE sur les bases - XGBoost

En comparant les MSE avec celles obtenues par les méthodes GLM, il est possible de remarquer que la méthode *XGBoost* a des MSE similaires aux modèles GLM dans le cas de la garantie Vol. Dans les autres cas, la MSE augmente mais pas de manière significative.

Pour finir, une comparaison des indices de Gini est effectuée dans le tableau suivant :

	DDE	Incendie	Vol	BDG	RC
Modèle XGBoost	0.4433974	0.5213907	0.4528791	0.4865415	0.4470658
Modèle GLM	0.4197186	0.5577626	0.4451197	0.4629143	0.4185361

TABLE 4.24 – Tableau de comparaison des indices de Gini (modèle GLM/modèle XGBoost)

Pour la garantie Incendie, l'indice de Gini est plus faible que pour le modèle GLM le plus performant. En combinant ce résultat au constat fait sur la comparaison des MSE, le modèle *XGBoost* pour cette garantie apparaît moins efficace que le modèle GLM.

Dans le cas de la garantie Vol, l'indice de Gini présente une augmentation relative d'environ 2% par rapport à celui du cas GLM. Cependant, la MSE reste stable. Cette faible amélioration de l'indice de Gini n'est de plus pas assez significative pour pallier la perte de flexibilité et d'interprétabilité au niveau du modèle XGBoost par rapport au GLM, qui reste donc le plus performant dans ce cas.

Dans le cas des garanties BDG, DDE et RC, les augmentations relatives de l'indice de Gini sont d'approximativement $5 - 7\%$ pour une augmentation relative de moins de 1% de la MSE. Le modèle GLM reste là aussi le plus efficace pour ces garanties (gain de l'indice de Gini insuffisant par rapport à la perte de flexibilité du modèle).

En conclusion de ce chapitre 4, au vu des comparaisons des MSE et indices de Gini obtenus, les modèles *XGBoost* n'ont pas une efficacité suffisante pour dépasser celle des modèles GLM dans l'estimation des fréquences, dans le cadre de l'étude.

Le tableau suivant résume les modélisations les plus efficaces pour la fréquence, pour chacune des garanties de l'étude.

	DDE	Incendie	Vol	BDG	RC
Modèle Sélectionné	GLM - BN	GLM - BN	GLM - BN	GLM - BN	GLM - BN

TABLE 4.25 – Modèles de fréquence sélectionnés, par garantie

Chapitre 5

Application des modèles pour la modélisation du coût

Dans ce chapitre sont étudiés les résultats de l'application des modèles utilisés pour estimer le montant des coûts des sinistres. Dans les différentes parties, chaque modélisation est comparée à celle de référence (GLM).

L'intégralité des modélisations et analyses sous-jacentes sont effectuées avec le logiciel R.

5.1 Modèle Linéaire Généralisé

Comme pour la modélisation de la fréquence, la première étape consiste en une modélisation au travers des GLM, pour chacune des garanties. Les étapes sont identiques à celles de l'étude de la fréquence.

5.1.1 Avant la modélisation

Avant d'effectuer les modélisations, il est convenu de tester l'adéquation de deux lois par rapport à la distribution des coûts (par garantie).

Il est constaté, sur chaque garantie, à l'aide du même type de graphiques que ci-dessous (figure 5.1 et figure 5.2), que la loi Log-normale ainsi que la loi Gamma peuvent être adaptées afin de modéliser les coûts.

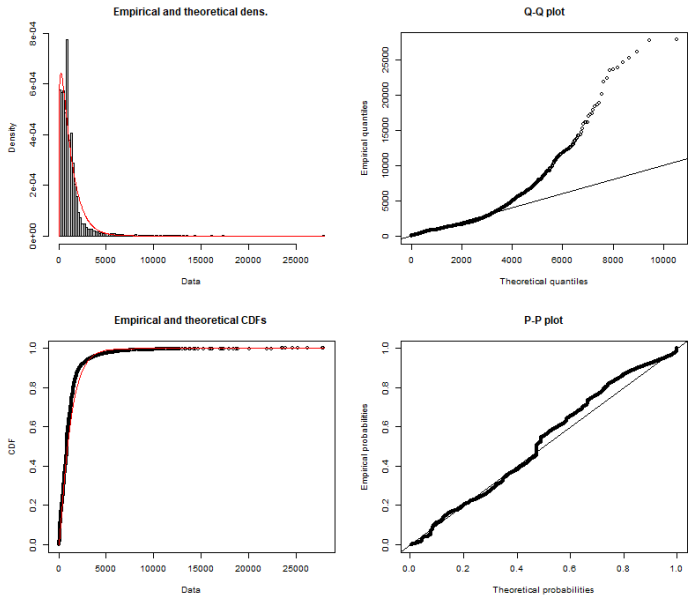

FIGURE 5.1 – Analyse de la distribution du coût DDE - loi Gamma

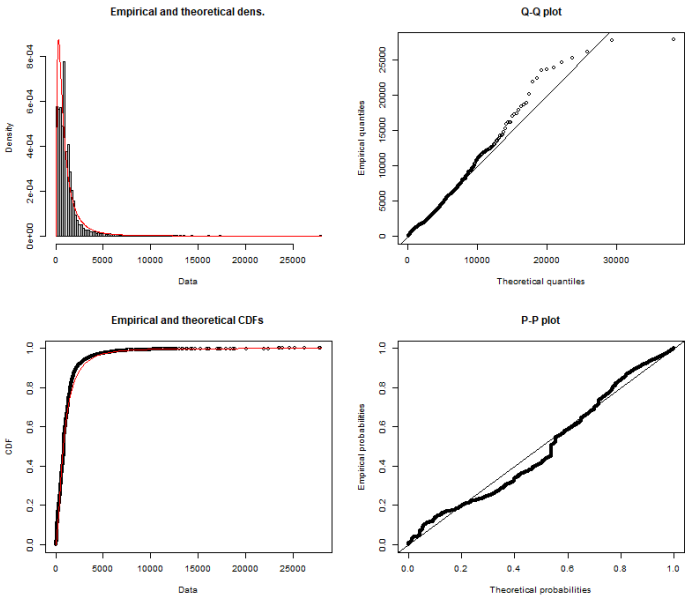

FIGURE 5.2 – Analyse de la distribution du coût DDE - loi Log-normale

A partir des graphiques d'analyse et des tableaux se trouvant dans la partie M.1 de l'annexe M, le modèle le mieux adapté est déterminé, avant de procéder à quelque amélioration que ce soit. Ces hypothèses sont réunies dans le tableau suivant :

	DDE	**Incendie**	**Vol**	**BDG**	**RC**
Modèle pré-sélectionné	Log-normale	Log-normale	Log-normale	Gamma	Log-normale

TABLE 5.1 – Tableau des GLM les plus adaptés par garantie (avant modélisation)

5.1.2 Choix des variables et calibration de leurs effets

Comme pour la modélisation des fréquences, le choix des variables est établi par les méthodes *Forward*, *Backward* et progressive. Après cette sélection, les tests de "type III", permettant de constater la significativité des variables sélectionnées avant regroupement des modalités, sont effectués.

Les différentes modalités des variables, sélectionnées par les méthodes précédentes, sont réunies en utilisant le même principe que dans le cas de la fréquence.

Pour la plupart des modélisations, soit très peu de variables sont sélectionnées, soit peu de modalités sont réellement significatives (dans certains cas, les deux). Cependant, toutes les variables (avec les regroupements de modalités) sont significatives. Ceci est mis en évidence dans des tableaux du même type que celui qui suit (où la garantie DDE avec la loi Log-normale sert d'exemple). Les autres tableaux sont regroupés en annexe M.

Variables	**Df**	**Statistique**	**p-value**
nbPiecesPrincipales	3	34.218	$1.782e-07$
typedistribution	2	54.506	$1.459e-12$
sin_ant	1	87.938	$< 2.2e-16$
fran_globale	2	41.685	$8.875e-10$
pourcentageOV	1	20.376	$6.361e-06$
typeResQualiSous	2	247.269	$< 2.2e-16$

TABLE 5.2 – Analyse de type III pour la loi de Log-normale et la garantie DDE

5.1.3 Analyse des résidus

L'étape suivante consiste à analyser les résidus. Cependant, en raison de la méthode mise en place lors de la modélisation du GLM avec la loi Log-normale [1], les résidus obtenus avec cette loi sont biaisés. En effet, ils ont tendance à laisser penser que la loi Log-normale est plus adaptée alors que ce n'est pas forcément le cas.

Pour illustrer ce phénomène, il faut observer les résidus obtenus pour la garantie DDE.

1. Une explication du processus derrière la création du GLM avec cette loi peut se retrouver en annexe G.

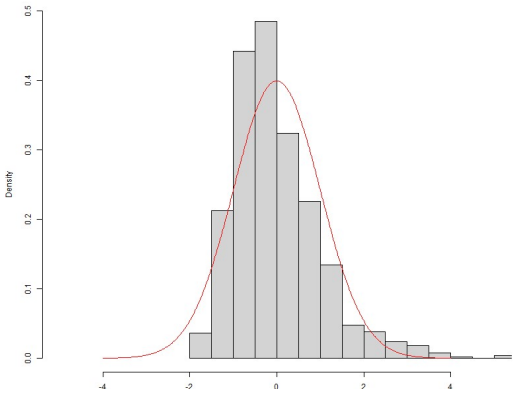

FIGURE 5.3 – Histogramme des Résidus de Pearson - Gamma - DDE

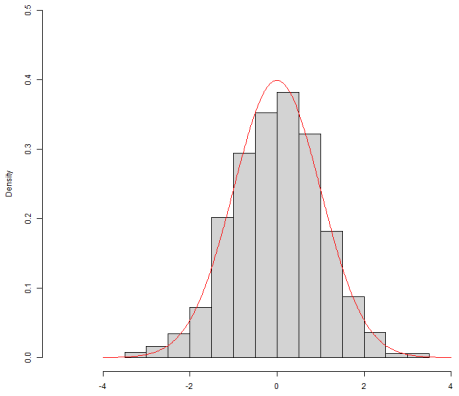

FIGURE 5.4 – Histogramme des Résidus de Pearson - Log-normale - DDE

Les histogrammes ci-dessus montrent que la loi Log-normale a effectivement des résidus qui suivent presque parfaitement une loi normale tandis que, dans le cas de la loi Gamma, une légère tendance à la sous-estimation globale du coût peut être observée.

Ces observations se confirment au travers de la visualisation des résidus suivants (figure 5.5 et figure 5.6). En effet, les résidus de Pearson dans le cas de la loi Log-normale sont, presque dans leur totalité, compris entre -2 et 2 et pour la loi Gamma, les points sortant de l'intervalle $[-2; 2]$ sont uniquement situés au dessus de 2. Ceci permet de conclure à une sous-estimation des valeurs par le modèle.

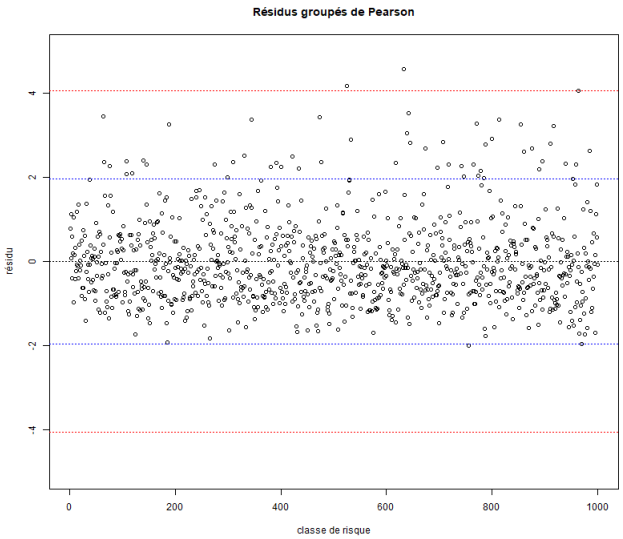

FIGURE 5.5 – Résidus de Pearson en fonction de la classe de risque - Gamma - DDE

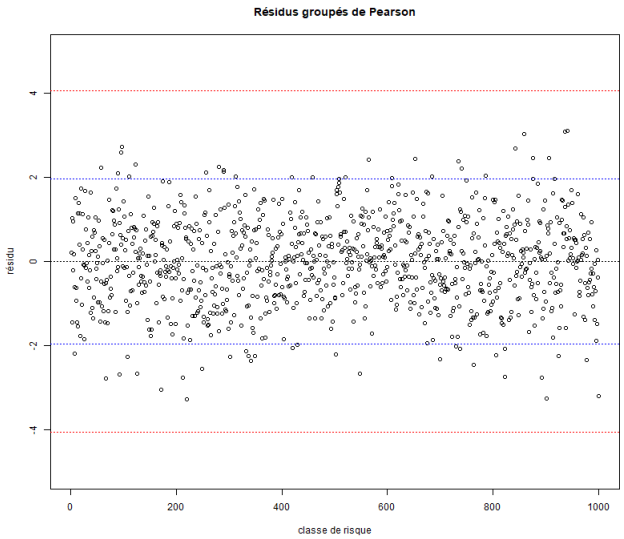

FIGURE 5.6 – Résidus de Pearson en fonction de la classe de risque - Log-normale - DDE

Le biais, lors de l'utilisation de la loi Log-normale, est un peu moins évident quand les résidus quantiles sont étudiés. Ce biais reste tout de même présent.

Au travers des histogrammes et des résidus quantiles représentés ci-dessous (figures 5.7, 5.8, 5.9 et 5.10), la sous-estimation des coûts avec la loi Gamma est encore plus prononcée. Cependant, ces résidus quantiles, dans la cas de la loi Log-normale, ne sont plus uniquement compris entre -2 et 2.

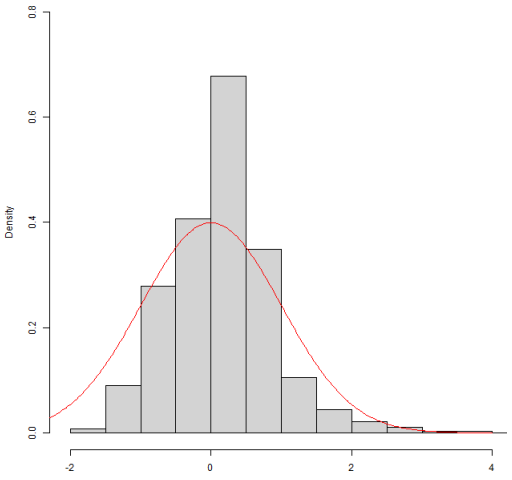

FIGURE 5.7 – Histogramme des Résidus Quantiles - Gamma - DDE

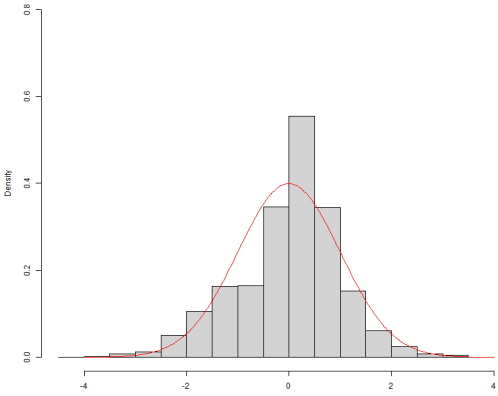

FIGURE 5.8 – Histogramme des Résidus Quantiles - Log-normale - DDE

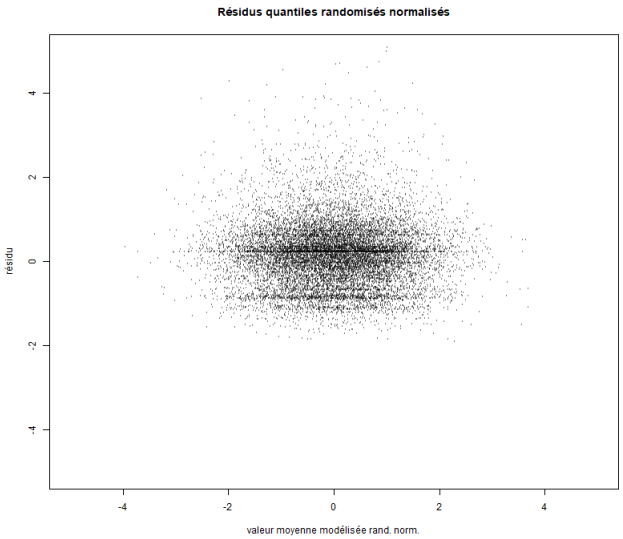

FIGURE 5.9 – Résidus des Quantiles - Gamma - DDE

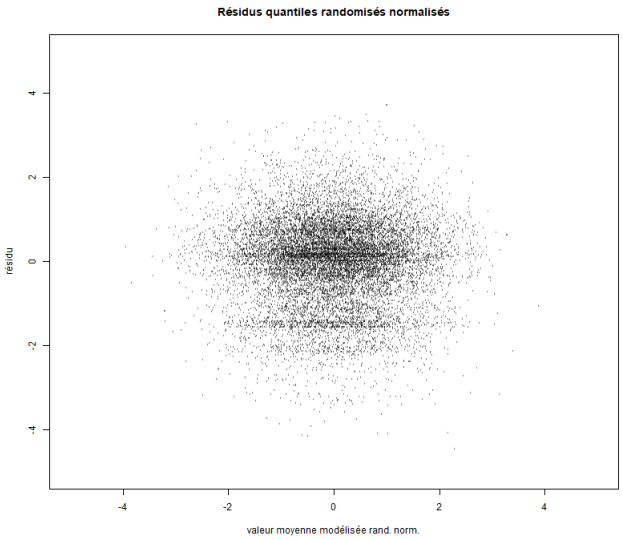

FIGURE 5.10 – Résidus des Quantiles - Log-normale - DDE

Les résidus de Pearson et les résidus quantiles ont été analysés pour l'ensemble des garanties. Les représentations graphiques de ces résidus (à l'exception de celles présentées dans cette partie) sont disponibles dans l'annexe N.

5.1.4 Sélection du GLM le plus efficace par garantie

Il est décidé, pour cette partie, de baser notre choix sur les RMSE et les résultats obtenus par la courbe de Lorenz.

Par garantie, des tableaux regroupant les RMSE pour les modèles Gamma et Log-normale sont créés.

De plus, les différentes courbes de Lorenz, pour chaque garantie, sont regroupées sur un même graphique afin d'observer quelle loi permet d'obtenir la meilleure estimation possible.

Pour affiner le choix lorsque les courbes sont relativement similaires, le calcul des différents indices de Gini s'avère nécessaire.

	DDE	Incendie	Vol	BDG	RC
Indice de Gini - Gamma	0.2494755	0.1118854	0.314756	0.2915579	0.2669111
Indice de Gini - Log-normale	0.2354648	0.09233819	0.2713674	0.3775043	0.2059338

TABLE 5.3 – Tableau des indices de Gini - GLM coût

Pour le coût BDG :

Le tableau des RMSE est le suivant :

	BDG	
	Gamma	Log-normale
RMSE - Apprentissage	418.2812	422.8844
RMSE - Validation	476.0823	479.8847

TABLE 5.4 – Tableau des RMSE - GLM - Coût - BDG

En se basant uniquement sur les RMSE, le choix de la loi la plus adaptée pour l'estimation du coût de cette garantie est la loi Gamma.

Afin de mettre en perspective ce constat, les courbes de Lorenz, concernant le coût BDG, sont représentées dans la figure suivante :

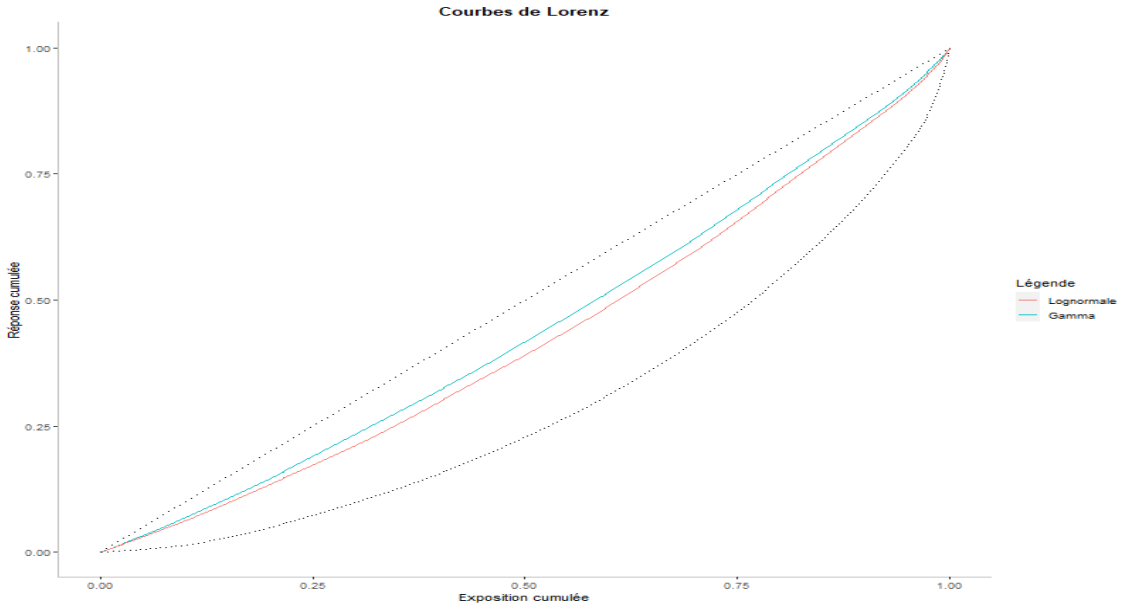

FIGURE 5.11 – Courbes de Lorenz - coût - BDG

En observant la figure 5.11 et les indices de Gini, le GLM utilisant une distribution Log-normale apparait comme le plus efficace. La courbe de Lorenz de ce modèle reste, malgré tout, toujours très proche de la courbe de mutualisation égale du coût pour tous les assurés.

Malgré ces deux résultats contradictoires, la loi Log-normale est conservée puisque l'augmentation relative de

l'indice de Gini par rapport au modèle Gamma est d'environ 30% alors que l'augmentation relative du RMSE sur la base de validation n'est que de seulement 1%.

Pour le coût DDE :

Les RMSE des modèles pour l'estimation des coûts de la garantie DDE sont réunis dans le tableau suivant :

	DDE	
	Gamma	Log-normale
RMSE - Apprentissage	1 485.519	1 497.363
RMSE - Validation	1 537.423	1 547.791

TABLE 5.5 – Tableau des RMSE - GLM - Coût - DDE

En analysant les éléments du tableau ci-dessus, la conclusion est que le modèle utilisant la loi Gamma est le plus pertinent pour ce critère.

Comme pour la garantie précédente, l'étude se poursuit en regardant les courbes de Lorenz représentées sur le graphique suivant :

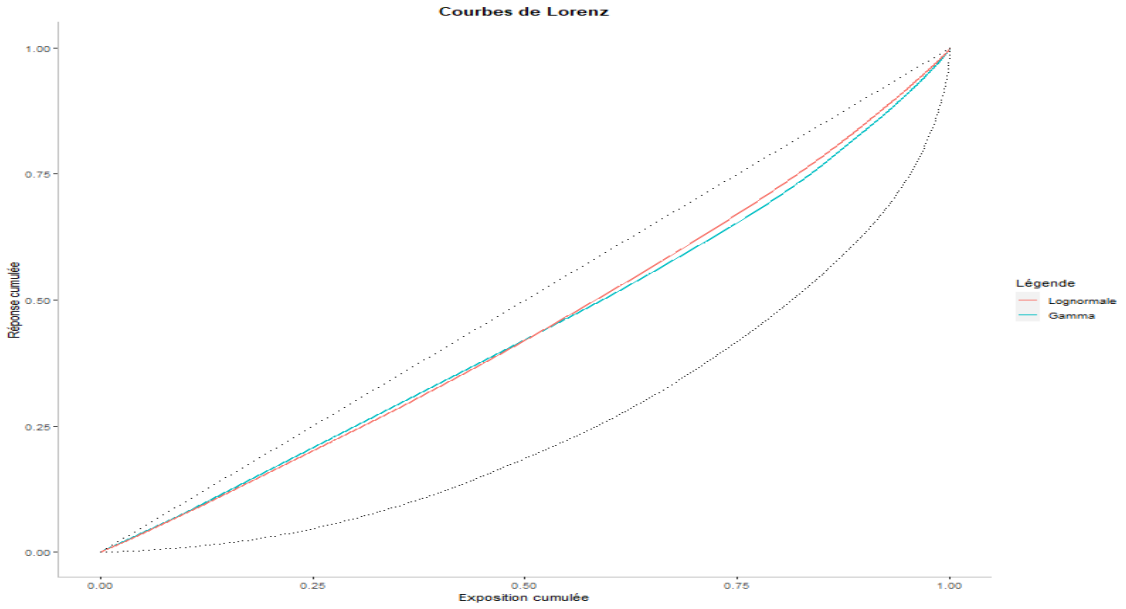

FIGURE 5.12 – Courbes de Lorenz - coût - DDE

A partir de la figure 5.12 et des indices de Gini calculés pour la garantie DDE, il semble préférable d'utiliser la distribution Gamma dans le GLM afin de prédire les coûts pour cette garantie.

Pour la garantie DDE, les deux indicateurs étudiés montrent que le modèle Gamma apparaît comme le plus efficace pour la modélisation.

Pour le coût Incendie :

Comme pour les garanties précédentes, les indicateurs de RMSE et indices de Gini (liés aux courbes de Lorenz) sont calculés.

	Incendie	
	Gamma	Log-normale
RMSE - Apprentissage	4 730.215	4 801.088
RMSE - Validation	4 508.29	4 594.252

TABLE 5.6 – Tableau des RMSE - GLM - Coût - Incendie

Ce tableau permet de consigner que le modèle Gamma semble être le mieux adapté pour cette garantie, même si la supériorité du modèle Gamma reste limitée.

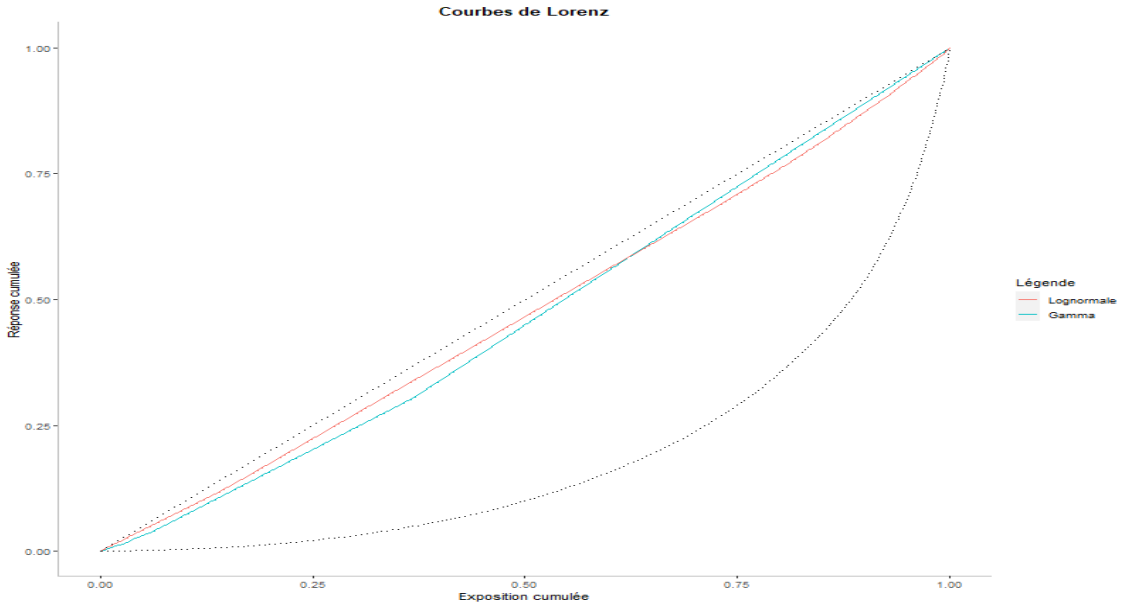

FIGURE 5.13 – Courbes de Lorenz - coût - Incendie

En se référant aux indices de Gini calculés dans le tableau 5.3, au graphique précédent permettant de les illustrer ainsi qu'aux RMSE, il est possible de déterminer lequel de ces deux modèles est celui qui capte un peu plus les variations du coût dans le portefeuille. Il s'agit de celui qui utilise une loi Gamma pour la distribution des sinistres pour cette garantie.

Pour le coût Vol :

Le tableau suivant réunit les RMSE des deux modèles réalisés pour la garantie Vol.

	Vol	
	Gamma	Log-normale
RMSE - Apprentissage	2 635.678	2 657.238
RMSE - Validation	2 402.897	2 385.457

TABLE 5.7 – Tableau des RMSE - GLM - Coût - Vol

Les RMSE confirment que le modèle Log-normal est plus adapté que le modèle GLM utilisant la loi Gamma.

Les deux courbes de Lorenz sont maintenant regroupées sur le graphique suivant :

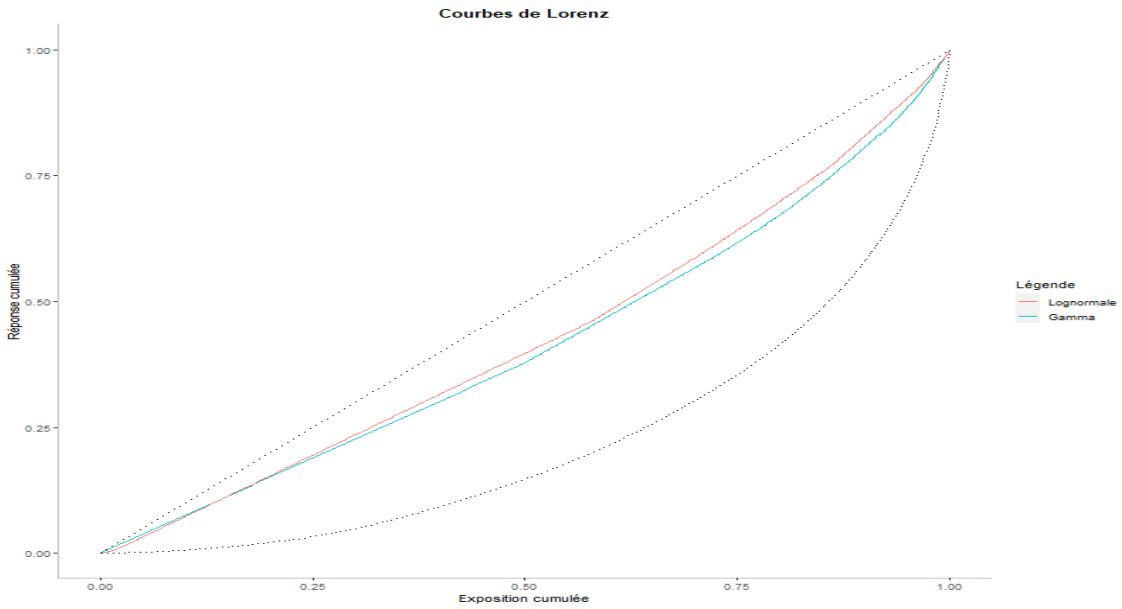

FIGURE 5.14 – Courbes de Lorenz - coût - Vol

En regardant les indices de Gini dans le tableau 5.3 pour cette garantie, il apparaît que la distribution Gamma est un peu mieux adaptée pour l'estimation.

Ainsi, les deux résultats proposent deux lois différentes pour la distribution. Or la RMSE augmente de moins de 1% pour une augmentation d'approximativement 16% de l'indice de Gini. Le choix de la loi sous-jacente se porte donc sur la loi Gamma.

Pour le coût RC :

Pour cette dernière garantie étudiée, le tableau des RMSE donne les résultats suivants :

	RC	
	Gamma	Log-normale
RMSE - Apprentissage	1 484.643	1 488.501
RMSE - Validation	1 636.005	1 628.876

TABLE 5.8 – Tableau des RMSE - GLM - Coût - RC

En regardant les RMSE sur la base de validation pour les deux modèles, le choix se porte sur le modèle avec loi Log-normale.

Les deux courbes de Lorenz sont également réunies sur le graphique ci-dessous.

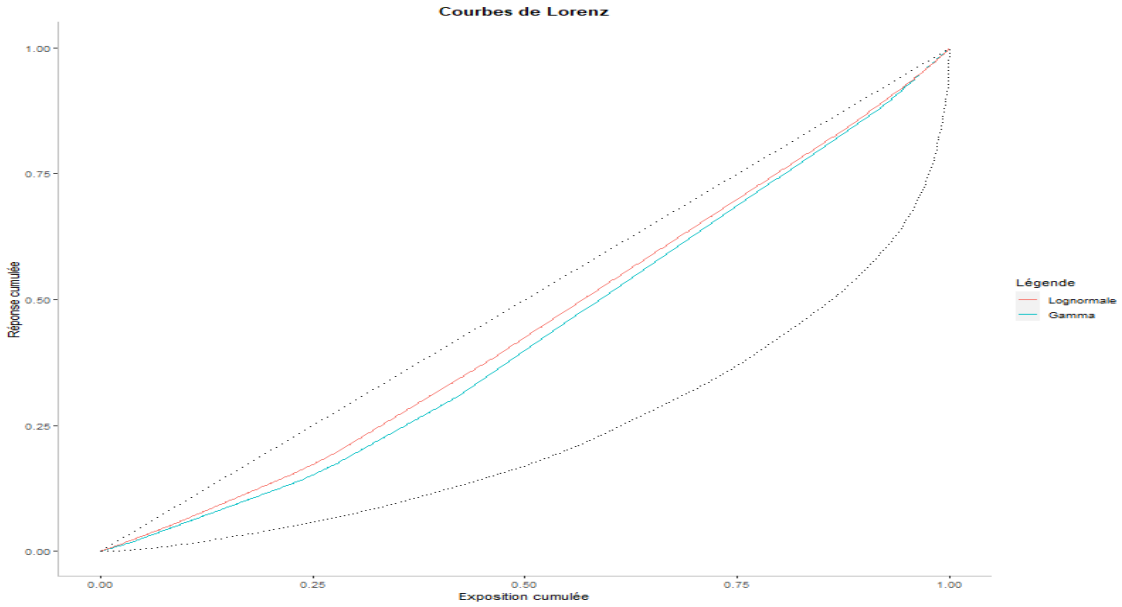

FIGURE 5.15 – Courbes de Lorenz - coût - RC

Son étude montre que les deux modélisations GLM effectuent des estimations des coûts très proches. Cependant, l'analyse des indices de Gini laisse plutôt à penser que le modèle utilisant la loi Gamma est le plus efficace des deux modèles testés.

Or la diminution de l'indice de Gini du GLM Log-normal par rapport à celui du modèle Gamma est de 23% tandis que la baisse de la RMSE ne représente que 0.4%. Par arbitrage, le choix retenu pour cette garantie est fixé sur le modèle GLM Gamma.

En conclusion, pour toutes les garanties, les RMSE restent élevées et les courbes de Lorenz proches de la courbe représentant une mutualisation du coût sur le portefeuille. Une meilleure appréciation du coût passe donc peut être par la mise en œuvre d'autres modèles.

5.1.5 Modèles de coût global

Ayant pour certaines garanties relativement peu de données et peu de modalités significatives dans les modèles, il est décidé de réaliser deux modèles "toutes garanties confondues". Le même processus que précédemment est appliqué à ces modèles.

Afin d'obtenir une distinction des coûts par rapport à la nature du sinistre, une variable illustrant le type du sinistre est créée et implémentée dans les modèles GLM. Cette variable s'avère significative pour le modèle utilisant la loi de distribution Gamma et pour celui se servant de la distribution Log-normale.

Variables	**Df**	**Statistique**	**p-value**
Zone	3	37.621	$3.400e-08$
typeHabetage	1	16.980	$3.779e-05$
typedistribution	1	4.542	0.0330652
sin_ant	1	17.015	$3.709e-05$
fran_globale	1	12.888	0.0003307
pourcentageOV	2	46.536	$7.851e-11$
typeResQualiSous	3	257.269	$< 2.2e-16$
garantie	4	2218.443	$< 2.2e-16$
montantCapitalMobilierp	2	17.154	0.0001884

TABLE 5.9 – Analyse de type III pour la loi de Gamma toutes garanties

Variables	**Df**	**Statistique**	**p-value**
nbPiecesPrincipales	3	24.104	$2.377e-05$
Zone	2	24.0465	$6.003e-06$
typedistribution	1	29.968	$4.392e-08$
sin_ant	1	108.007	$< 2.2e-16$
fran_globale	1	9.113	0.002538
pourcentageOV	2	47.328	$5.284e-11$
typeResQualiSous	2	274.841	$< 2.2e-16$
garantie	4	1967.940	$< 2.2e-16$
montantCapitalMobilierp	2	9.299	0.009568

TABLE 5.10 – Analyse de type III pour la loi Log-normale toutes garanties

Pour analyser l'efficience des modèles, les courbes de Lorenz ont d'abord été tracées à partir des données de validation toutes garanties confondues et les RMSE sont réunies dans le tableau suivant.

	Toutes Garanties	
	Gamma	Log-normale
RMSE - Apprentissage	1 937.943	1 952.16
RMSE - Validation	1 929.06	1 935.32

TABLE 5.11 – Tableau des RMSE - GLM - Coût - Global

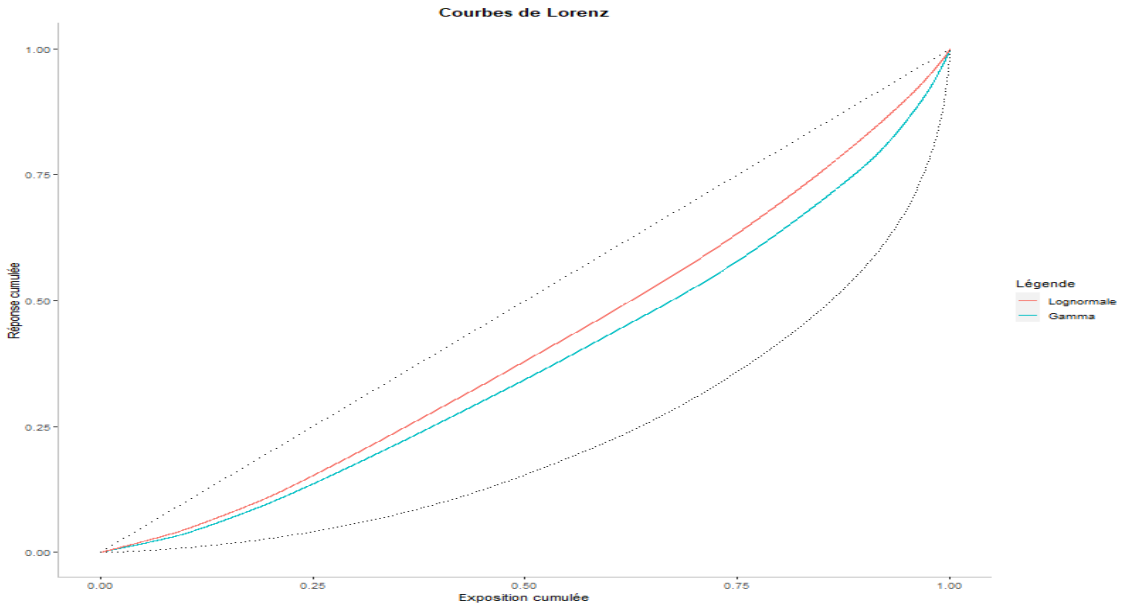

FIGURE 5.16 – Courbes de Lorenz - coût - Global

Il est possible de constater à travers le graphique précédent que le modèle global est efficace. Les deux indicateurs se rejoignent pour confirmer que le modèle "toutes garanties confondues" utilisant la loi Gamma est le plus performant.

Cependant, le but étant d'évaluer la pertinence du modèle global par rapport à chaque modèle par garantie, d'autres RMSE et courbes de Lorenz (grâce aux indices de Gini associés) ont ainsi dû être analysées.

Les tableaux suivants regroupent les différentes RMSE et indices de Gini, pour chaque garantie, afin de déterminer si le modèle global a permis de gagner en efficience par rapport à l'utilisation d'un modèle spécifique à la garantie étudiée.

	DDE	Incendie	Vol	BDG	RC
Modèle Spécifique - Gamma	1 537.423	4 508.29	2 402.897	476.0823	1 636.005
Modèle Global - Gamma	1 544.154	4 690.938	2 407.435	481.903	1 627.014
Modèle Spécifique - Log-normale	1 547.791	4 594.252	2 385.457	479.8847	1 628.876
Modèle Global - Log-normale	1 548.291	4 709.741	2 417.439	483.581	1 629.254

TABLE 5.12 – Tableau de comparaison des RMSE base de validation (modèle global/spécifique)

	DDE	Incendie	Vol	BDG	RC
Modèle spécifique - Gamma	0.2494755	0.1118854	0.314756	0.2915579	0.2669111
Modèle Global - Gamma	0.2017297	0.2059591	0.2038371	0.212794	0.1876072
Modèle spécifique - Log-normale	0.2354648	0.09233819	0.2713674	0.3775043	0.2059338
Modèle Global - Log-normale	0.1857381	0.1746979	0.1732983	0.1895579	0.1681829

TABLE 5.13 – Tableau de comparaison des indices de Gini (modèle global/spécifique)

Au travers des résultats obtenus, il est possible de conclure qu'à l'exception de la garantie Incendie (pour les deux lois) et la garantie RC (pour la loi Gamma), le modèle global ne permet pas d'améliorer l'efficacité de l'estimation du coût des sinistres.

Dans le cas de la garantie Incendie, puisque l'indice de Gini est significativement plus élevé par l'utilisation du modèle global (respectivement augmentation relative d'environ 84% (Gamma) et 89% (Log-normale) pour une augmentation du RMSE de 4% et de 2.5%), si un choix devait être effectué entre l'utilisation du GLM spécifique à la garantie et celui global, à ce moment de l'étude, il se porterai plutôt sur le modèle "toutes garanties confondues". Cette évolution significative pourrait venir de deux faits : le premier étant que les sinistres présents dans la base sinistre Incendie sont insuffisants pour capter tous les effets des variables explicatives sur le coût et le deuxième étant que, usuellement, les garanties DDE et Incendie ont des impacts similaires au niveau des variables tarifantes. Ainsi, au lieu de créer du bruit, comme pour les autres garanties, la discrimination plus accentuée du modèle général (essentiellement due à la garantie la plus présente, DDE) permet d'améliorer la compréhension des effets des variables explicatives sur le coût incendie.

Dans la situation de la garantie RC, pour la distribution Gamma - modèle global, il est constaté une diminution relative de la RMSE de moins de 1%. Cependant, la diminution relative de l'indice de Gini est d'environ 30%. Le modèle "toutes garanties confondues" ne sera donc pas considéré pour modéliser la sévérité de la garantie RC.

Par la réalisation de ce modèle global, il apparaît possible de trouver une solution permettant d'améliorer les performances du modèle pour une des garanties. Pour poursuivre cette démarche de recherche de performance, d'autres modèles vont être testés par garantie et "toutes garanties confondues".

Avant de poursuivre les analyses, les choix des modèles les plus efficaces en fonction de chaque garantie sont regroupés dans le tableau récapitulatif suivant :

	DDE	Incendie	Vol	BDG	RC
Modèle Sélectionné	Gamma	Gamma (toutes garanties)	Gamma	Log-normale	Gamma

TABLE 5.14 – Tableau des GLM les mieux adaptés, par garantie

5.2 Arbre CART

Pour essayer d'améliorer l'évaluation du coût, comme dans le cas de la fréquence, une mise en œuvre des modèles CART est expérimentée. Dans le cas de la modélisation fréquence, les variables et les regroupements réalisés lors de la création des modèles GLM (ceux qui étaient les plus pertinents par garantie) ont été utilisés, afin de servir de base pour la réalisation des modèles de *Data Science*. Cette approche a été choisie pour avoir une harmonisation des regroupements en vue de la comparaison entre les modèles et également car une majorité des variables explicatives étaient conservées dans les modèles par garantie. Cette même méthodologie est mise en œuvre comme première approche, pour les modèles suivants. Cependant, suite à l'application des GLM pour l'estimation des coûts, très peu de variables ont été conservées et beaucoup de regroupements ont été effectués. De plus, les bases pour certaines

garanties n'ayant que peu d'observations, le choix est fait de conserver également les variables n'étant pas utilisées dans les GLM, dans le cadre d'une seconde approche. La réalisation de cette dernière permettra ainsi d'observer si l'utilisation de variables supplémentaires améliorera les modèles qui seront réalisés ensuite et donc si ces nouveaux modèles réussiront à capter de manière plus significative les effets des variables écartées dans les modèles GLM pour l'estimation des différents coûts.

Les modèles utilisant l'ensemble des sinistres, toutes garanties confondues, sont aussi étudiés.

Après avoir obtenu les arbres maximaux, leur élagage est effectué en retenant les paramètres de complexité suivants :

	DDE	Incendie	Vol	BDG	RC	Global
cp - Approche 1	1.804153e-03	0.0002677001	5.02888e-05	0.003605905	5.549728e-06	0.002815501
cp - Approche 2	1.804153e-03	0.00241364	0.007684858	0.006127646	0.00175931	0.002815501

TABLE 5.15 – Tableau regroupant les différentes cp utilisées pour l'élagage - CART

Les arbres obtenus lors des différentes modélisations sont consignés en annexe O.

Pour vérifier la pertinence de l'utilisation du modèle CART sur ces garanties par rapport à l'utilisation du GLM, une comparaison des RMSE et des indices de Gini, entre le modèle GLM désigné comme le plus performant et le modèle CART, est présentée dans les tableaux suivants :

	DDE	Incendie	Vol	BDG	RC
RMSE - Approche 1	1 534.523	4 509.869	2 423.11	477.824	1 634.765
RMSE - Approche 2	1 516.723	4 534.496	2 402.455	477.7486	1 632.9
RMSE - GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.16 – Tableau des RMSE base de validation - CART coût

	DDE	Incendie	Vol	BDG	RC
Gini - Approche 1	0.1736876	0.08684865	0.313368	0.2563355	0.2663519
Gini - Approche 2	0.1708604	0.1216667	0.2063629	0.2324474	0.2182258
Gini - GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.17 – Tableau des indices de Gini - CART coût

Pour toutes les garanties et pour les deux approches, les indices de Gini sont inférieurs à ceux obtenus par les modèles GLM. Cependant, à part pour la première approche de la garantie Vol, l'ensemble des RMSE obtenues par CART sont inférieures aux RMSE des GLM retenus. Ainsi, pour chaque garantie, un arbitrage doit être réalisé entre ces deux indicateurs pour conclure sur l'amélioration ou non de la performance par les modèles d'arbres de régression.

Dans les cas de la garantie DDE, de la garantie BDG (pour les deux approches) et de l'approche 2 de la garantie Vol, une diminution relative de la RMSE de moins de 1% par rapport celle du GLM est constatée alors que la perte relative pour l'indice de Gini est entre 30% et 39% (pour les deux approches). Ainsi, l'utilisation des modèles CART ne conduit pas à une amélioration de l'estimation du coût, pour les garanties concernées.

Pour la garantie Incendie, la diminution relative des RMSE est plus importante (4% pour la première approche et 3.5% pour la deuxième), pourtant le modèle GLM reste toujours le plus performant puisque les diminutions pour les indices de Gini sont toujours très significatives (58% et 41% respectivement).

Enfin, en ce qui concerne la deuxième approche de la garantie RC, le même constat peut être fait que pour les précédentes garanties (diminution de 0.2% de la RMSE pour une baisse de 18% de l'indice de Gini). Dans le cas de la première approche, la réduction de la RMSE est de 0.08% pour une décrue de 0.20% de l'indice de Gini. La dernière diminution étant légèrement supérieure à celle de la RMSE, l'arbitrage est fait de considérer le modèle GLM comme le plus performant.

Pour "toutes garanties confondues" :

Concernant le modèle global, les deux approches ont abouti à la réalisation du même arbre de régression.

Le but de ce paragraphe étant l'étude de la pertinence de ce modèle par garantie en comparaison à celle des GLM, une analyse détaillée est effectuée et présentée dans les tableaux suivants :

	DDE	Incendie	Vol	BDG	RC
Modèle global - CART	1 548.503	4 530.407	2 417.811	503.7382	1 635.777
Modèle GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.18 – Tableau de comparaison des RMSE base de validation (modèle global CART/GLM)

	DDE	Incendie	Vol	BDG	RC
Modèle global - CART	0.1896539	0.0000	0.2138844	0.07471641	0.07092169
Modèle GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.19 – Tableau de comparaison des indices de Gini (modèle global CART/GLM)

Comme dans le cas des modèles CART par garantie, le modèle "toutes garanties confondues" donne des indices de Gini inférieurs à ceux des GLM sur toutes les études réalisées. Les RMSE des garanties DDE, Vol et BDG sont supérieures à celles des GLM tandis que celles des garanties Incendie et RC sont inférieures. Pour les trois premières garanties citées, il est possible de conclure, sans analyse supplémentaire, que les modèles GLM sont plus efficaces dans l'estimation des coûts. Dans le cas de la garantie Incendie, une perte de RMSE de 3.42% est constatée par rapport à celle du GLM, pour une chute de l'indice de Gini représentant 100%. Le modèle CART "toutes garantie confondues" n'est donc pas privilégié pour déterminer le coût de la garantie Incendie. Enfin, dans le cas de la garantie RC, la perte de RMSE est largement inférieure à 1% (approximativement 0.014%) contre une diminution de l'indice de Gini d'environ 73%. Ainsi, comme pour la garantie Incendie, l'arbitrage entre les deux indicateurs permet de déduire que le modèle GLM reste le plus pertinent.

En conclusion de cette partie, les calculs démontrent que la modélisation CART, dans le cadre de cette étude, ne permet pas d'améliorer (presque équivalence dans la première approche de la garantie RC) les résultats obtenus par la modélisation GLM.

En raison de l'instabilité des arbres CART qui est l'un des inconvénients de ce type de modélisation, les parties suivantes sont consacrées à la mise en œuvre de modèles de type Random Forest et XGBoost afin de voir si les performances obtenues seront supérieures à celles des CART mais aussi des GLM.

5.3 Les Forêts aléatoires

Dans cette partie, la même méthodologie que celle utilisée pour la fréquence est appliquée. Pour rappel, elle est développée dans la partie 4.3. Comme pour les modèles CART, les deux approches sont réalisées.

En appliquant la méthode *Grid Search* sur les différents modèles coût, les paramètres suivants sont selectionnés sur l'ensemble des modèles testés :

	DDE		Incendie		Vol		BDG		RC		Global	
	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2
Nombre d'arbres	140	120	80	600	380	180	180	700	160	40	100	380
Nombres de variables	3	5	3	3	3	3	7	3	3	3	3	10
Taille de l'arbre	20	20	20	20	20	20	100	20	20	100	100	20

TABLE 5.20 – Récapitulatif des paramètres optimaux - Random Forest - Coût

Une représentation graphique de l'évolution de la RMSE, en fonction du nombre d'arbres pour la forêt optimale pour chaque garantie et pour les deux approches, est consignée en annexe P. Pour l'ensemble des forêts optimales, il est constaté qu'une stabilisation s'effectue avant le nombre d'arbres optimal trouvée précédemment. Ainsi, le choix fait du nombre d'arbres assure la robustesse du modèle.

Dans cette même annexe sont aussi présentés les graphiques affichant l'importance des variables. Comme pour la modélisation de la fréquence Incendie par *Random Forest*, certains modèles de coût créés ont eu des variables avec des importances négatives. La même méthode a alors été utilisée dans ces cas-là, en enlevant la variable ayant une importance négative. Afin de simplifier la lecture des graphiques dans l'annexe, seuls les graphiques d'évolution de la RMSE et d'importance des variables, sans les variables à importance négative, sont reproduits, pour chaque garantie.

Dans le tableau suivant sont résumées les trois variables les plus importantes pour chacune des garanties.

		Variable1	Variable 2	Variable 3
DDE	A1	typeResQualiSous	sin_ant	fran_globale
	A2	typeResQualiSous	sin_ant	typeHabetage
Incendie	A1	fran_globale	typedistribution	typeResQualiSous
	A2	Zone	fran_globale	typedistribution
Vol	A1	pourcentageOV	nbPiecesPrincipales	typeResQualiSous
	A2	pourcentageOV	nbPiecesPrincipales	typeHabetage
BDG	A1	fran_globale	Zone	sin_ant
	A2	typeResQualiSous	typedistribution	fran_globale
RC	A1	fran_globale	typeResQualiSous	Zone
	A2	fran_globale	typeResQualiSous	typeHabetage
Global	A1	garantie	typeResQualiSous	typeHabetage
	A2	garantie	typeResQualiSous	fran_globale

TABLE 5.21 – Variables les plus importantes - RF - Coût

Enfin, la pertinence des modèles *Random Forest* par rapport aux modèles GLM est étudiée dans les deux tableaux suivants.

	DDE	Incendie	Vol	BDG	RC
RMSE - Approche 1	1 537.729	4 510.345	2 423.454	483.993	1 634.818
RMSE - Approche 2	1 542.535	4 510.48	2 395.905	477.3401	1 632.622
RMSE - GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.22 – Tableau des RMSE base de validation - RF coût

	DDE	Incendie	Vol	BDG	RC
Gini - Approche 1	0.2081555	0.08309836	0.3096154	0.3244768	0.2681032
Gini - Approche 2	0.2246438	0.1580726	0.2873589	0.2059568	0.3389661
Gini - GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.23 – Tableau des indices de Gini - RF coût

Sans analyse plus approfondie des résultats obtenus, il est possible de constater que pour les premières approches pour les garanties BDG et Vol et pour les deux approches de la garantie DDE, les RMSE des *Random Forest* augmentent tandis que les indices de Gini diminuent. Pour ces trois cas, les modèles sont donc moins performants que les GLM. En ce qui concerne la garantie RC, les deux modèles sont considérés plus efficaces dans l'estimation du coût que le GLM puisque les deux indicateurs montrent une amélioration. Or, l'approche 2 de cette garantie est caractérisée par une diminution de la RMSE (moins de 1%) et une augmentation de l'indice de Gini (27%) plus significative que dans le cas de l'approche 1. Ainsi, pour la garantie RC, le modèle *Random Forest* utilisant l'ensemble des variables explicatives est le plus performant.

Pour les deuxièmes approches des garanties BDG et Vol, la diminution de la RMSE est de moins de 0.50% alors que la chute constatée au niveau des indices de Gini est respectivement de 45% et 9%. En réalisant un arbitrage entre ces deux indicateurs, conclusion est faite que ces modèles restent moins efficaces que les modèles GLM.

Pour la garantie Incendie, la diminution significative des indices de Gini (Approche 1, 60% et Approche 2, 23%) est trop importante pour justifier l'utilisation de l'un de ces modèles à la place du GLM, en particulier en prenant en compte la chute de la RMSE qui est, dans les deux approches, d'environ 3.85%.

Pour "toutes garanties confondues" :

Les mêmes étapes que pour les modèles par garantie spécifique sont réalisées dans les deux approches pour le modèle global "toutes garanties confondues".

	DDE	Incendie	Vol	BDG	RC
Modèle Global - RF Approche 1	1 546.163	4 564.004	2 572.604	493.758	1 627.813
Modèle Global - RF Approche 2	1 543.662	4 598.818	2 597.546	477.4478	1 628.226
Modèle GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.24 – Tableau de comparaison des RMSE base de validation (modèle global RF/GLM)

	DDE	Incendie	Vol	BDG	RC
Modèle Global - RF Approche 1	0.2156593	0.1496013	0.1473126	0.1714064	0.1583464
Modèle Global - RF Approche 2	0.1973317	0.1123101	0.1436079	0.1472163	0.1793928
Modèle GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.25 – Tableau de comparaison des indices de Gini (modèle global RF/GLM)

Les RMSE augmentent et les indices de Gini diminuent (par rapport aux modèles GLM) pour les garanties DDE, Vol et la première approche de la garantie BDG. Les modèles GLM de ces premières garanties restent donc les plus performants.

Pour la deuxième approche de la garantie BDG, la baisse relative de 0.51% de la RMSE n'est pas assez significative pour compenser la diminution d'approximativement 61% de l'indice de Gini. Ainsi, le modèle GLM BDG reste aussi le plus efficace pour caractériser le coût de cette garantie.

En ce qui concerne la garantie Incendie, pour les deux approches, il est constaté respectivement une baisse de la RMSE de 2.71% et de 1.96% pour une chute de l'indice de Gini de 27% et de 45% environ. Donc, aucune de ces deux nouvelles modélisations ne permettent d'améliorer les résultats obtenus à partir des modèles GLM.

Le premier constat pour la garantie RC est que la RMSE subit une baisse plus importante que dans le cas de la modélisation *Random Forest* spécifique à la garantie. Mais les indices de Gini avec le modèle global "toutes garanties confondues" s'affaiblissent par rapport au modèle spécifique.

En comparant dans un premier temps le modèle "toutes garanties confondues" avec le modèle GLM, il y a une chute de 0.5% de la RMSE avec les deux approches alors que la baisse pour les indices de Gini est beaucoup plus forte (Approche 1 : environ 41% et Approche 2 : environ 33%). Ces modélisations ne sont donc pas considérées comme plus efficaces que celles du GLM. La comparaison avec le modèle *Random Forest* spécifique à la garantie est réalisée à titre indicatif. Pour les modèles "toutes garanties confondues", une amélioration de la RMSE d'environ 0.3% est obtenue pour une dégradation de 53% ou de 47% de l'indice de Gini selon les approches. Ceci permet de conclure que le modèle choisi, paramétré avec uniquement les montants de sinistre de la garantie, reste le plus performant.

Pour conclure, l'utilisation de *Random Forest* a permis d'améliorer la modélisation du coût pour la garantie RC. Pour les autres garanties, une amélioration de la RMSE est parfois constatée mais cela ne permet pas de compenser les dégradations de l'indice de Gini. Une autre approche de *Data Science* utilisant aussi les arbres va être mise en œuvre dans la suite toujours dans le but de tenter de surpasser les modèles GLM.

5.4 eXtreme Gradient Boosting : XGBoost

La méthodologie mise en place pour la réalisation des différents *XGBoost* pour l'estimation du coût est la même que pour les *XGBoost* pour l'estimation de la fréquence (cf. partie 4.4). Elle reprend les approches, A1 et A2, développées dans les parties précédentes de ce chapitre 5.

L'application de la méthode *Grid Search* pour les différentes garanties et approches donne les paramètres suivants :

	DDE		Incendie		Vol		BDG		RC		Global	
	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2	A1	A2
Taux d'apprentissage	0.1	0.1	0.1	0.1	0.05	0.05	0.1	0.05	0.05	0.05	0.05	0.05
Profondeur de l'arbre	30	30	10	30	30	50	10	30	10	30	50	50
Pourcentage de variables	1	1	0.3	1	1	1	1	1	1	1	1	1

TABLE 5.26 – Récapitulatif des paramètres optimaux - XGBoost - Coût

Concernant les modèles optimaux créés à partir des paramètres précédents, l'importance des variables explicatives est étudiée. Les graphiques résumant les différents effets sont présentés dans l'annexe Q. Le tableau suivant permet de visualiser les trois variables du modèle optimal, pour chaque garantie et pour chacune des deux approches :

		Variable1	Variable 2	Variable 3
DDE	A1	Locataire - Principale	Appartement-Chambre	Zone 9-5-6-7-8
	A2	Locataire - Principale	Zone 9-5-6-7-8	Appartement-Chambre
Incendie	A1	fran_globale 200-300-400	Locataire-Propriétaire	typedistribution RCS-Agence-Internet
	A2	Appartement-Chambre	typedistribution RCS-Agence-Internet	fran_globale 200-300-400
Vol	A1	pourcentageOV 0%	nbPiecesPrincipales 3-1-2	Locataire - Principale
	A2	pourcentageOV 0%	Appartement-Chambre	Locataire - Principale
BDG	A1	fran_globale 200-300-400	Locataire	Zone 7-8-9
	A2	fran_globale 200-300-400	Appartement-Chambre	Zone 7-8-9
RC	A1	fran_globale 200-0-75-150	Propriétaire	Zone 9-7-8-10-1-2
	A2	fran_globale 200-0-75-150	Appartement-Chambre	typedistribution RCS-Agence
Global	A1	garantie Incendie	garantie BDG	Locataire - Principale
	A2	garantie Incendie	garantie BDG	Locataire - Principale

TABLE 5.27 – Variables les plus importantes - XGBoost - Coût

Pour finaliser cette partie de l'étude, la pertinence des modèles pour chaque approche est comparée à celle obtenue pour les modèles GLM. Les résultats sont consignés dans les deux tableaux suivants :

	DDE	Incendie	Vol	BDG	RC
RMSE - Approche 1	1 556.399	4 507.757	2 407.209	483.456	1 634.336
RMSE - Approche 2	1 550.5	4 590.634	2 436.144	487.6812	1 647.323
RMSE - GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.28 – Tableau des RMSE base de validation - XGBoost coût

	DDE	Incendie	Vol	BDG	RC
Gini - Approche 1	0.2607366	0.1117732	0.2832557	0.3210083	0.2657947
Gini - Approche 2	0.2536284	0.2958713	0.3392619	0.370297	0.4289614
Gini - GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.29 – Tableau des indices de Gini - XGBoost coût

Pour les deux approches de la garantie DDE, la RMSE et l'indice de Gini augmentent. Le gain fait au niveau de l'indice de Gini est de 4.51% (approche 1) et 1.66% (approche 2) par rapport au GLM. Cette hausse significative n'est pas contrecarrée par l'augmentation des RMSE qui sont d'environ 1%. Par arbitrage, les deux modèles peuvent être vus comme plus performants que le modèle GLM choisi. Si une comparaison est effectuée de l'approche 2 par l'approche 1, il est possible de constater une baisse relative de 0.38% de la RMSE pour une chute de 2.73% de l'indice de Gini. Ainsi, l'approche la plus performante des deux sera l'approche 1. Cette approche n'apporte pas d'amélioration suffisamment significative de l'indice de Gini pour pallier les problèmes d'interprétabilité et flexibilité des modèles XGBoost par rapport aux modèles GLM, choix est fait de conserver le modèle GLM pour modéliser le coût.

En ce qui concerne les deux approches de la garantie BDG et la première approche de la garantie Vol, ces méthodes ne permettent pas d'égaler ou d'améliorer l'estimation par rapport aux GLM puisque les deux indicateurs se détériorent. Pour la deuxième approche de la garantie Vol, la hausse de la RMSE est d'un peu plus de 1% pour une majoration de 7.79% de l'indice de Gini. Cette hausse de RMSE moins importante permet de déduire que le modèle *XGBoost* de la deuxième approche est plus efficace dans l'estimation des coûts que les modèles GLM. Or comme dans le cas de la garantie DDE, la plus-value obtenue au niveau de l'indice de Gini n'est pas retenue comme suffisante par rapport à la perte dans la facilité d'explication du modèle et dans la contribution des variables explicatives dans l'estimation. Le modèle GLM est donc conservé.

Pour la garantie Incendie, la première approche n'améliore pas la performance du modèle GLM puisque la diminution relative de 3.9% de la RMSE n'est pas suffisante pour contrer la chute de 46% de l'indice de Gini. Par contre, la deuxième approche laisse apparaître une chute de 2.14% de la RMSE avec une hausse de 44% de l'indice de Gini. Comme une amélioration des deux indicateurs est constatée par rapport aux résultats obtenus avec les GLM et que celle de l'indice de Gini est significative, déduction est faite que cette approche est la plus performante.

Pour la garantie RC, la première approche est similaire aux modèles GLM puisque les deux indicateurs varient de moins de 1%. La seconde approche permet une augmentation de l'indice de Gini de 61% pour une dégradation de seulement 0.69% de la RMSE par rapport au modèle GLM. En prenant en compte cette évolution flagrante de l'indice de Gini, la deuxième approche pour cette garantie est plus pertinente que la modélisation par GLM.

Pour "toutes garanties confondues" :

Les mêmes étapes que pour les modèles par garantie spécifique sont réalisées pour les deux approches, pour le modèle global "toutes garanties confondues".

	DDE	Incendie	Vol	BDG	RC
Modèle Global - XGBoost Approche 1	1 555.304	4 658.478	2 466.951	485.9179	1 669.346
Modèle Global - XGBoost Approche 2	1 554.252	4 719.267	2 516.372	490.8996	1 656.167
Modèle GLM	1 537.423	4 690.938	2 402.897	479.8847	1 636.005

TABLE 5.30 – Tableau de comparaison des RMSE base de validation (modèle global XGBoost/GLM)

	DDE	Incendie	Vol	BDG	RC
Modèle Global - XGBoost Approche 1	0.254827	0.2859541	0.3621171	0.3159959	0.4196129
Modèle Global - XGBoost Approche 2	0.2665662	0.3409151	0.3769019	0.3535124	0.4629331
Modèle GLM	0.2494755	0.2059591	0.314756	0.3775043	0.2669111

TABLE 5.31 – Tableau de comparaison des indices de Gini (modèle global XGBoost/GLM)

Pour la garantie BDG, l'augmentation des RMSE ainsi que la chute des indices de Gini, par rapport aux indicateurs obtenus par GLM, permettent de conclure que ces modèles *XGBoost*, quelle que soit l'approche, ne sont pas assez performants pour challenger le modèle GLM.

Les approches de la garantie DDE montrent une augmentation d'environ 1% des RMSE, dans les deux cas, pour une augmentation de l'indice de Gini de respectivement 2.14% et 6.85% (approche 1 et approche 2). Il est ainsi possible de déduire que ces deux modèles sont plus efficaces que le modèle GLM. Comme les pourcentages d'augmentation de la RMSE sont équivalents à ceux obtenus par l'approche 1 du modèle *XGBoost* spécifique à la garantie et que l'augmentation des indices de Gini est plus significative dans le cas de la deuxième approche (modèle Global), le meilleur modèle est donc obtenu grâce à l'apprentissage "toutes garanties confondues" et utilisant toutes les variables (A2). Néanmoins, comme pour les modèles spécifiques à la garantie, l'évolution de l'indice est insuffisante pour retenir ce modèle pour le coût.

Pour la garantie Incendie, les variations des RMSE sont de moins de 1% et les hausses des indices de Gini sont respectivement de 38.84% et 65.53% (approche 1 et approche 2). L'approche 2 "toutes garanties confondues" est donc la plus efficace pour la modélisation recherchée. Une comparaison par rapport à l'approche 2 du modèle XGBoost spécifique montre une augmentation de 2.80% de la RMSE et une amélioration de l'indice de Gini de 15%. Ainsi, l'utilisation du modèle XGBoost "toutes garanties confondues", utilisant toutes les variables explicatives, est préférable pour la modélisation du coût de cette garantie.

Dans le cas de la garantie Vol, la hausse des RMSE est de 2.67% pour l'approche 1 et de 4.72% pour l'approche 2. L'augmentation des indices de Gini est de 15.05% et 19.74% (approche 1 et approche 2). L'approche 2 est préférée à l'approche 1 et au GLM puisque la différence relative de RMSE entre l'approche 1 et l'approche 2 est de 2% tandis que celle de l'indice de Gini est de 4.08%. Le meilleur modèle "toutes garanties confondues" pour estimer les coûts Vol est donc l'approche 2. Si une comparaison par rapport au meilleur modèle *XGBoost* est réalisée avec uniquement les sinistres Vol, il est constaté une baisse de la RMSE de 3.29% et de l'indice de Gini de 11.09%. Ainsi le modèle le plus performant pour l'estimation du coût est le *XGBoost* global utilisant l'ensemble des variables disponibles.

En ce qui concerne la garantie RC, les approches du modèle "toutes garanties confondues" ont une variation relative de la RMSE par rapport au modèle GLM d'environ $1-2\%$ pour une hausse relative de l'indice de Gini toujours par rapport au GLM de respectivement 57.21% et 73.44% (approche 1 et 2). Ainsi, l'approche 2 est donc la plus performante. Pour comparer ce modèle par rapport au modèle *XGBoost* spécifique à la garantie qui a été défini précédemment comme plus efficace que le modèle GLM, il est constaté une hausse de la RMSE de moins de 1% pour une hausse de 7.92% de l'indice de Gini. Ainsi l'approche 2 "toutes garanties confondues" est celle qui donne la meilleure estimation pour le coût RC.

Dans le cas des garanties DDE et BDG, les modélisations les plus pertinentes sont respectivement le GLM Gamma et le GLM Log-normal. La modélisation XGBoost "toutes garantie confondues" et utilisant toutes les variables explicatives est déterminée comme étant la plus efficace pour l'estimation des coûts des garanties Incendie et Vol. Enfin, en ce qui concerne la garantie RC, deux types de modèles sont plus efficaces que le GLM, l'approche 2 du modèle *Random Forest* spécifique et l'approche 2 du modèle *XGBoost* "toutes garanties confondues". La variation relative de la RMSE de l'approche 2 de la méthode *Random Forest* est de -1.42% par rapport à l'approche 2 de la méthode *XGBoost* "toutes garanties confondues", tandis que la différence relative de l'indice de Gini est de -27%. Ainsi, la méthode *Random Forest* est moins efficace pour l'estimation du coût RC.

En conclusion de ce chapitre 5, les modélisations les plus efficaces pour le coût ont été résumées dans le tableau suivant :

	Modèle coût choisi
DDE	GLM - Gamma
Incendie	XGBoost Toutes garanties - Toutes variables
Vol	XGBoost Toutes garanties - Toutes variables
BDG	GLM - Log-normal
RC	XGBoost Toutes garanties - Toutes variables

Chapitre 6

Comparaison avec le tarif actuel

Dans ce dernier chapitre de l'étude, les tarifs obtenus grâce aux modélisations de fréquence et de coût les plus performantes pour chaque garantie vont être comparés aux tarifs actuellement mis en place au service des partenariats de l'ÉQUITÉ. Pour rappel, à la prime pure calculée par les modélisations viendra s'ajouter une prime pure pour les orphelins. Pour finir, 30% de la prime finale a été dédié à la prise en compte des sinistres graves, des frais et de la réassurance.

Pour effectuer cette comparaison par garantie, pour chaque contrat de la base de validation, le nouveau tarif est calculé en multipliant la fréquence modélisée par le coût estimé, obtenus par les modèles déterminés dans les chapitres précédents. L'écart relatif entre ce nouveau tarif et le tarif actuellement utilisé est ensuite obtenu par l'application de la formule suivante :

$$ecart = \frac{tarif\ modelise - tarif\ actuel}{tarif\ actuel} \times 100$$

L'histogramme ci-dessous montre la répartition des écarts pour toutes les garanties confondues. Des écarts significatifs sont constatés. En effet, environ 30% du portefeuille d'étude a un écart avec le tarif actuel de plus de 20% et 30% des tarifs ont une différence de moins de -20%.

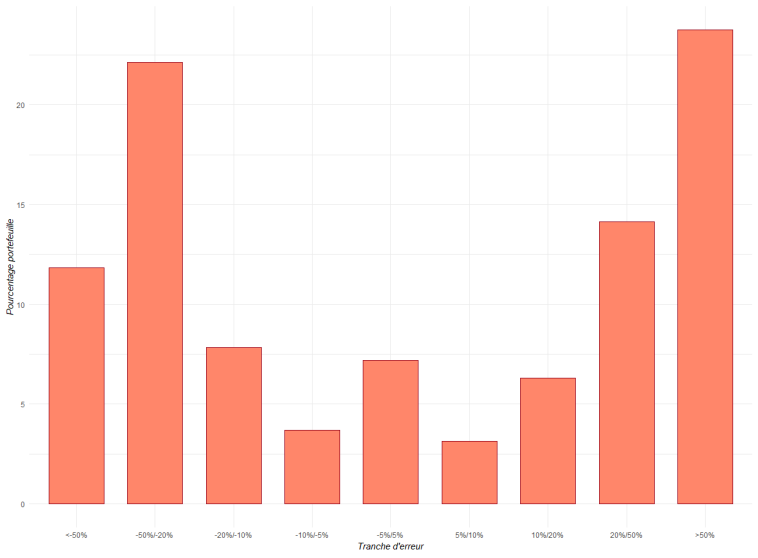

FIGURE 6.1 – Comparaison tarif estimé/tarif actuel sur l'ensemble des garanties

Afin de pouvoir étudier les effets de la tarification, il est pertinent de poursuivre l'étude en regardant la répartition des écarts pour chaque garantie.

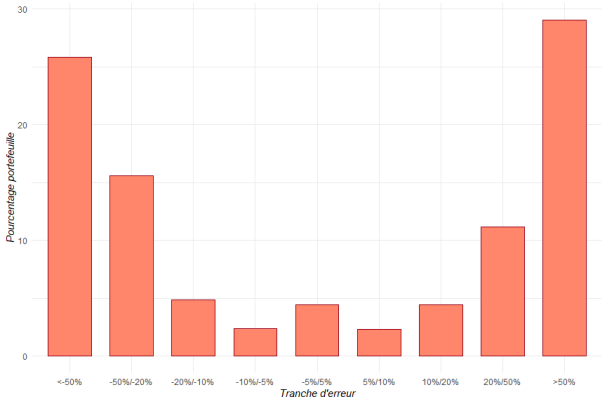

FIGURE 6.2 – Comparaison tarif estimé/tarif actuel pour la garantie BDG

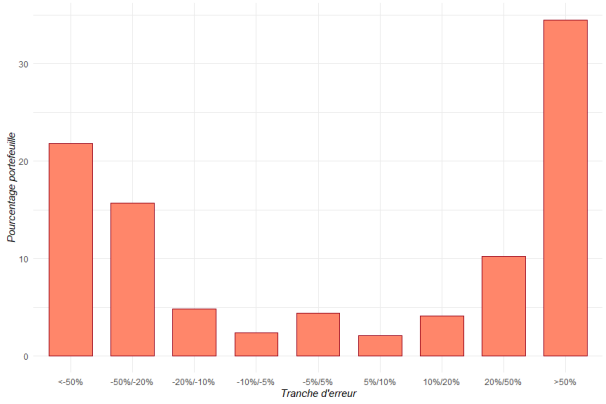

FIGURE 6.3 – Comparaison tarif estimé/tarif actuel pour la garantie DDE

Pour les garanties BDG et DDE, une sous-estimation et une surestimation par rapport aux tarifs actuels peuvent être constatées. Ces constats sont un rappel des observations effectuées sur le graphique de comparaison pour toutes garanties confondues.

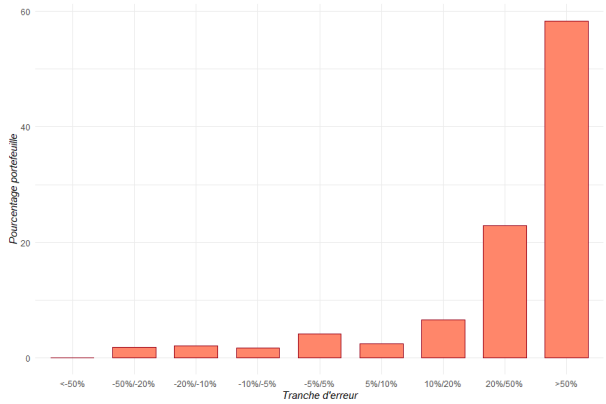

FIGURE 6.4 – Comparaison tarif estimé/tarif actuel pour la garantie Incendie

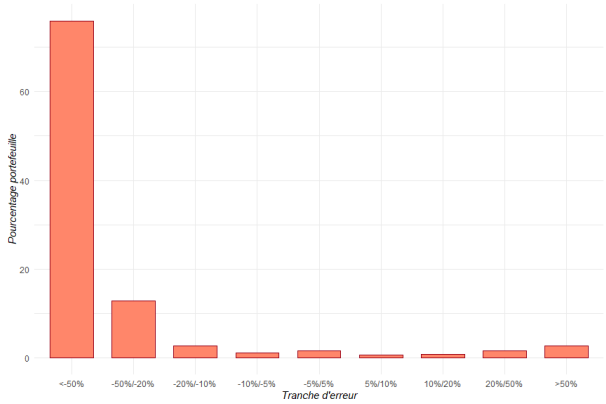

FIGURE 6.5 – Comparaison tarif estimé/tarif actuel pour la garantie Vol

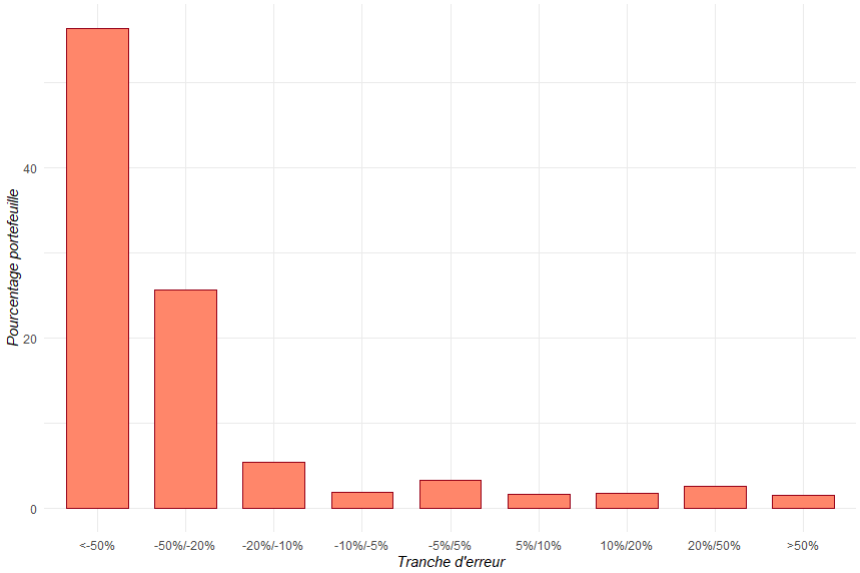

FIGURE 6.6 – Comparaison tarif estimé/tarif actuel pour la garantie RC

En ce qui concerne la garantie Incendie, le graphique permet de conclure a une estimation plus élevée de la prime sur l'ensemble du portefeuille étudié (cf figure 6.4). Pour les garanties Vol et RC, c'est le constat inverse, une diminution de tarif est constatée par rapport au tarif fourni actuellement (cf figures 6.5 et 6.6).

La dernière étape dans ces comparaisons est d'identifier les caractéristiques des unités présentant les décalages les plus significatifs. Pour identifier ces caractéristiques, les effets de certaines des variables sont examinés (principalement type Habitation, qualité souscripteur, nombre de pièces principales et franchise). Ceci permet d'avoir une première appréciation sur les différences avec le tarif actuellement en vigueur, en fonction de profils. Dans un souci de clarté, seul le raisonnement pour la garantie DDE est détaillé dans la suite. Pour les autres garanties, une simple présentation des résultats est donnée.

Par la forme du graphique de comparaison de la garantie DDE, il est décidé d'étudier les répartitions des profils des observations dans les catégories suivantes : moins de -50% et plus de 50%.

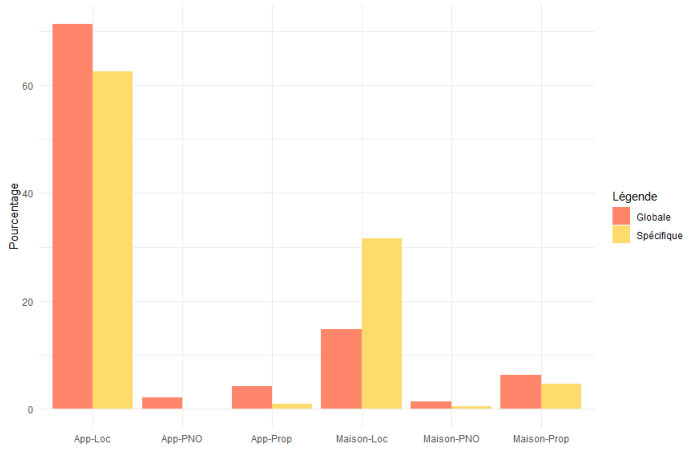

FIGURE 6.7 – Comparaison tarif en fonction de type Habitation et Qualité du Souscripteur - moins de -50% DDE

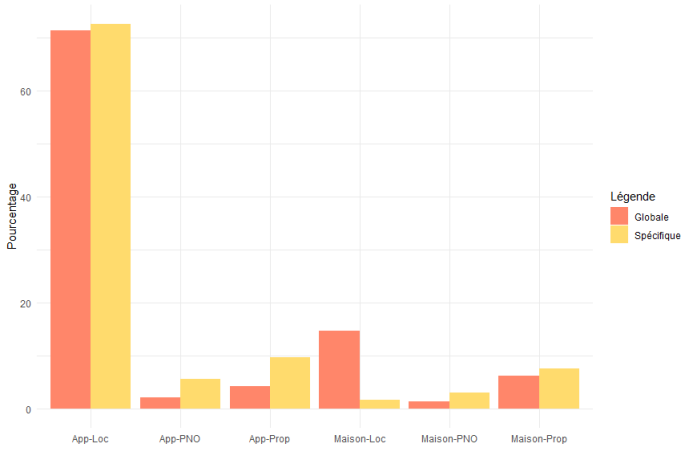

FIGURE 6.8 – Comparaison tarif en fonction de type Habitation et Qualité du Souscripteur - plus 50% DDE

L'étude de ces deux graphiques montre l'existence d'une sous-tarification du tarif estimé par rapport à celui en vigueur dans le cas des profils *Maison-Locataire* ainsi que des surestimations dans le cas des profils PNO et Propriétaire tant appartement que maison.

Les groupes qui se sont démarqués vont être étudiés plus en détail en analysant les autres variables. Dans un premier temps, les potentiels effets du nombre de pièces principales vont être mis en évidence.

Etude de la sous-tarification du profil *Maison-Locataire* :

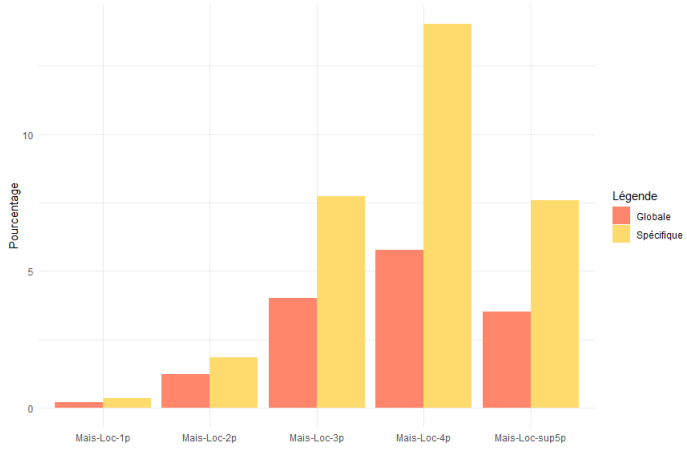

FIGURE 6.9 – Comparaison tarif par nombre de pièces (Maison-Locataire) - moins de −50% DDE

La tendance de sous-tarification pour le profil étudié est plus importante et significative pour les bâtiments les plus grands. En effet, une sur-représentation est remarquée pour les plus de 2 pièces principales par rapport à la représentation totale dans l'ensemble du portefeuille.

L'analyse sur les *Maison-Locataire-plus 2 pièces principales* se poursuit par l'étude de la variable correspondant à la franchise. Les graphiques qui suivent concernent les 3, 4 et plus de 4 pièces principales.

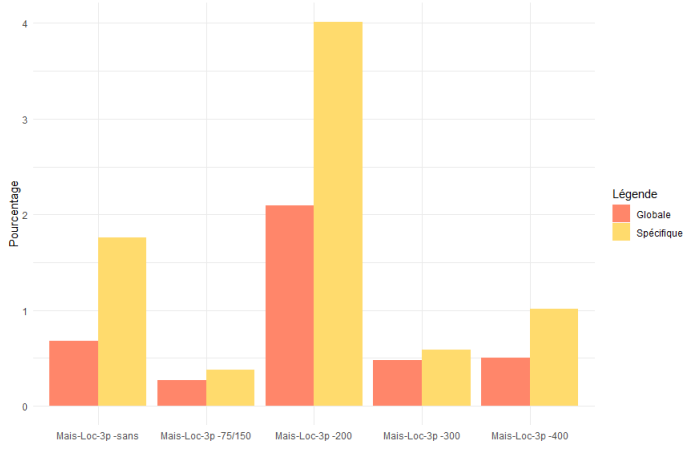

FIGURE 6.10 – Comparaison tarif par franchise (Maison-Locataire-3 pièces principales) - moins de −50% DDE

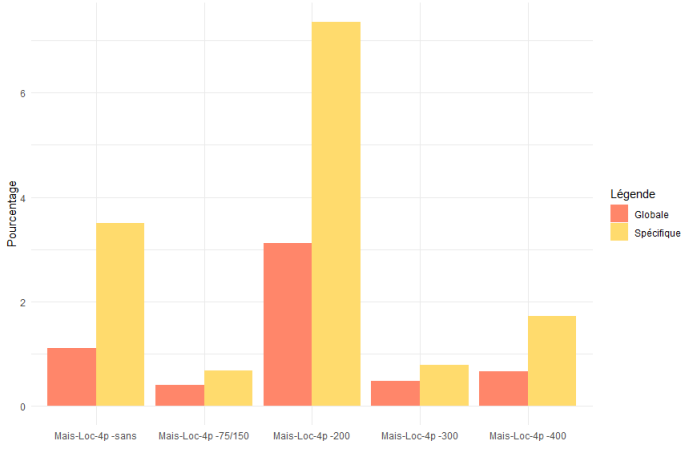

FIGURE 6.11 – Comparaison tarif par franchise (Maison-Locataire- 4 pièces principales) - moins de −50% DDE

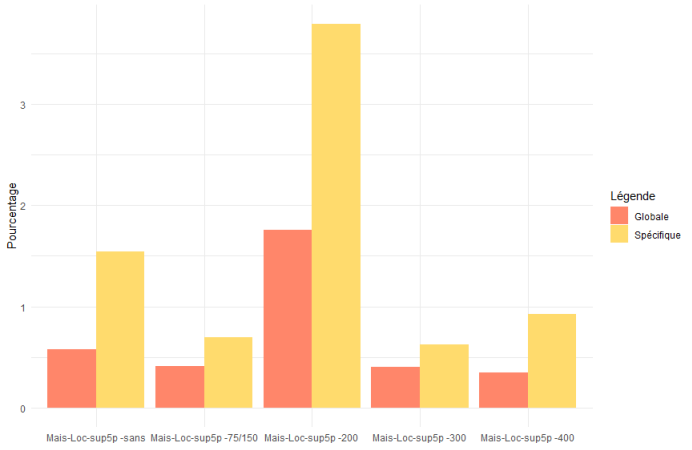

FIGURE 6.12 – Comparaison tarif par franchise (Maison-Locataire- plus 4 pièces principales) - moins de −50% DDE

Dans l'ensemble des trois cas sélectionnés, il existe une présence significative des sans franchise, 200 et 400.

Pour simplifier, l'analyse de la franchise est poursuivie sans séparer les cas mais en gardant directement l'ensemble des plus de 2 pièces principales. Le même phénomène est constaté.

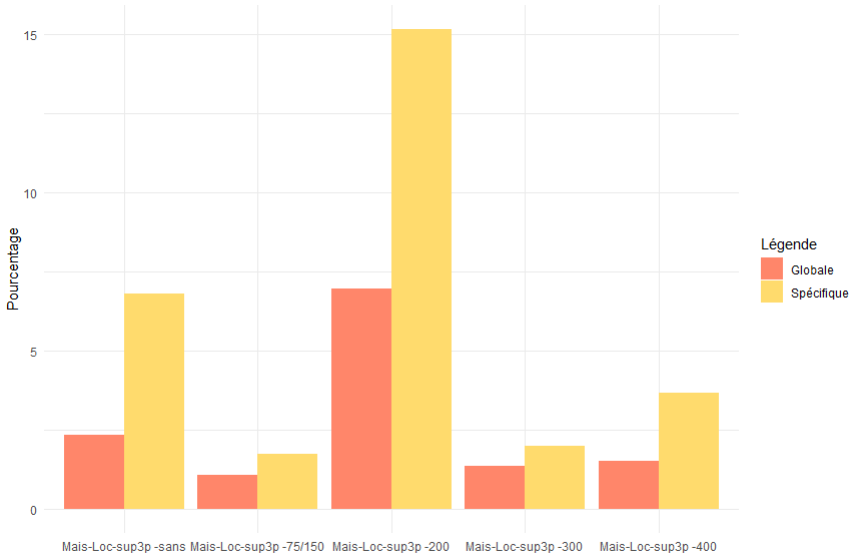

FIGURE 6.13 – Comparaison tarif par franchise (Maison-Locataire-plus de 2 pièces principales) - moins de −50% DDE

Des tests sont ensuite réalisés sur d'autres variables utilisées lors de la tarification pour avoir une vision beaucoup plus précise du type de profil principal en situation de sous-tarification.

Le pourcentage d'OV a permis d'accentuer ces profils. Dans tous les cas de franchise mis en avant dans cette partie, la catégorie sans OV est prédominante (cf figures 6.14, 6.15 et 6.16).

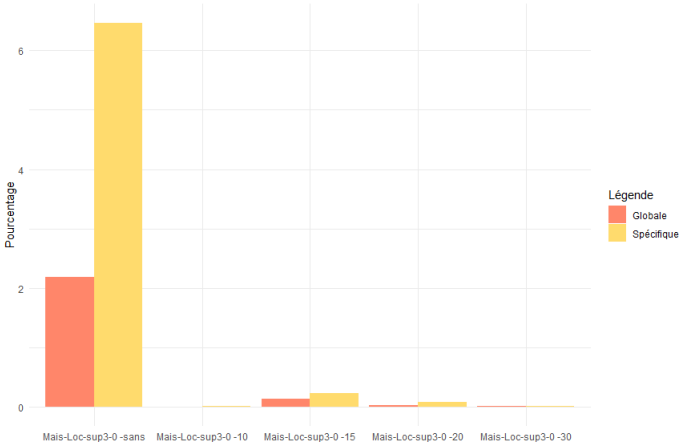

FIGURE 6.14 – Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - sans franchise) - moins de −50% DDE

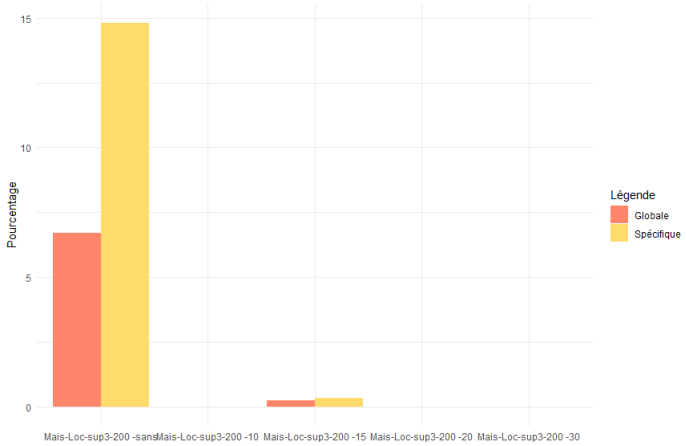

FIGURE 6.15 – Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - 200) - moins de −50% DDE

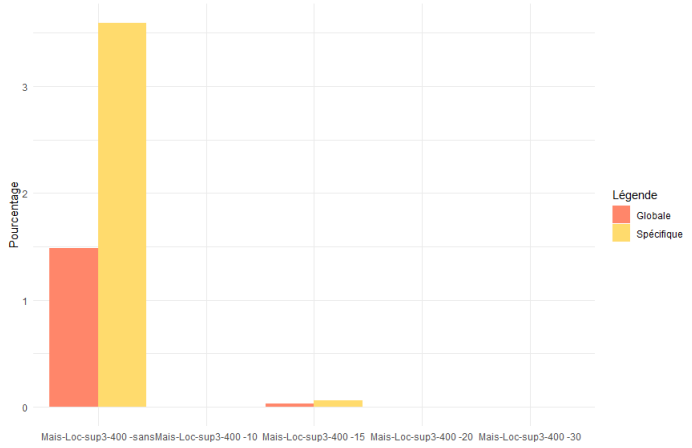

FIGURE 6.16 – Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - 400) - moins de −50% DDE

La conclusion de cette première piste d'analyse sur les profils dans le groupe d'écart moins de −50% est que les profils *Maison-Locataire-plus de 2 pièces principales - sans franchise/200/400 - sans OV* sont les plus présents.

Etude des sur-tarifications :

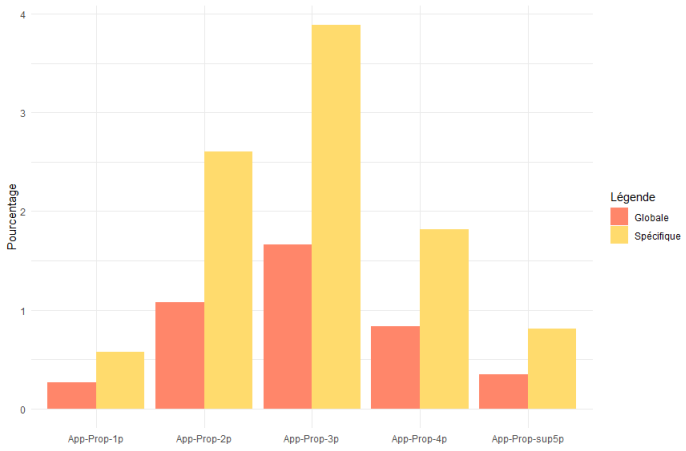

FIGURE 6.17 – Comparaison tarif par nombre de pièces (Appartement-Propriétaire) - plus de 50% DDE

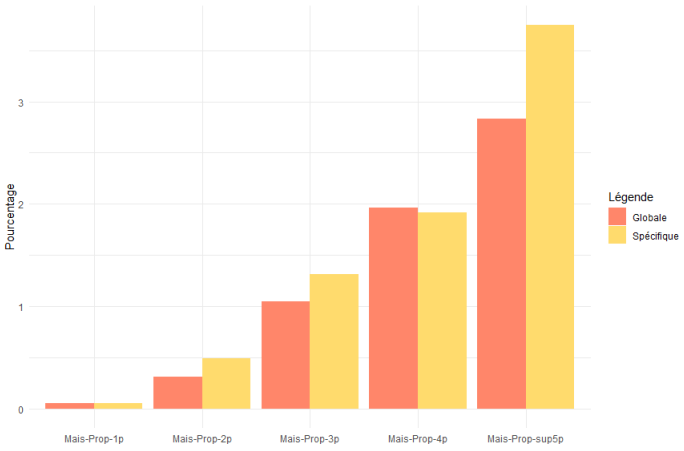

FIGURE 6.18 – Comparaison tarif par nombre de pièces (Maison-Propriétaire) - plus de 50% DDE

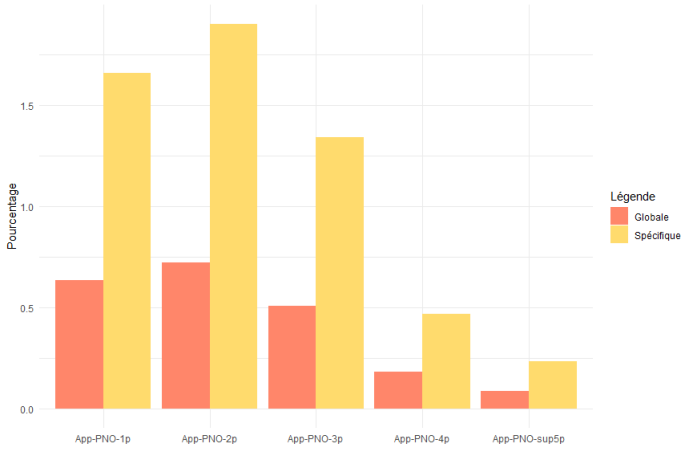

FIGURE 6.19 – Comparaison tarif par nombre de pièces (Appartement-PNO) - plus de 50% DDE

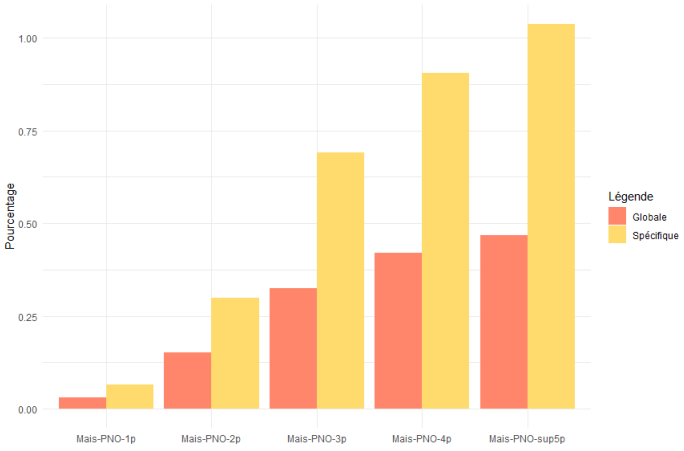

FIGURE 6.20 – Comparaison tarif par nombre de pièces (Maison-PNO) - plus de 50% DDE

Dans le cas des profils *Maison-Propriétaire*, la catégorie *plus de 4 pièces principales* est plus présente que sur l'ensemble du portefeuille. Il y a donc une accentuation de ce profil dans les écarts correspondant à une sur-tarification par rapport au tarif actuel. La tentative pour préciser le profil en sur-proportion n'a pas abouti par l'analyse prenant en compte d'autres variables. Le profil *Maison-Propriétaire-plus de 4 pièces principales* est donc surestimé de manière générale.

Pour les profils avec PNO, la surestimation est présente sur l'ensemble des segments du nombre de pièces. Comme précédemment, l'étude de cette variable n'a pas abouti à une conclusion plus précise sur les profils. Ce constat reste le même pour le profil *Appartement-Propriétaire*. D'autres variables sont maintenant prises en compte afin d'affiner le contrôle des profils en sur-tarification.

La franchise est étudiée pour les trois types de profils cités.

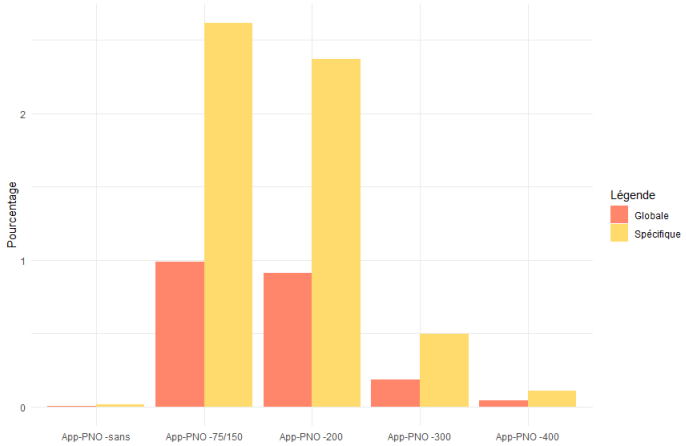

FIGURE 6.21 – Comparaison tarif par franchise (Appartement-PNO) - plus de 50% DDE

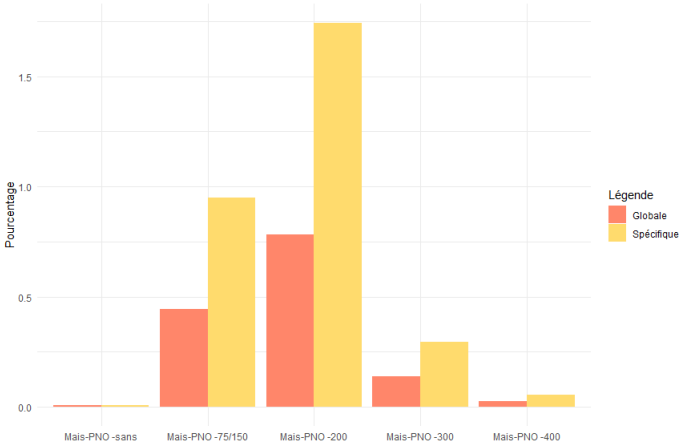

FIGURE 6.22 – Comparaison tarif par franchise (Maison-PNO) - plus de 50% DDE

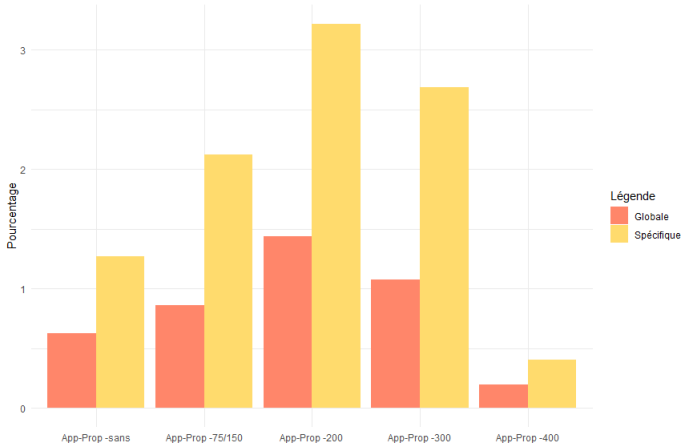

FIGURE 6.23 – Comparaison tarif par franchise (Appartement-Propriétaire) - plus de 50% DDE

Dans tous les cas, l'étude de cette variable ne permet pas d'accentuer les profils qui seraient en surestimation. Pour les deux profils de *PNO* (Appartement et Maison), conclusion est faite que le tarif modélisé est plus élevé que le tarif actuel, entraînant une surestimation globale.

Pour le cas du profil *Appartement-Propriétaire*, une dernière analyse est effectuée pour tenter de mettre en évidence un biais dans le tarif actuel. Elle s'appuie sur la variable de pourcentage d'OV.

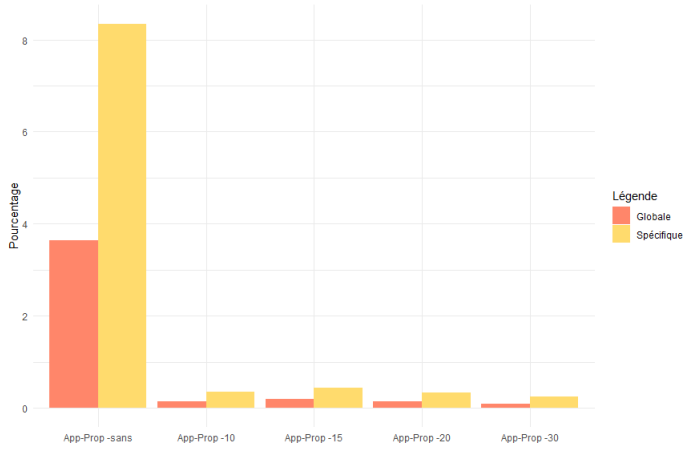

FIGURE 6.24 – Comparaison tarif par pourcentage d'OV (Appartement-Propriétaire) - plus de 50% DDE

Malgré le pic flagrant pour les contrats sans pourcentage OV, le graphique illustre bien que l'ensemble des modalités de la variable sont fortement représentées dans ce type de profil et dans le groupe d'écart de plus de 50%. La même conclusion que dans les cas *PNO* s'impose : les profils *Appartement-Propriétaire* sont en général surestimés dans le tarif obtenu grâce aux modèles.

Dans le cas de la garantie BDG, les mêmes groupes d'écart que pour la garantie DDE sont étudiés. Les analyses effectuées sur cette garantie et pour les écarts concernés n'ont pas été concluantes pour illustrer des profils en sur-tarification ou sous-tarification par rapport au tarif actuel.

Pour les autres garanties (Incendie, Vol et RC), le groupe d'écart dominant (plus de 50% pour l'Incendie et moins de -50% pour les autres) est étudié. Pour ces garanties, seule l'étude des graphiques pour la variable nombres de pièces principales a permis d'obtenir une conclusion sur les profils qui étaient hors de la catégorie d'écart.

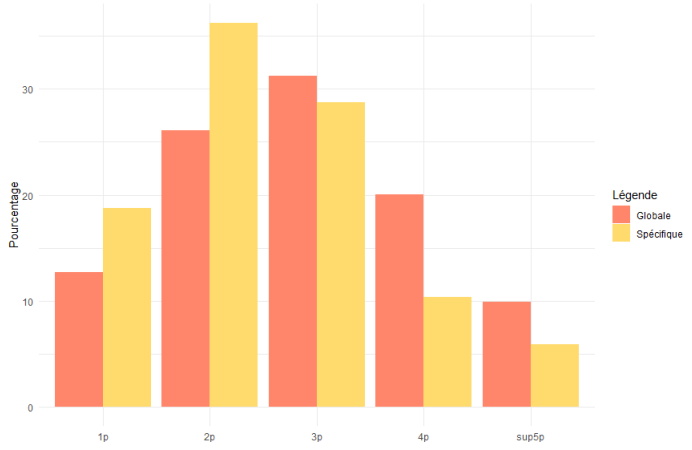

FIGURE 6.25 – Comparaison tarif par nombre de pièces principales - plus de 50% Incendie

Pour la garantie Incendie, dans l'étude du groupe "plus de 50%" d'écart, les habitations de moins de 3 pièces principales sont plus présentes que sur le portefeuille global. A l'inverse, une présence moins importante est observée pour les habitations avec plus de 3 pièces principales. La tendance générale de sur-tarification observée (cf figure 6.4) est donc plus significative pour les moins de 3 pièces principales que sur les plus de 3 pièces principales.

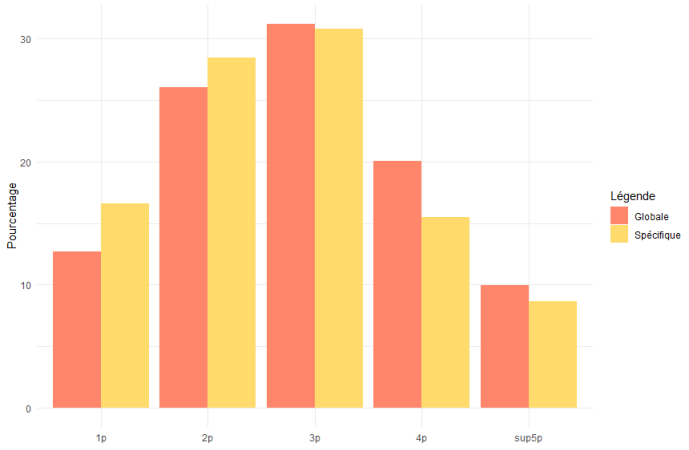

FIGURE 6.26 – Comparaison tarif par nombre de pièces principales - moins de −50% RC

Pour la garantie RC, une sous-tarification globale par rapport au tarif actuel est constatée (cf figure 6.6). Cependant, les habitations avec moins de 3 pièces principales ont tendance à être plus sous-tarifé par rapport au tarif actuel que les habitations de plus de 3 pièces principales (cf figure 6.26).

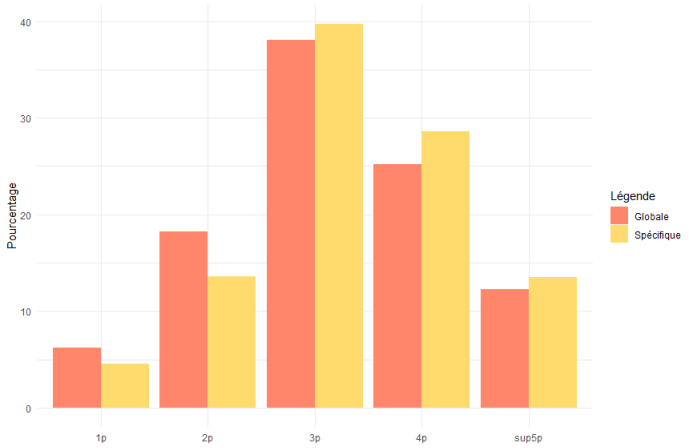

FIGURE 6.27 – Comparaison tarif par nombre de pièces principales - moins de −50% Vol

Comme pour la garantie RC, une sous-tarification globale est présente pour la garantie Vol (cf figure 6.5). Cependant, un effet inverse du cas de la RC est observé. L'étude du graphique 6.27 illustre une tendance de sous-tarification plus marquée pour les plus de 3 pièces principales par rapport aux "petites" habitations (moins de 3 pièces principales).

Le tableau suivant résume les conclusions des premières études effectuées sur la détermination des profils les plus présents par écart :

	Sous-tarification	Sur-tarification
DDE	Maison-Locataire-plus de 2 pièces principales-sans franchise/200/400-sans OV	Maison-Propriétaire-plus de 4 pièces principales
		Appartement/Maison-PNO
		Appartement-Propriétaire
Incendie		Moins de 3 pièces principales
Vol	Plus de 3 pièces principales	
RC	Moins de 3 pièces principales	

TABLE 6.1 – Récapitulatif des profils par garantie

Conclusion et Ouverture

Dans un marché saturé par la forte concurrence entre les différents acteurs, la segmentation apparaît comme un enjeu majeur pour les sociétés d'assurance afin d'estimer le plus finement et précisément possible la tarification des risques inhérents à leur activité. L'objectif d'un assureur est ainsi de pouvoir trouver le bon équilibre entre conservation de la compétitivité tarifaire et rentabilité technique.

Pour mener à bien cette optimisation tarifaire, différents modèles statistiques sont utilisés, classiquement les GLM. La majorité des assureurs se base sur ce type de modèles pour effectuer leur tarification. Mais depuis quelques années, des méthodes de *Data Science* se sont développées. Ces méthodes peuvent s'avérer pertinentes et même plus efficaces qu'une approche par GLM classique. Il y a donc un intérêt à les mettre en œuvre pour des études sur les tarifs, en plus des GLM, afin d'identifier au cas par cas si ces démarches innovantes permettent d'obtenir de meilleurs résultats.

Dans cette optique, ce mémoire se recentre sur la mise en place de divers modèles paramétriques (GLM) et non paramétriques (CART, *Random Forest* et *Gradient Boosting*) dans le but de comparer leur efficacité sur les différentes garanties MRH soumises à cette étude, à la demande de la direction des partenariats de l'ÉQUITÉ (filiale de GENERALI FRANCE), au niveau de l'estimation de la fréquence et du coût.

En ce qui concerne la modélisation, du fait d'une faible corrélation entre les fréquences et les coûts des sinistres, pour chacune des garanties étudiées, une approche "fréquence-coût" s'est montrée pertinente pour la tarification de cette assurance non-vie.

Pour les modèles de fréquence, une première approche basée sur les GLM, utilisant comme loi de distribution les lois de Poisson et Binomiale Négative, a été mise en œuvre. L'analyse des résidus et des indicateurs de performance calculés (MSE et indices de Gini) montrent que l'efficacité des modèles, dans cette approche, est similaire et qu'aucun ne semble se démarquer significativement l'un de l'autre. En complément, des versions "zéro-tronquée" de ces lois ont été tentées en raison d'une masse en zéro observée sur la distribution des fréquences. L'utilisation de ces dernières a prouvé leur inadéquation pour l'estimation des fréquences dans le cas précis de cette étude. Le GLM avec la loi sous-jacente Binomiale Négative est ainsi retenu pour modéliser la fréquence, pour chaque garantie.

Une deuxième approche basée sur les arbres CART a été appliquée à toutes les garanties, à titre de comparaison. Bien que restant performants, ces modèles ont toujours montré une moindre efficacité par rapport aux GLM, dans le cas précis de l'étude. En revanche, ce constat est à nuancer en prenant en compte leur faible complexité. En effet, des méthodes de *Data Science* plus sophistiquées, comme les Forêts Aléatoires et les *Gradient Boosting*, existent. L'application de ces méthodes à l'estimation des fréquences n'a pas permis de conclure à une efficacité supérieure à celle des GLM. Cependant, certains modèles *XGBoost* parviennent à les égaler. Néanmoins, leur mise en œuvre plus complexe tant au niveau des calculs que de l'interprétabilité des résultats obtenus ne permet pas de les rendre compétitifs par rapport aux modèles classiques GLM pour la modélisation des fréquences.

Concernant la construction des modèles de coût, des méthodes GLM utilisant soit une loi de distribution Gamma soit une loi de distribution Log-normale, ont été appliquées. Pour l'ensemble des garanties à l'exception de la garantie BDG, le GLM utilisant la loi Gamma s'est montré le plus efficace. Les performances des modèles étant limitées, en particulier dans le cas de la garantie Incendie, des modèles "toutes garanties confondues" ont été créés. L'analyse des modèles globaux par garantie a démontré une amélioration des performances des indicateurs dans le cas de la garantie Incendie. Cette amélioration peut découler de deux éléments complémentaires : un nombre de sinistres Incendie insuffisant pour déterminer efficacement l'ensemble des effets des variables dans l'estimation du coût et l'influence des sinistres DDE plus fréquents dont les effets sur les coefficients GLM sont souvent similaires à ceux des sinistres Incendie. Pour cette garantie, le modèle GLM Gamma "toutes garanties confondues" a donc été retenu.

Au vu de cette observation singulière, des modèles globaux ont aussi été testés pour toutes les autres méthodes utilisées pour la modélisation des coûts.

Comme pour la fréquence, des modèles CART, *Random Forest* et *XGBoost* ont aussi été mis en œuvre pour toutes nos garanties ainsi que "toutes garanties confondues". En revanche, contrairement à la fréquence où les modèles GLM sélectionnaient les mêmes variables et des regroupements similaires, dans le cas du coût, peu de variables étaient sélectionnées et des regroupements importants ont dû être réalisés. Il a donc été décidé d'effectuer la même approche que dans le cas de la fréquence, en prenant les regroupements des variables désignées par le modèle GLM le plus pertinent, mais aussi d'amener une seconde approche en tentant de créer les différents modèles de *Data Science* à partir de toutes les variables disponibles. Parmi ces types de modèles créés, certains, comme le *Random Forest* toutes variables de la garantie RC ou les modèles *XGBoost* toutes garanties confondues et toutes variables pour les garanties RC, Vol et Incendie, se sont avérés plus performants dans l'estimation que les GLM, malgré la perte d'interprétabilité inhérente à ce type de modèle. Ainsi, ces modèles XGBoost ont été retenus pour les garanties RC, Vol et Incendie. Pour les garanties BDG et DDE, les modèles GLM spécifiques respectivement Log-normal et Gamma restent ceux choisis. L'intérêt de ces nouvelles approches de tarification est démontré pour certaines des garanties. Il apparaît alors pertinent de mettre en place ces études afin d'affiner la tarification au cas par cas.

Au travers des raisonnements menés, il ressort que les GLM sont souvent autant, voire plus performants que les modèles CART, *Random Forest* ou *XGBoost*, en plus d'une mise en œuvre plus aisée et d'une interprétabilité supérieure. Cependant, casuellement, il est possible que l'apport de ces modèles développés grâce à la *Data Science* soit satisfaisant pour palier leurs désavantages.

Après avoir sélectionné les modèles les plus pertinents par garantie vient le temps de la comparaison des tarifs déterminés dans cette étude avec ceux actuellement en vigueur au sein de l'ÉQUITÉ. La comparaison tarifaire "toutes garanties confondues" a montré des écarts significatifs sur une part importante du portefeuille d'analyse, autant en surestimation qu'en sous-estimation, par rapport au tarif actuel. Toutefois, il a été jugé nécessaire d'approfondir cette comparaison tarifaire par garantie. En effet, le schéma global de l'étude des écarts de tarifs est relativement similaire à ceux spécifiques aux garanties BDG et DDE, montrant une sous-tarification et une surtarification par rapport au tarif actuel. Dans le cas des garanties restantes, une surestimation ou sous-estimation générale est constatée.

Ces comparaisons ont ensuite été analysées plus précisément pour vérifier si des profils spécifiques étaient plus présents dans les groupes d'écart représentant un pourcentage significatif du portefeuille d'étude. Cette analyse, ébauchée à la fin de ce mémoire (notamment sur la garantie DDE), devra être complétée pour les autres garanties afin de permettre de cibler finement les biais potentiels du tarificateur actuel. Cela permettra d'envisager d'autres études, particulièrement au niveau de S/C par garantie, dans le but d'effectuer une modernisation du tarif proposé par l'ÉQUITÉ.

De plus, ayant des bases sinistres restreintes pour certaines garanties, les modèles entraînés par garantie peuvent perdre en performance puisqu'ils risquent de ne pas capter les impacts des variables explicatives sur les coûts des sinistres. Cette constatation pourrait aussi expliquer pourquoi les modèles *XGBoost* "toutes garanties confondues" et "toutes variables" se sont avérés les plus performants sur les garanties Incendie, Vol et RC. Le phénomène est encore plus flagrant pour la garantie Incendie dont le GLM le plus performant était aussi celui "toutes garanties confondues" utilisant la loi Gamma. Lorsque les données seront plus étoffées, il serait pertinent de réaliser à nouveau la modélisation par garantie afin de pouvoir, éventuellement, affiner le tarif modélisé.

Bibliographie

[1] O. ALLAIRE : Comparaison de différentes méthodes pour la modélisation de la prime pure d'un produit risque aggravé en assurance automobile. 2020.

[2] F. Assureurs : L'assurance habitation en 2021. 2022.

[3] G. BOUCHTA : Mise en œuvre de méthodes innovantes de tarification. 2017.

[4] A. Charpentier et M. Denuit : Mathématiques de l'assurance non-vie, tome 1 et tome 2. *Economica, Paris*, 2005.

[5] C. Fesquet : Utilisation de facteurs exogènes pour les zoniers en tarification mrh. 2021.

[6] S. LEFEVRE : Elaboration d'un racier et tarification des produits en assurance santé animale. 2019.

[7] I. MEZRAG : Construction d'un zonier en assurance mrh. 2021.

[8] J. PARIENTE : Modélisation du risque géographique en assurance habitation. 2017.

[9] F. PLANCHET et A. MISERAY : Tarification iard, introduction aux techniques avancées. *Cours ISFA*, 2017.

[10] F. PLANCHET et G. SERDECZNY : Modeles fréquence-coût : quelles perspectives d'évolution. *Mars*, 2014.

[11] F.-Z. Zouggagh : Tarifcation automobile à l'aide de modèles de machine learning et apport des données télématiques. 2018.

Annexe A

Glossaire

Abréviation	Signification
AIC	Akaike Information Criterion
BDG	Bris De Glace
BIC	Bayesian Information Criterion
BN	Loi Binomiale Négative
CART	Classification And Regression Tree
CAT	catastrophes naturelles
cp	complexity parameters - paramètres de complexité
DDE	Dégats Des Eaux
EMV	Estimateur de Maximum de Vraisemblance
FFB	Fédération Française du Bâtiment
GLM	Modèles Linéaires Généralisés
i.i.d	Indépendant, Identiquement Distribué
kNN	K-nearest neighbors
LFGN	Loi Forte des Grands Nombres
MRH	Multi-Risque Habitation
MSE/RMSE	erreur quadratique moyenne et sa racine carrée
OV	Objets de Valeurs
PNO	Propriétaire Non Occupant
RC	Responsabilité Civile
RF	Random Forest

Annexe B

Algorithme de Newton-Raphson

Cette méthode qui est utilisée dans les GLM pour approcher les valeurs du vecteur des coefficients β est basée sur le Hessien (la matrice des dérivés secondes), défini par :

$$H_{i,j} = \frac{\partial^2 \mathcal{L}}{\partial \beta_i \partial \beta_j} \tag{B.1}$$

L'algorithme de Newton-Raphson est un algorithme de recherche de zéro d'une fonction réelle à partir d'approximations linéaires successives.

Le principe de cette méthode est simple :

- Partir d'un point x_0 proche de la solution
- La tangente de la fonction en x_0 est utilisée pour trouver le point suivant x_1 (ça sera le point de jonction entre l'abscisse et la tangente)
- Ce principe est réitéré

Plus simplement, la formule de récurrence sera :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{B.2}$$

Dans le cas de l'approximation des coefficients GLM, la partie fractionnaire de la formule de récurrence sera représentée par le Hessien.

Annexe C

Provisionnement Non-Vie

Dans cette annexe est développé le provisionnement par *Chain-Ladder* qui a été utilisé pour le provisionnement de l'étude. De plus, les différents *C-C plots*, ayant permis de justifier l'application de *Chain-Ladder* sur les données, sont représentés.

C.1 Théorie de *Chain-Ladder*

Il est supposé que les sinistres peuvent se dérouler sur $n+1$ années.
Plusieurs hypothèses doivent être respectées avant d'appliquer la méthode de *Chain-Ladder* :

- Pour j fixé, l'alignement des couples $(C_{i,j}, C_{i,j+1})_{i=1,...,n-j-1}$ sur une droite passant par l'origine doit être respecté,
  où $C_{i,j}$ est le règlement cumulé pour l'année i et pour le délai de développement j.
- Pour $j = 0, ..., n$, les $(f_{i,j})$ doivent être sensiblement constants.

Les facteurs de développement individuels peuvent être écrits comme suit :

$$f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \tag{C.1}$$

Le facteur de développement pour le délai de règlement j est donné par la formule suivante :

$$f_j = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}} \tag{C.2}$$

Ainsi, pour avoir la charge à l'ultime, la formule utilisée est la suivante :

$$S_i = C_{i,n} = C_{i,n-i} \prod_{n-i}^{n-1} f_h \tag{C.3}$$

C.2 Application sur les données

Les différents tableaux et *C-C plots* pour la justification de l'application de la méthode sont présentés ci-dessous.
Le premier est celui des facteurs de développement individuels en fonction de chaque année et du délai de règlement.

	0	1	2	3	4	5	6
2015	62,93942365	1,269708917	1,021550088	1,037404912	1,008741131	1,002517108	1,009429428
2016	26,0026918	1,369912842	1,05073413	1,033285467	1,003163106	1,013531144	
2017	39,89644461	1,317712115	1,077352473	1,025770386	1,006341405		
2018	24,10267303	1,377276856	1,067839985	1,030707783			
2019	38,83485899	1,553304971	1,080921516				
2020	49,09969975	1,466753303					
2021	65,89711908						

FIGURE C.1 – Tableau des facteurs de développement individuels

Une analyse plus poussée des facteurs de développement individuels par délai de règlement est ensuite faite afin de prouver la seconde hypothèse du modèle.

	0	1	2	3	4	5	6
Moyenne	43,82470156	1,392444834	1,059679638	1,031792137	1,006081881	1,008024126	1,009429428
Ecart-Type	16,46866719	0,102746849	0,024310607	0,004870681	0,002798054	0,007788099	
Coeff variation	37,6%	7,4%	2,3%	0,5%	0,3%	0,8%	0,0%

FIGURE C.2 – Analyse des colonnes du tableau précédent

Les *C-C plots* créés montrent bien que la première hypothèse pour le modèle est applicable puisque les R^2 des courbes de tendance linéaire sont proches de 1.

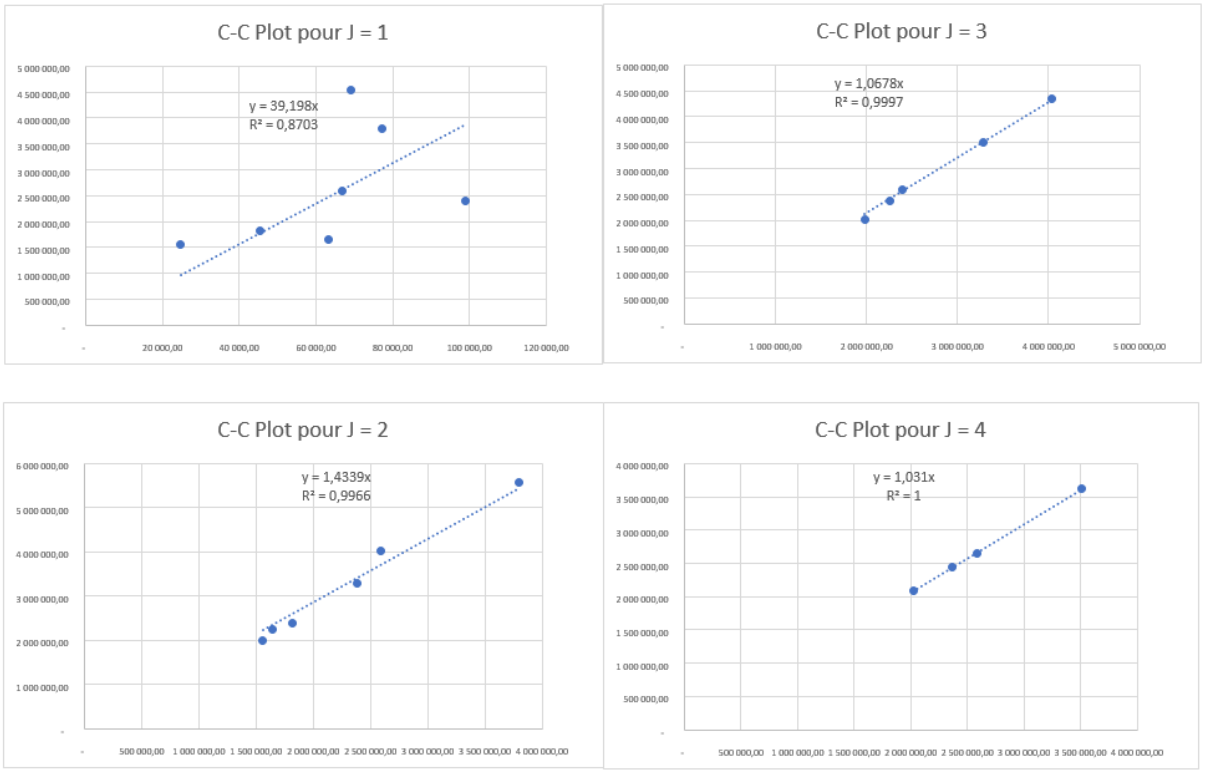

FIGURE C.3 – *C-C plots* pour le provisionnement

Annexe D

Graphiques complémentaires pour l'étude des indices *As If*

Dans cette annexe sont présentés les différents graphiques qui ont permis de choisir l'indice pour passer les montants de règlement en *As If*.

D.1 Indice FFB

Dans un premier temps, une analyse avec les sinistres toutes garanties confondues est faite.

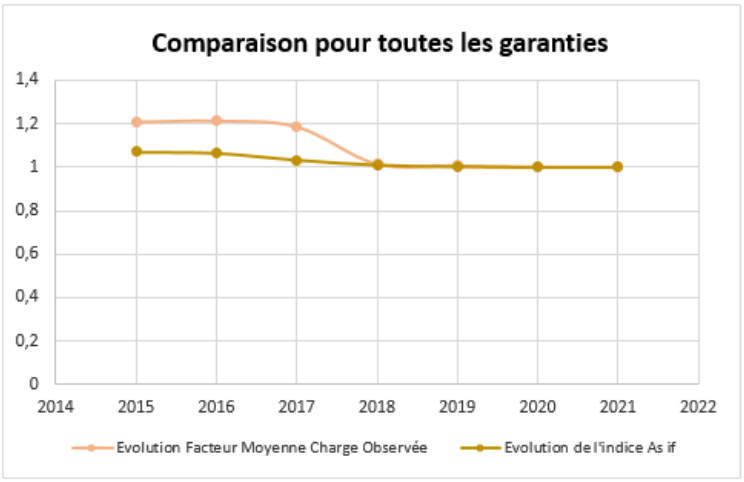

FIGURE D.1 – Comparaison des données et de l'indice FFB sur toutes les garanties

Cependant, afin de pousser plus loin l'analyse, les mêmes graphiques, garantie par garantie, sont étudiés.

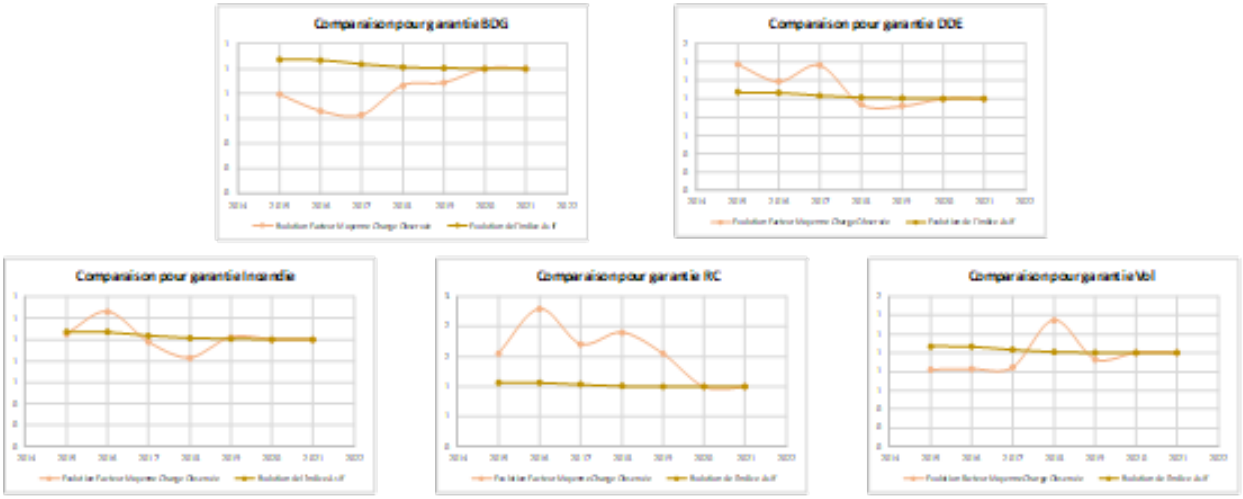

FIGURE D.2 – Comparaison des données et de l'indice FFB par garantie

D.2 Indice de l'inflation

L'analyse avec les sinistres, toutes garanties confondues, est faite en prenant cette fois l'indice d'inflation.

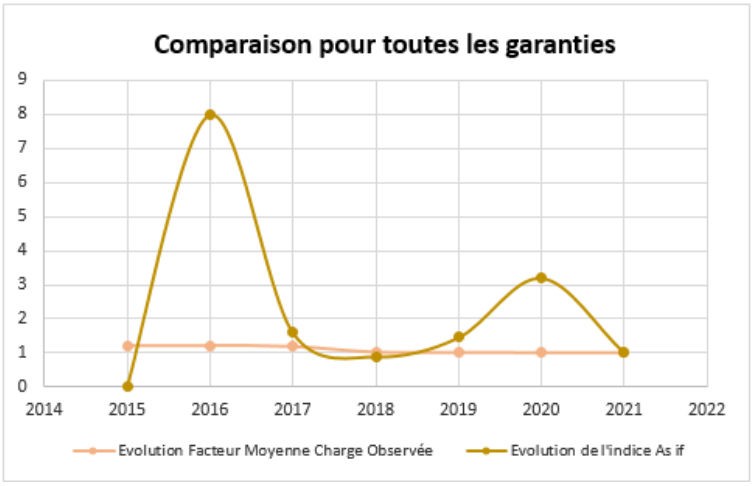

FIGURE D.3 – Comparaison des données et de l'indice de l'inflation sur toutes les garanties

Enfin, l'analyse poussée en regardant les mêmes graphiques, pour chaque garantie, est reprise pour cet indice.

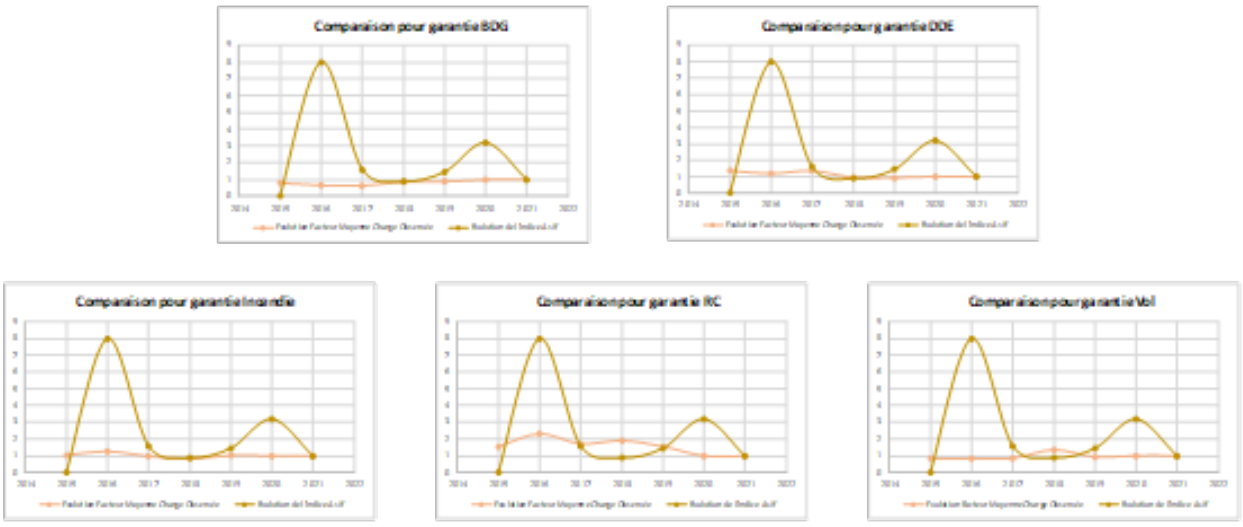

FIGURE D.4 – Comparaison des données et de l'indice de l'inflation par garantie

Annexe E

Les tests de corrélation

E.1 Pearson

Contrairement aux deux corrélations suivantes, la corrélation de Pearson est une corrélation de type paramétrique.

Le coefficient de corrélation de Pearson est le coefficient de corrélation le plus souvent utilisé et donc qualifié de "coefficient de corrélation".

Ce coefficient peut être défini par la formule suivante :

$$\frac{\sum(x - m_X)(y - m_Y)}{\sqrt{\sum(x - m_X)^2 \sum(y - m_Y)^2}} \tag{E.1}$$

où x et y sont les vecteurs des variables et m_X et m_Y sont les moyennes de respectivement x et y.

E.2 Kendall

La corrélation de Kendall est un coefficient de corrélation qui est basé sur les rangs des variables. Cette corrélation est une corrélation non paramétrique.

La corrélation de Kendall est définie par la formule suivante :

$$\frac{Nombre\ de\ paires\ concordantes - Nombre\ de\ paires\ discordantes}{\frac{n(n-1)}{2}} \tag{E.2}$$

où n est le nombre d'observations.

Une paire d'information (x_1, y_1) et (x_2, y_2) est

- concordante si ($x_1 > x_2$ et $y_1 > y_2$) ou ($x_1 < x_2$ et $y_1 < y_2$)
- discordante si ($x_1 > x_2$ et $y_1 < y_2$) ou ($x_1 < x_2$ et $y_1 > y_2$)

E.3 Spearman

La corrélation de Spearman est habituellement utilisée pour établir une relation entre les variables étudiés de type affine.

Le coefficient de corrélation de Spearman est estimé à partir des rangs des variables et non pas des valeurs prises par ces variables.

Il est défini de la manière suivante :

$$r_s = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{E.3}$$

où rg_X et rg_Y sont les rangs des variables et σ_{rg_X} et σ_{rg_Y} les écart-types des rangs des variables.

Cette corrélation peut être interprétée comme étant la corrélation de Pearson des variables des rangs.

Annexe F

Les lois de fréquences

F.1 La loi de Poisson

Cette loi de probabilité discrète est définie par son paramètre λ.
Supposons une variable Y qui suit une loi de Poisson alors :

$$\mathbb{P}(Y = k) = exp(-\lambda)\frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{N} \tag{F.1}$$

La loi de Poisson fait partie de la famille exponentielle définie par l'équation 2.6.
En notant les probabilités de la manière suivante, $\mathbb{P}(Y = k) = exp(klog\lambda - \lambda - log(k!))$, il peut être identifié, en notant $\theta = log(\lambda)$ et $\phi = 1$:

- $a(\phi) = 1$
- $b(\theta) = \exp\theta = \lambda$
- $c(k, \phi) = -log(k!)$

La loi de Poisson a une moyenne et une variance égales ($\mathbb{E}[X] = \mathbb{V}[X] = \lambda$). C'est cette égalité qui est caractéristique de cette loi. Son utilisation est réservée au cas où un phénomène de sur-dispersion n'est pas observé. Elle est un peu plus restrictive que la loi Binomiale Négative.

F.2 La loi Binomiale Négative

La loi Binomiale Négative est un mélange entre une loi de Poisson et une loi Gamma. Dans ce cas, le paramètre λ de la loi n'est pas déterministe mais est défini comme une loi aléatoire suivant une loi Gamma.
Cette loi a plusieurs paramètres μ (la moyenne) et k (paramètre de dispersion). Si une variable aléatoire Y suit une telle loi, alors :

$$f(y) = \frac{\Gamma(k+y)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y, \quad \forall y \in \mathbb{N} \tag{F.2}$$

$$= exp\left(ylog\left(\frac{\mu}{k+\mu}\right) + klog\left(\frac{k}{k+\mu}\right) + log\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right)\right) \tag{F.3}$$

En notant $\theta = log\left(\frac{\mu}{k+\mu}\right)$ et $\phi = 1$, il est possible identifier les valeurs de la famille exponentielle :

- $a(\phi) = 1$
- $b(\theta) = -klog(1 - exp(\theta))$
- $c(k, \phi) = log\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right)$

La moyenne et la variance de cette loi peuvent être définies comme :

- $\mathbb{E}[Y] = k\frac{exp(\theta)}{1-exp(\theta)} = \mu$
- $\mathbb{V}[Y] = k\frac{exp(\theta)}{(1-exp(\theta))^2} = \mu + \frac{\mu^2}{k}$

La variance de cette loi est toujours supérieure à l'espérance. Cette loi peut donc être utilisée quand il y a de la sur-dispersion sur les données.

F.3 Les lois "zéro-tronquée"

Ces lois sont des mélanges de deux composantes :

- loi de type Binomiale qui gére les 0 (ne proviennent plus de la loi de comptage)
- loi de comptage tronquée

La probabilité suivante caractérise ces lois :

$$\mathbb{P}(N = k) = \begin{cases} f_{zero}(0) & \text{si } k = 0 \\ (1 - f_{zero}(0))\frac{f_{count}(k)}{1-f_{count}(0)} & \text{si } k > 0 \end{cases} \tag{F.4}$$

Avec f_{count} la densité de la loi de la partie de fréquence non nulle et f_{zero} la densité qui défini s'il y a un zéro. Ce type de distribution est utilisé dans le cas où les fréquences sont souvent égales à 0.

Annexe G

Les lois de coût

G.1 La loi Normale

La loi Normale est une loi de probabilité continue qui est paramétrée par deux facteurs, sa moyenne μ et sa variance σ^2.
La densité de cette loi a la forme suivante :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad \forall y \in \mathbb{R} \tag{G.1}$$

$$= exp\left(\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2} - \frac{\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)}{2}\right) \tag{G.2}$$

Cette loi fait partie de la famille exponentielle. Par identification et en notant $\theta = \mu$ et $\phi = \sigma^2$, il est obtenu :

- $a(\phi) = \phi$
- $b(\theta) = \frac{\theta^2}{2}$
- $c(y, \phi) = -\left(\frac{\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)}{2}\right)$

La moyenne et la variance de cette loi peuvent être définies comme :

- $\mathbb{E}[Y] = \mu$
- $\mathbb{V}[Y] = \sigma^2$

En pratique, cette loi n'est pas souvent utilisée pour la modélisation du coût en raison de sa symétrie et de la possibilité d'avoir des valeurs négatives. C'est pour cette raison que d'autres lois sont généralement préférées, comme la loi Gamma.

G.2 La loi Gamma

La loi Gamma est une loi de probabilité continue qui est paramétrée par deux facteurs, un de forme α et l'autre d'intensité β.

Avec ces paramètres, une loi Gamma a la densité suivante :

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} exp(-\beta y), \quad \forall y \in \mathbb{R}^+ \tag{G.3}$$

$$= exp\left(\frac{-y\frac{\beta}{\alpha} + log(\beta)}{\alpha^{-1}} + (\alpha - 1)log(y) - log(\Gamma(\alpha))\right) \tag{G.4}$$

La fonction Γ a l'expression suivante :

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} exp(x) dx, \quad \forall \alpha > 0 \tag{G.5}$$

Cette loi fait partie de la famille exponentielle. Par identification et en notant $\theta = \frac{-\beta}{\alpha}$ et $\phi = \frac{1}{\alpha}$, il est obtenu :

- $a(\phi) = \phi$
- $b(\theta) = -log(-\alpha\theta)$
- $c(k, \phi) = (\alpha - 1)log(y) - log(\Gamma(\alpha))$

La moyenne et la variance de cette loi peuvent être définies comme :

- $\mathbb{E}[Y] = \frac{\alpha}{\beta}$
- $\mathbb{V}[Y] = \frac{\alpha}{\beta^2}$

La loi Gamma est fréquemment utilisée pour la modélisation de la sévérité en raison de sa positivité.

G.3 La loi Log-normale

Dans les modèles GLM réalisés, la loi Log-normale a pu être utilisée. Cette loi ne fait pas partie de la famille exponentielle. Pour cette raison, une manipulation doit être effectuée afin de réaliser les GLM avec cette loi.

Ce GLM s'obtient en effectuant un GLM gaussien, non pas sur le coût, mais sur le logarithme du coût.

En effet, par définition, l'équivalence suivante existe :

$$Y \sim \mathcal{LN}(\mu, \sigma^2) \Leftrightarrow log(Y) \sim \mathcal{N}(\mu, \sigma^2) \tag{G.6}$$

Le principal défaut de cette méthode vient de l'espérance et de la variance :

- $\mathbb{E}[Y] = e^{\mu + \frac{\sigma^2}{2}} \neq e^{\mathbb{E}[logY]}$
- $\mathbb{V}[Y] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) \neq e^{\mathbb{V}[logY]}$

Ainsi, afin d'obtenir les prédictions de coût, il faudra multiplier les prédictions du modèle créé grâce à la fonction logarithmique par $e^{\frac{\sigma^2}{2}}$.

Annexe H

Tableaux complémentaires pour les GLM fréquence

H.1 Tableaux pour adéquation des lois avant modélisation

	Poisson	**Binomiale Négative**
Déviance	23 363	20 342
AIC	28 270	28 241

TABLE H.1 – Tableau des indicateurs avant modélisation pour la garantie BDG

	Poisson	**Binomiale Négative**
Déviance	104 274	89 327
AIC	133 724	133 323

TABLE H.2 – Tableau des indicateurs avant modélisation pour la garantie DDE

	Poisson	**Binomiale Négative**
Déviance	15 984	14 700
AIC	19 061	19 047

TABLE H.3 – Tableau des indicateurs avant modélisation pour la garantie Incendie

	Poisson	**Binomiale Négative**
Déviance	18 505	15 193
AIC	22 164	22 126

TABLE H.4 – Tableau des indicateurs avant modélisation pour la garantie Vol

	Poisson	Binomiale Négative
Déviance	28 686	23 324
AIC	34 368	34 298

TABLE H.5 – Tableau des indicateurs avant modélisation pour la garantie RC

H.2 Tableaux complémentaires pour la significativité des variables sélectionnées

Variables	Df	Statistique	p-value
Zone	9	132.5434	$< 2.2e-16$
typeHabetage	4	8.4939	0.07507
typeResQualiSous	3	28.2328	$3.246e-06$
typedistribution	3	24.3086	$2.153e-05$
sin_ant	1	27.8991	$1.278e-07$
fran_globale	5	14.3903	0.01331
pourcentageOV	4	30.7625	$3.423e-06$
montantCapitalMobilierp	10	71.9315	$1.876e-11$

TABLE H.6 – Analyse de type III pour la loi de Poisson et la garantie Vol

Variables	Df	Statistique	p-value
Zone	9	130.121	$< 2.2e-16$
typeHabetage	4	8.333	0.08012
typeResQualiSous	3	26.764	$6.598e-06$
typedistribution	3	23.688	$2.902e-05$
sin_ant	1	28.357	$1.009e-07$
fran_globale	5	14.421	0.01314
pourcentageOV	4	30.324	$4.205e-06$
montantCapitalMobilierp	10	67.455	$1.371e-10$

TABLE H.7 – Analyse de type III pour la loi Binomiale Négative et la garantie Vol

Variables	Df	Statistique	p-value
bdgprisoption	1	5.2912	0.021434
nbPiecesPrincipales	9	98.2355	$< 2.2e-16$
Zone	9	24.4819	0.003601
typeHabetage	4	16.4885	0.002429
typeResQualiSous	3	26.4894	$7.532e-06$
typedistribution	3	32.0847	$5.023e-07$
sin_ant	1	73.7957	$< 2.2e-16$
fran_globale	5	575.5286	$< 2.2e-16$
pourcentageOV	4	13.0264	0.011148
montantCapitalMobilierp	10	42.4914	$6.130e-06$

TABLE H.8 – Analyse de type III pour la loi de Poisson et la garantie BDG

Variables	Df	Statistique	p-value
bdgprisoption	1	5.286	0.021498
fran_globale	5	563.018	$< 2.2e-16$
nbPiecesPrincipales	9	96.685	$< 2.2e-16$
montantCapitalMobilierp	10	41.835	$8.022e-06$
typedistribution	3	31.365	$7.123e-07$
sin_ant	1	72.028	$< 2.2e-16$
Zone	9	23.773	0.004675
typeResQualiSous	3	25.940	$9.818e-06$
typeHabetage	4	15.862	0.003211
pourcentageOV	4	12.948	0.011530

TABLE H.9 – Analyse de type III pour la loi Binomiale Négative et la garantie BDG

Variables	Df	Statistique	p-value
nbPiecesPrincipales	9	186.7429	$< 2.2e-16$
Zone	9	34.2792	$7.979e-05$
typeHabetage	4	41.8142	$1.823e-08$
typeResQualiSous	3	39.0033	$1.733e-08$
typedistribution	3	6.4136	0.09313
nboptions	11	24.4404	0.01100
sin_ant	1	95.9252	$< 2.2e-16$
fran_globale	5	91.9642	$< 2.2e-16$

TABLE H.10 – Analyse de type III pour la loi de Poisson et la garantie RC

Variables	Df	Statistique	p-value
nbPiecesPrincipales	9	183.2901	$< 2.2e-16$
Zone	9	33.4956	0.0001095
typeHabetage	4	40.3853	$3.603e-08$
typeResQualiSous	3	38.3223	$2.416e-08$
typedistribution	3	6.4203	0.0928579
nboptions	11	24.0227	0.0126379
sin_ant	1	94.0894	$< 2.2e-16$
fran_globale	5	89.0610	$< 2.2e-16$

TABLE H.11 – Analyse de type III pour la loi Binomiale Négative et la garantie RC

Variables	Df	Statistique	p-value
nbPiecesPrincipales	9	95.1679	$< 2.2e-16$
Zone	9	26.1401	0.0019379
typeHabetage	4	152.9281	$< 2.2e-16$
typeResQualiSous	3	144.9766	$< 2.2e-16$
typedistribution	3	19.2066	0.0002478
nboptions	11	57.0084	$3.315e-08$
sin_ant	1	63.4058	$1.682e-15$
pourcentageOV	4	8.2941	0.0813810
montantCapitalMobilierp	10	28.5262	0.0014861

TABLE H.12 – Analyse de type III pour la loi de Poisson et la garantie Incendie

Variables	Df	Statistique	p-value
nbPiecesPrincipales	9	93.9074	$2.669e-16$
Zone	9	25.7584	0.0022374
typeHabetage	4	152.0434	$< 2.2e-16$
typeResQualiSous	3	143.5372	$< 2.2e-16$
typedistribution	3	19.0353	0.0002688
nboptions	11	52.1061	$2.608e-07$
sin_ant	1	63.5116	$1.594e-15$
pourcentageOV	4	7.5029	0.1115814
montantCapitalMobilierp	10	28.1843	0.0016865

TABLE H.13 – Analyse de type III pour la loi Binomiale Négative et la garantie Incendie

Variables	Df	Statistique	p-value
fran_globale	5	167.387	$< 2.2e-16$
Zone	9	507.546	$< 2.2e-16$
typeResQualiSous	3	268.555	$< 2.2e-16$
typeHabetage	4	506.903	$< 2.2e-16$
nbPiecesPrincipales	9	346.103	$< 2.2e-16$
sin_ant	1	557.245	$< 2.2e-16$
montantCapitalMobilierp	10	76.287	$2.670e-12$
pourcentageOV	4	49.622	$4.329e-10$
typedistribution	3	31.072	$8.208e-07$
nboptions	11	46.557	$2.577e-06$

TABLE H.14 – Analyse de type III pour la loi Binomiale Négative et la garantie DDE

Annexe I

Illustrations complémentaires pour les GLM fréquence

I.1 Garantie DDE

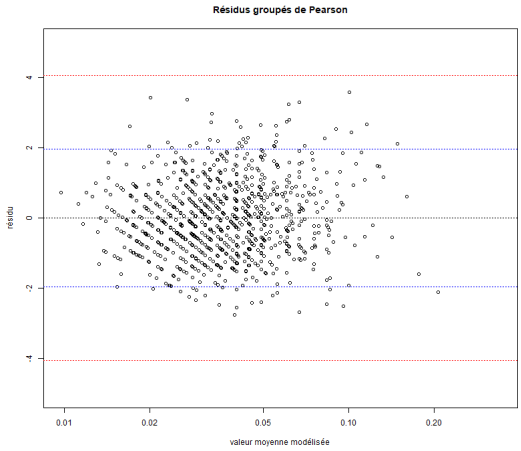

FIGURE I.1 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - DDE

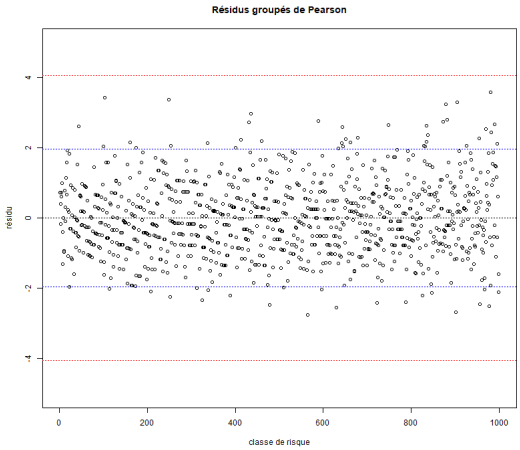

FIGURE I.2 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - DDE

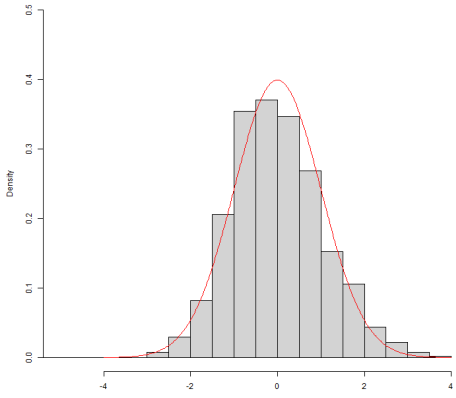

FIGURE I.3 – Distribution empirique des résidus groupés de Pearson - BN - DDE

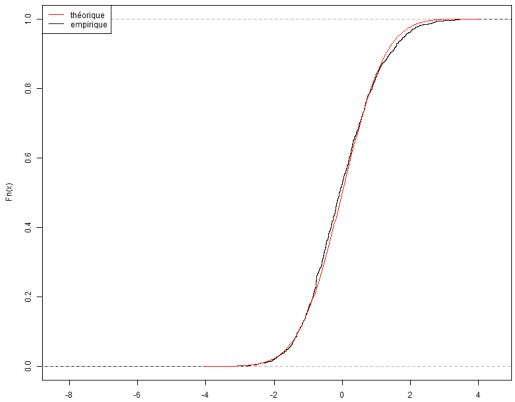

FIGURE I.4 – Fonction de répartition empirique des résidus de Pearson - BN - DDE

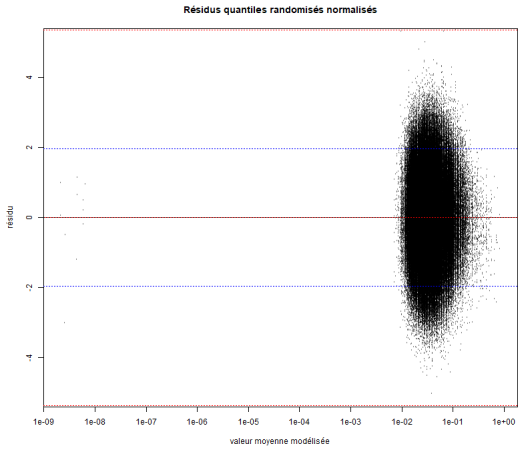

FIGURE I.5 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - DDE

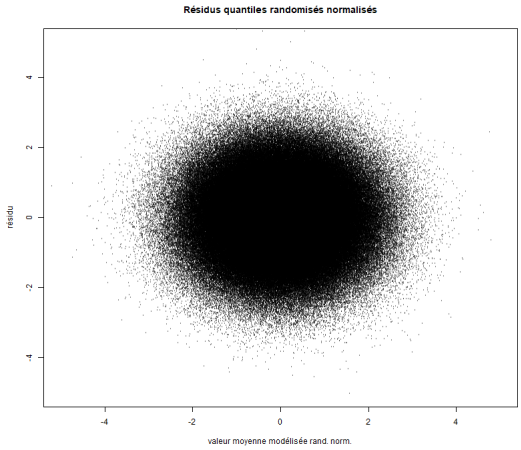

FIGURE I.6 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - DDE

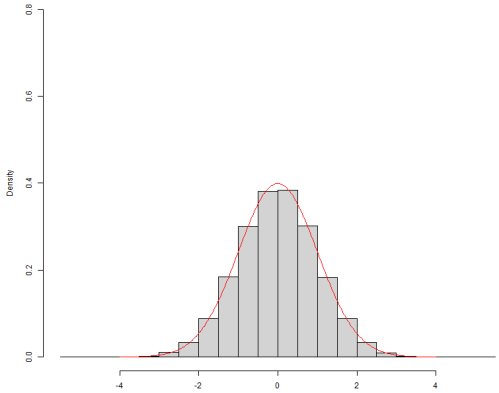

FIGURE I.7 – Distribution empirique des résidus quantiles randomisés normalisés - BN - DDE

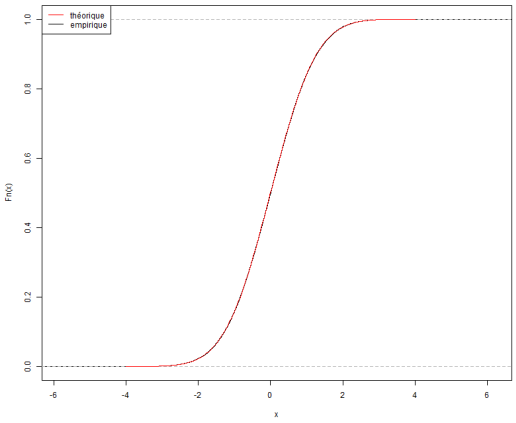

FIGURE I.8 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - DDE

I.2 Garantie BDG

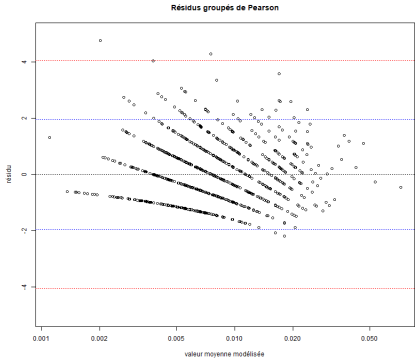

FIGURE I.9 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - BDG

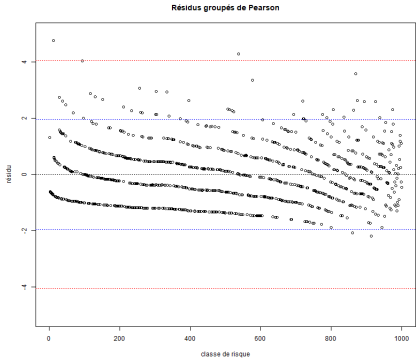

FIGURE I.10 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - BDG

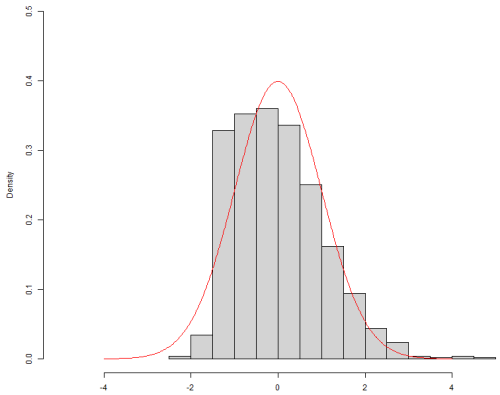

FIGURE I.11 – Distribution empirique des résidus groupés de Pearson - Poisson - BDG

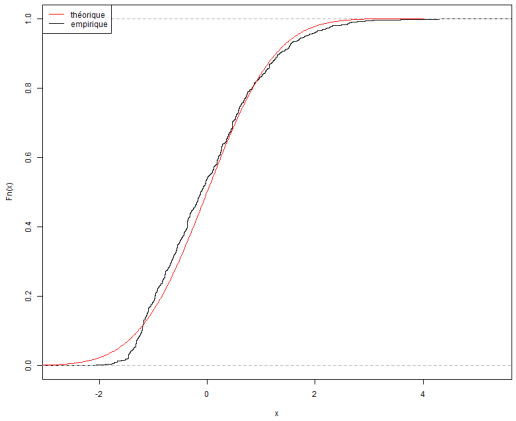

FIGURE I.12 – Fonction de répartition empirique des résidus de Pearson - Poisson - BDG

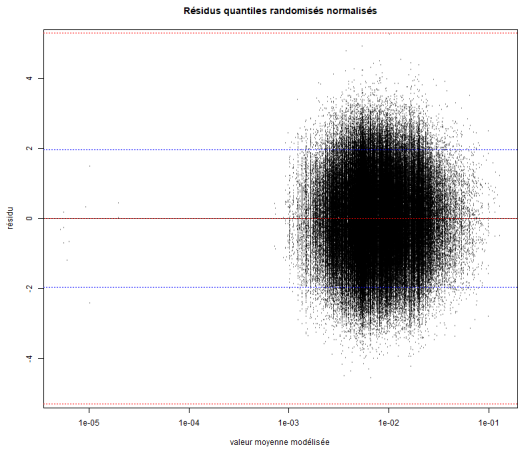

FIGURE I.13 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - BDG

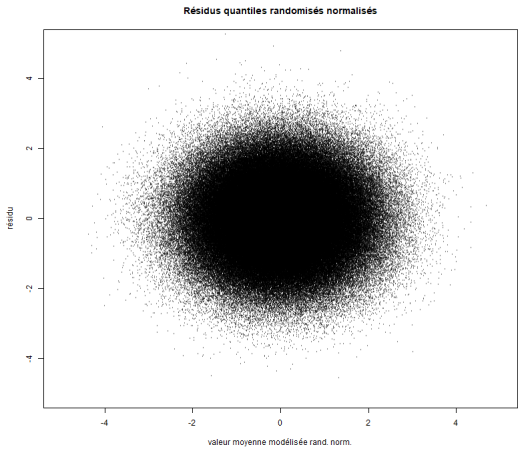

FIGURE I.14 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - BDG

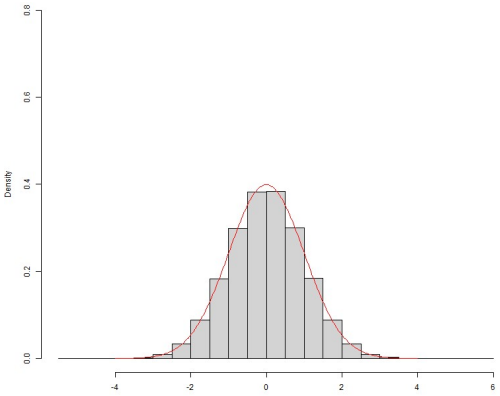

FIGURE I.15 – Distribution empirique des résidus quantiles randomisés normalisés - Poisson - BDG

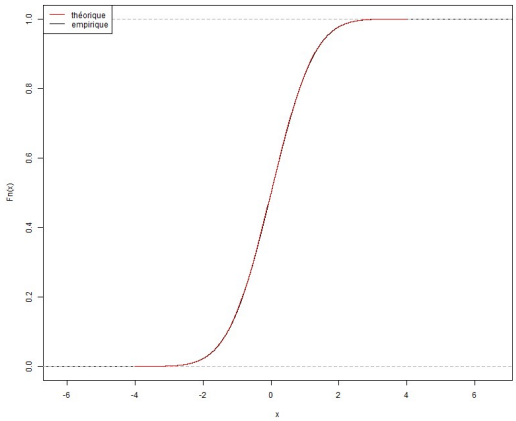

FIGURE I.16 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - BDG

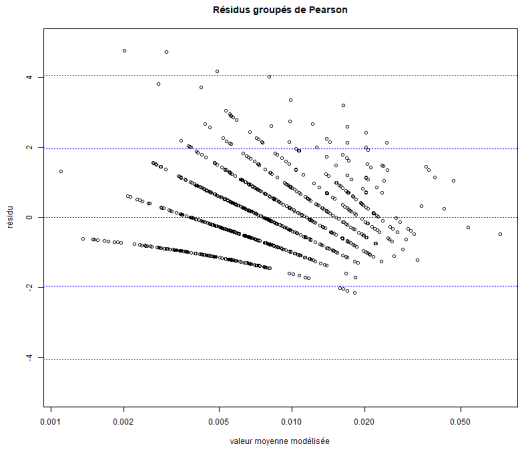

FIGURE I.17 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - BDG

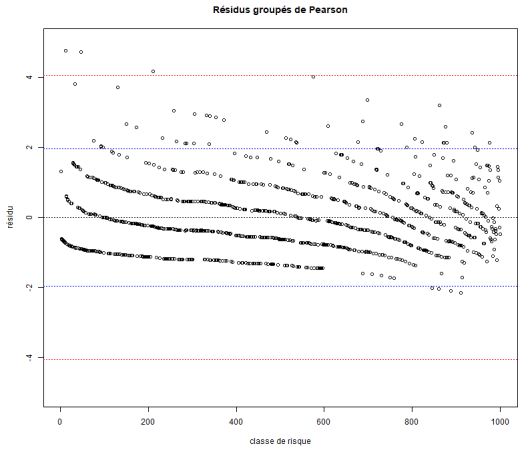

FIGURE I.18 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - BDG

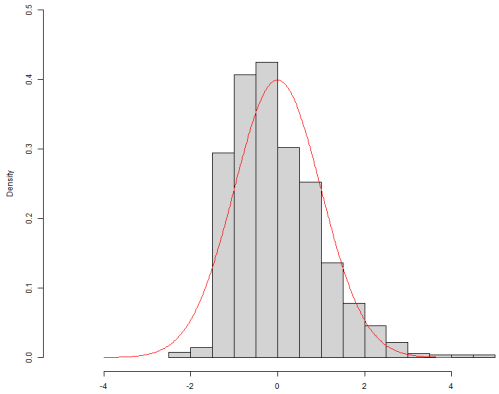

FIGURE I.19 – Distribution empirique des résidus groupés de Pearson - BN - BDG

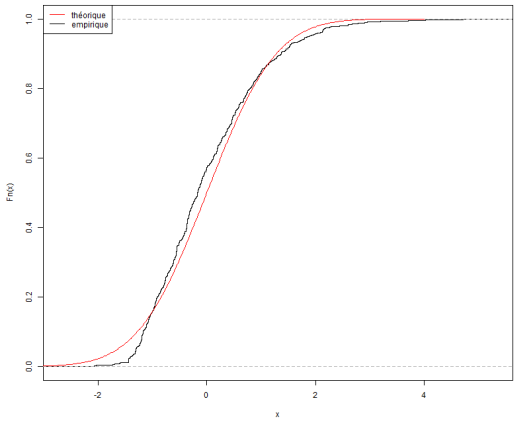

FIGURE I.20 – Fonction de répartition empirique des résidus de Pearson - BN - BDG

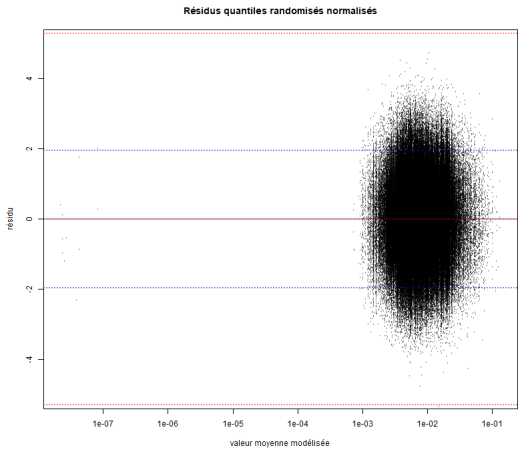

FIGURE I.21 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - BDG

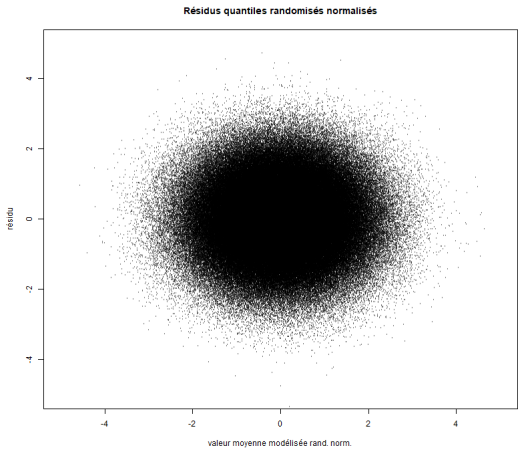

FIGURE I.22 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - BDG

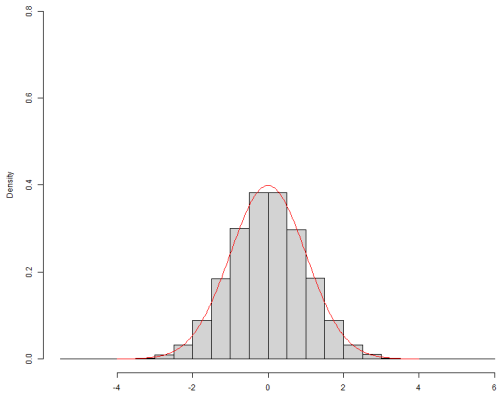

FIGURE I.23 – Distribution empirique des résidus quantiles randomisés normalisés - BN - BDG

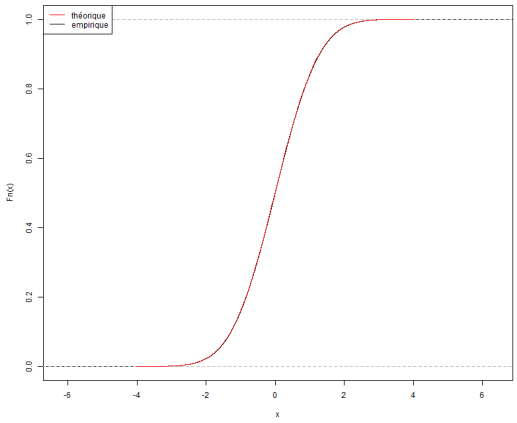

FIGURE I.24 – Fonction de répartition empirique des résidus quantiles randomisées normalisés - BN - BDG

I.3 Garantie Incendie

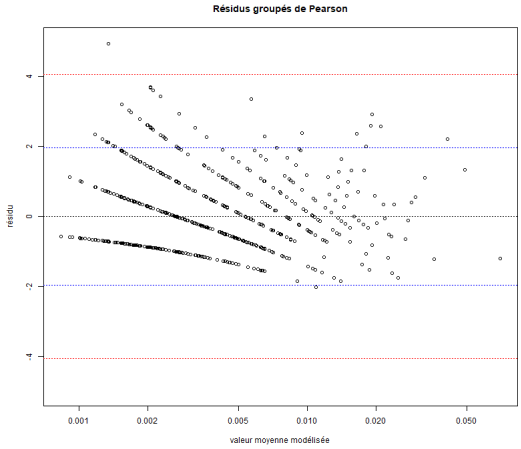

FIGURE I.25 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Incendie

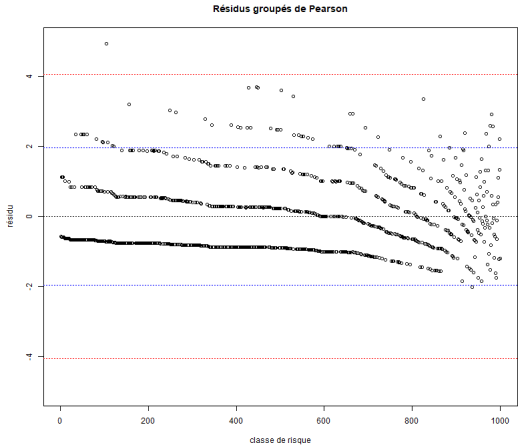

FIGURE I.26 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Incendie

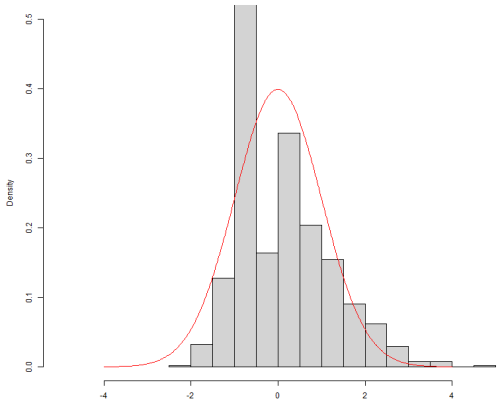

FIGURE I.27 – Distribution empirique des résidus groupés de Pearson - Poisson - Incendie

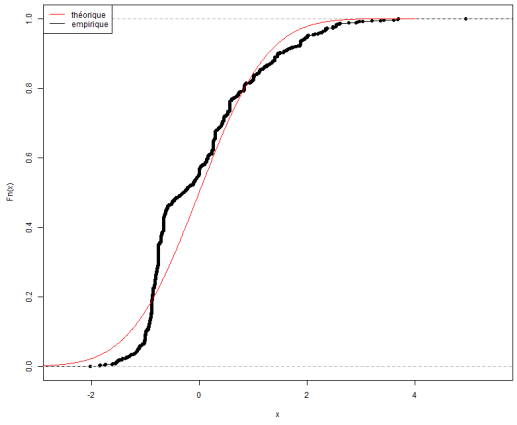

FIGURE I.28 – Fonction de répartition empirique des résidus de Pearson - Poisson - Incendie

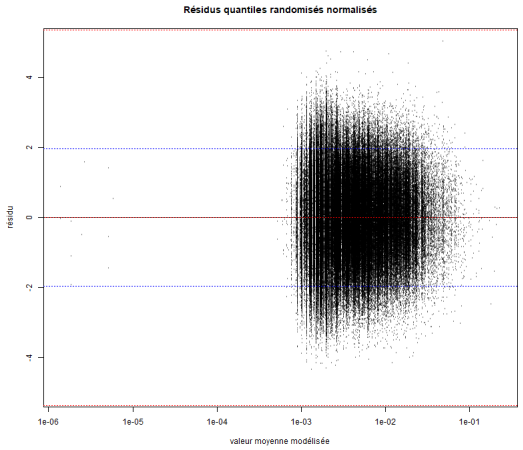

FIGURE I.29 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Incendie

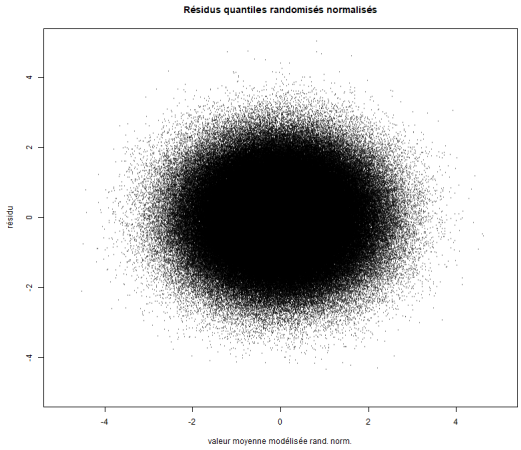

FIGURE I.30 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - Incendie

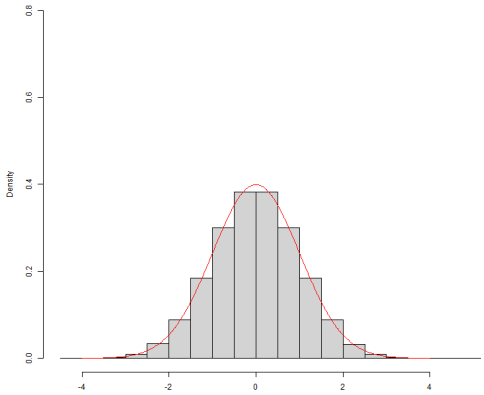

FIGURE I.31 – Distribution empirique des résidus quantiles randomisés normalisés - Poisson - Incendie

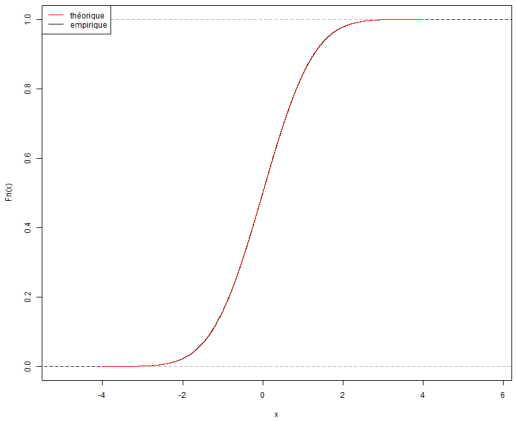

FIGURE I.32 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - Incendie

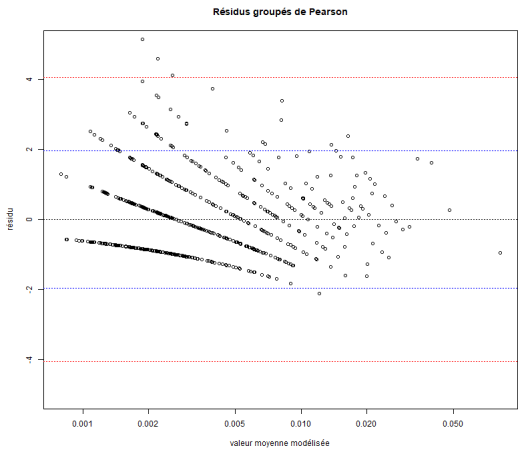

FIGURE I.33 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Incendie

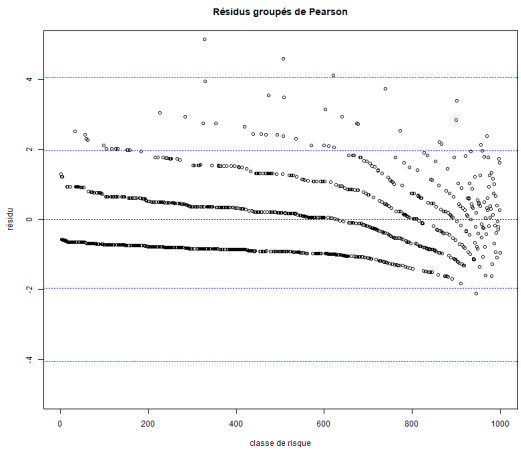

FIGURE I.34 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Incendie

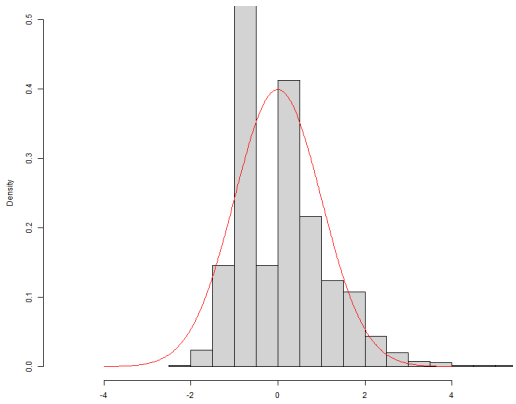

FIGURE I.35 – Distribution empirique des résidus groupés de Pearson - BN - Incendie

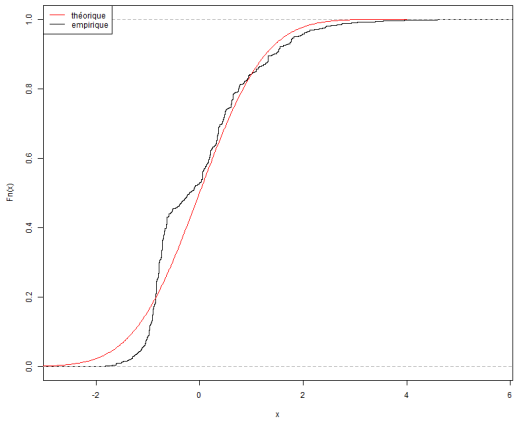

FIGURE I.36 – Fonction de répartition empirique des résidus de Pearson - BN - Incendie

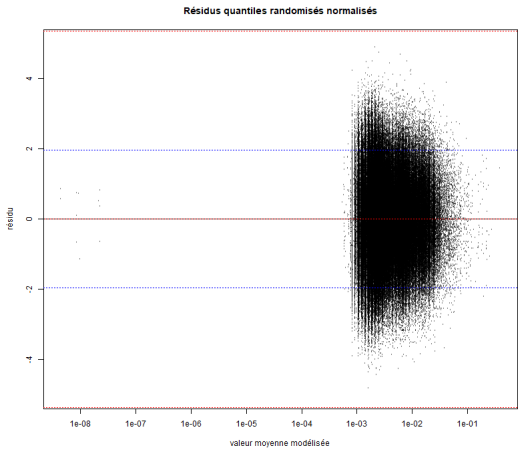

FIGURE I.37 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Incendie

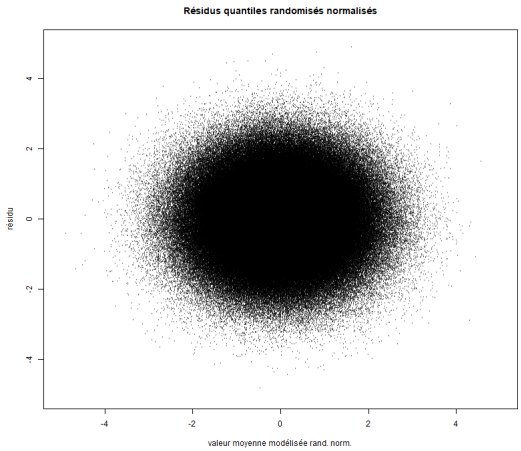

FIGURE I.38 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - Incendie

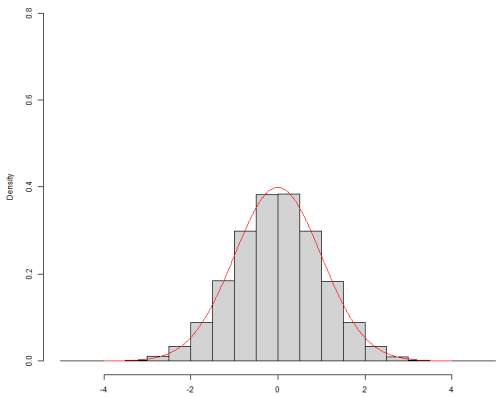

FIGURE I.39 – Distribution empirique des résidus quantiles randomisées normalisés - BN - Incendie

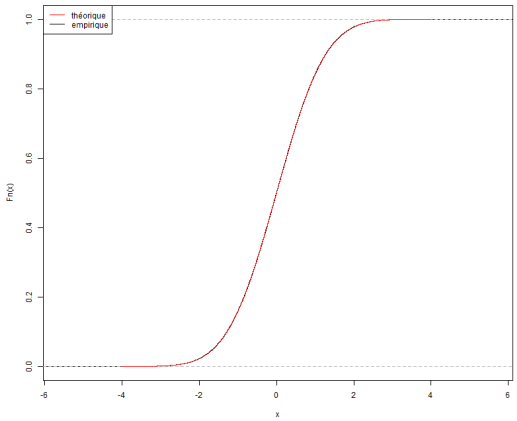

FIGURE I.40 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - Incendie

I.4 Garantie Vol

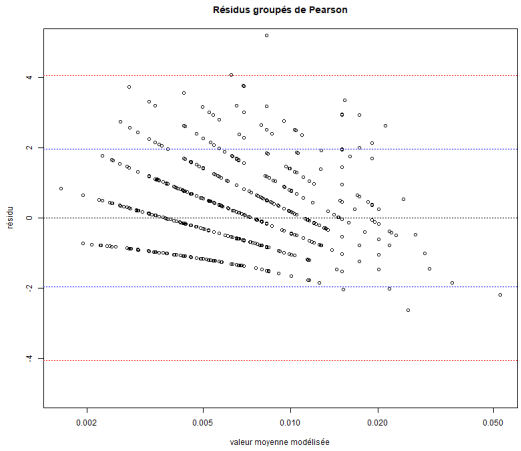

FIGURE I.41 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Vol

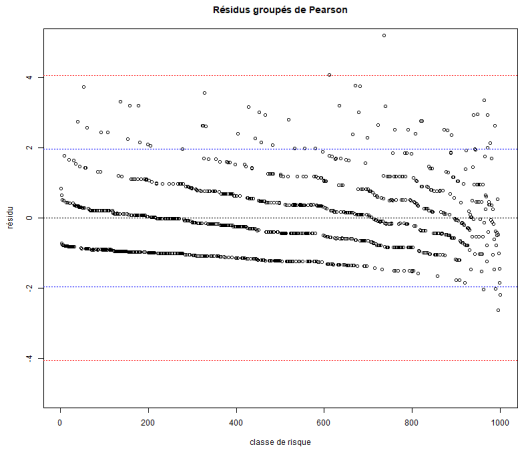

FIGURE I.42 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Vol

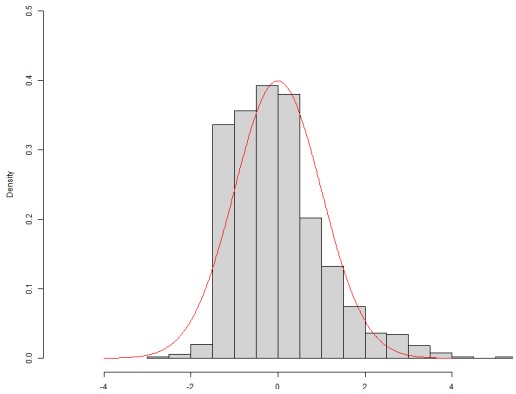

FIGURE I.43 – Distribution empirique des résidus groupés de Pearson - Poisson - Vol

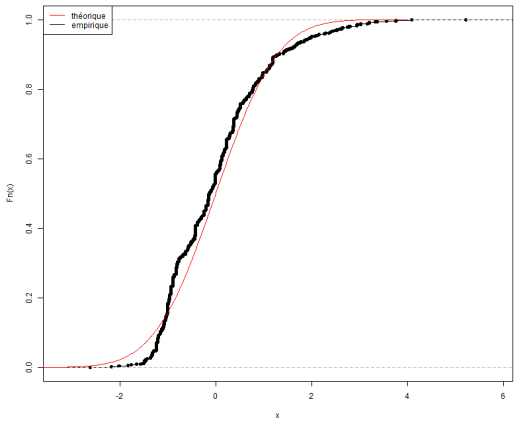

FIGURE I.44 – Fonction de répartition empirique des résidus de Pearson - Poisson - Vol

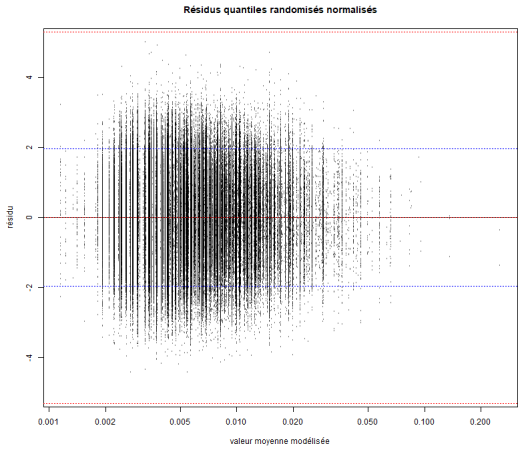

FIGURE I.45 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Vol

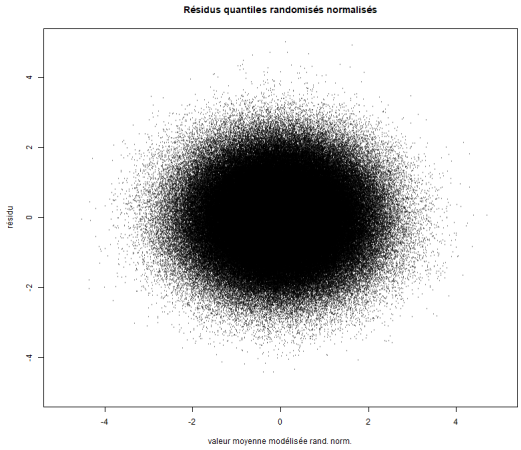

FIGURE I.46 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - Vol

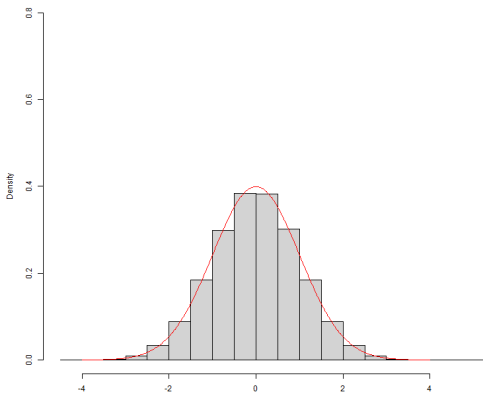

FIGURE I.47 – Distribution empirique des résidus quantiles randomisées normalisés - Poisson - Vol

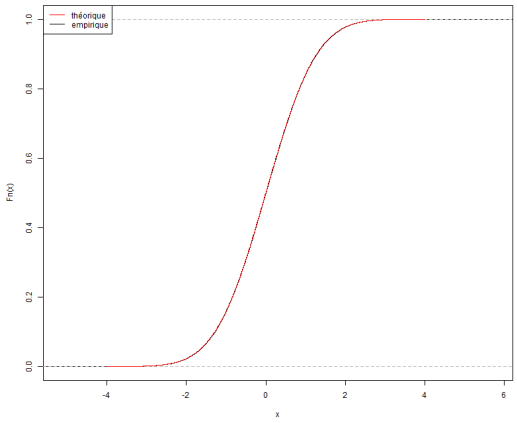

FIGURE I.48 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - Vol

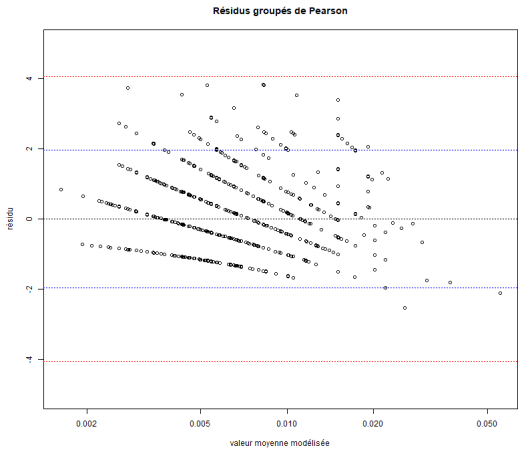

FIGURE I.49 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Vol

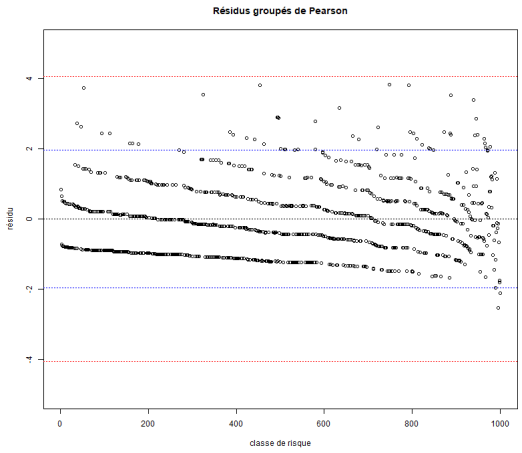

FIGURE I.50 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Vol

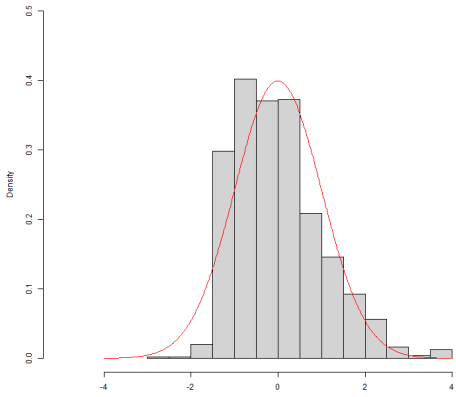

FIGURE I.51 – Distribution empirique des résidus groupés de Pearson - BN - Vol

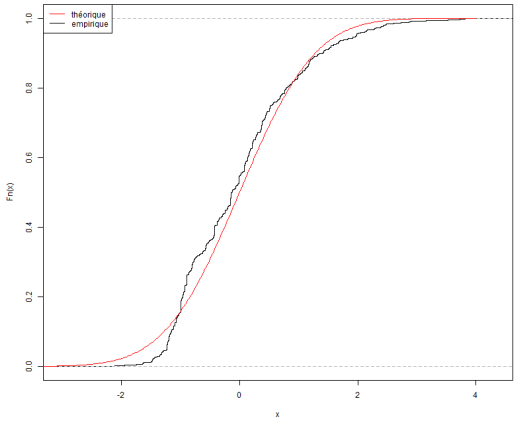

FIGURE I.52 – Fonction de répartition empirique des résidus de Pearson - BN - Vol

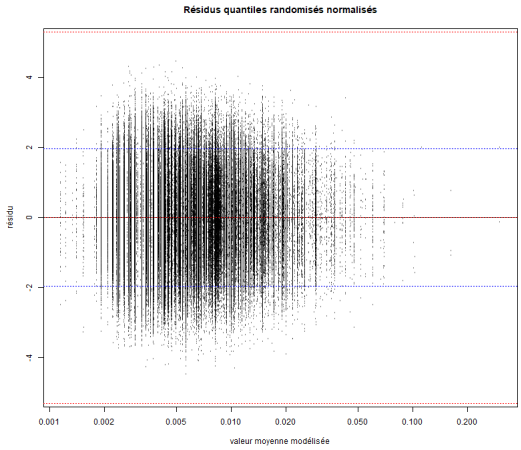

FIGURE I.53 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Vol

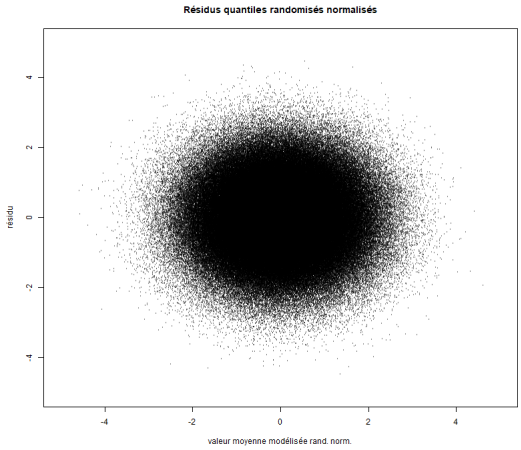

FIGURE I.54 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - Vol

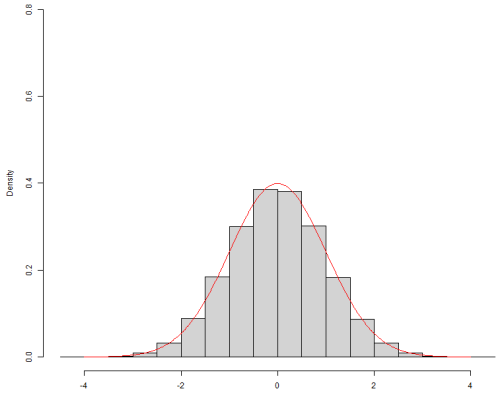

FIGURE I.55 – Distribution empirique des résidus quantiles randomisés normalisés - BN - Vol

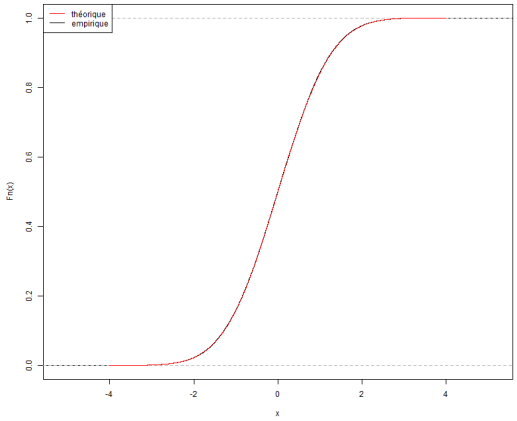

FIGURE I.56 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - Vol

I.5 Garantie RC

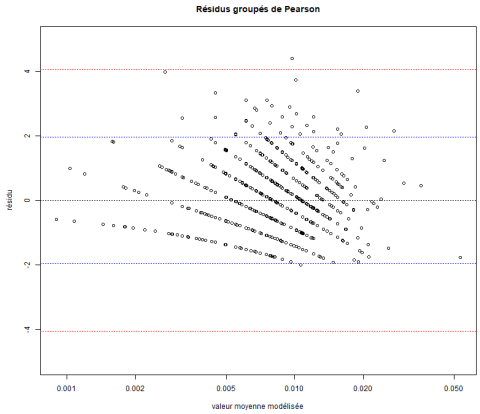

FIGURE I.57 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - RC

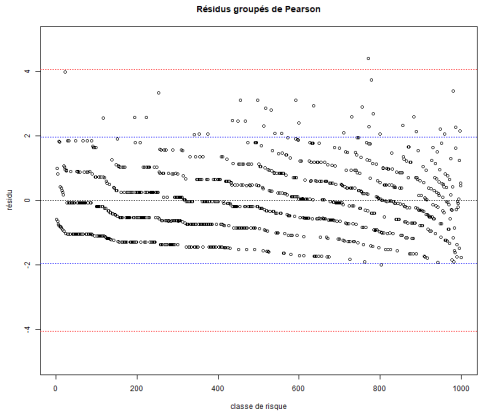

FIGURE I.58 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- Poisson - RC

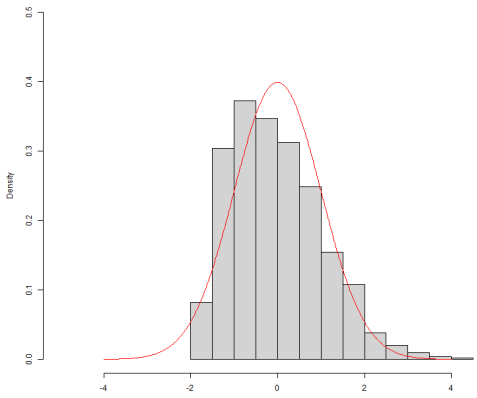

FIGURE I.59 – Distribution empirique des résidus groupés de Pearson - Poisson - RC

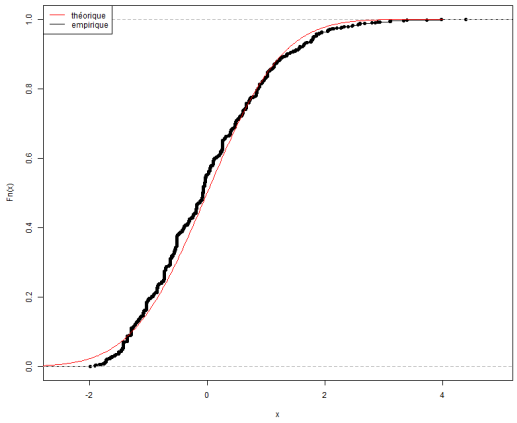

FIGURE I.60 – Fonction de répartition empirique des résidus de Pearson - Poisson - RC

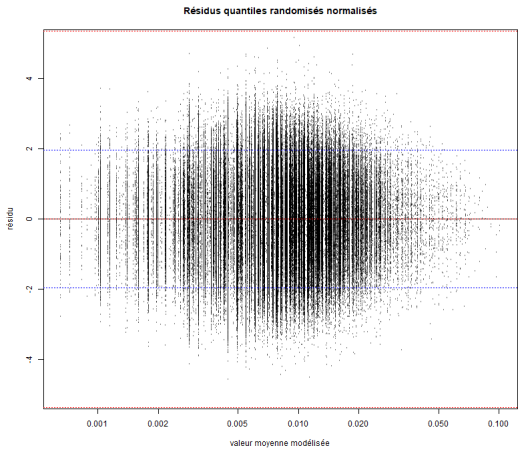

FIGURE I.61 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - RC

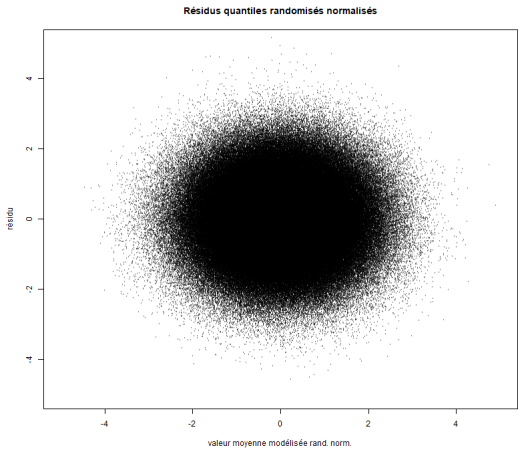

FIGURE I.62 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - RC

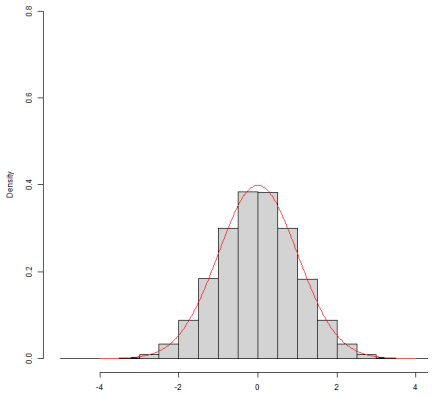

FIGURE I.63 – Distribution empirique des résidus quantiles randomisés normalisés - Poisson - RC

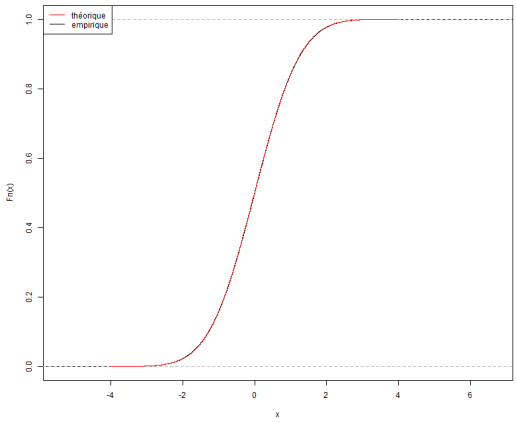

FIGURE I.64 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - RC

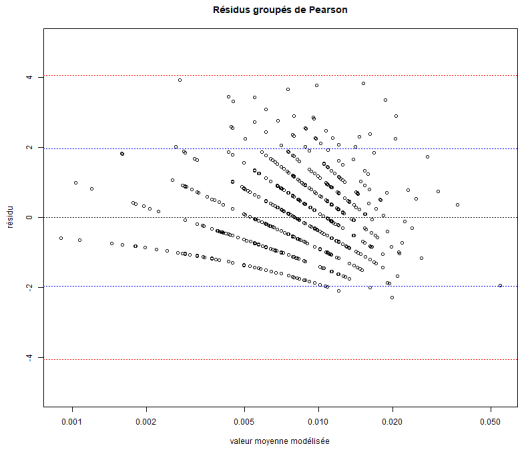

FIGURE I.65 – Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - RC

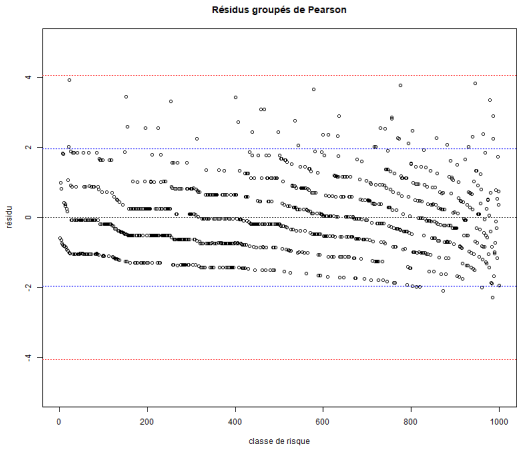

FIGURE I.66 – Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - RC

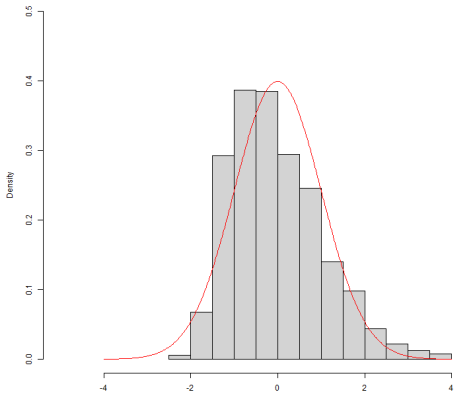

FIGURE I.67 – Distribution empirique des résidus groupés de Pearson - BN - RC

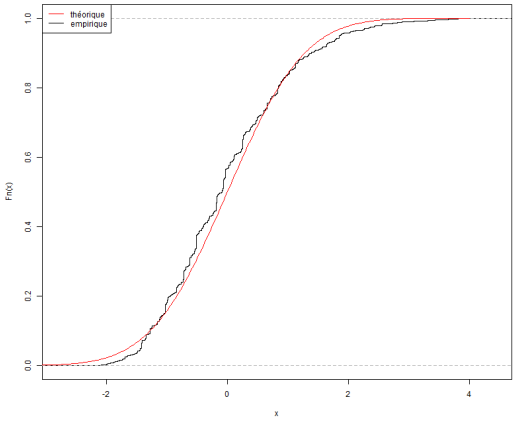

FIGURE I.68 – Fonction de répartition empirique des résidus de Pearson - BN - RC

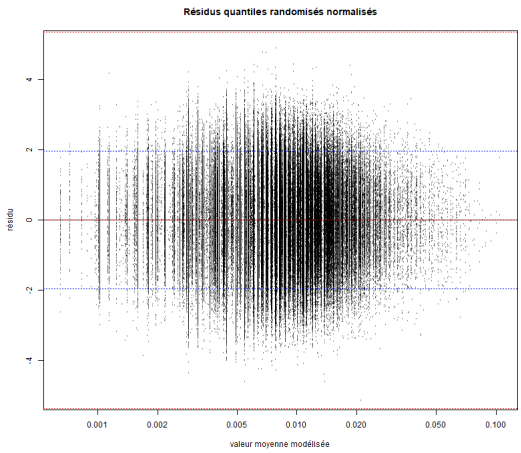

FIGURE I.69 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - RC

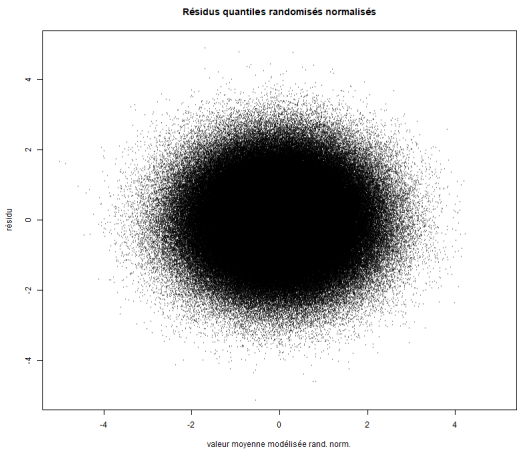

FIGURE I.70 – Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - RC

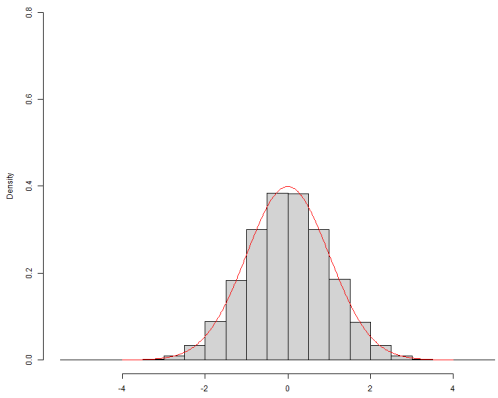

FIGURE I.71 – Distribution empirique des résidus quantiles randomisés normalisés - BN - RC

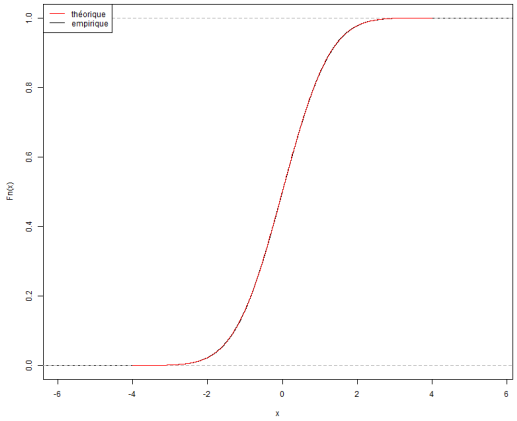

FIGURE I.72 – Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - RC

Annexe J

Eléments graphiques complémentaires pour les modèles CART fréquence

Dans cette partie, sont présentés les différents arbres finaux pour les garanties (à l'exception de la garantie DDE).

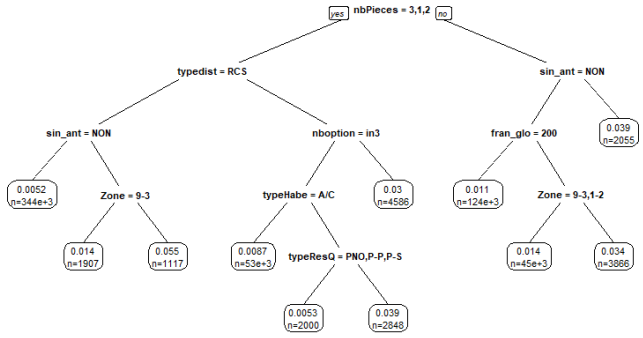

FIGURE J.1 – Arbre final pour la garantie RC

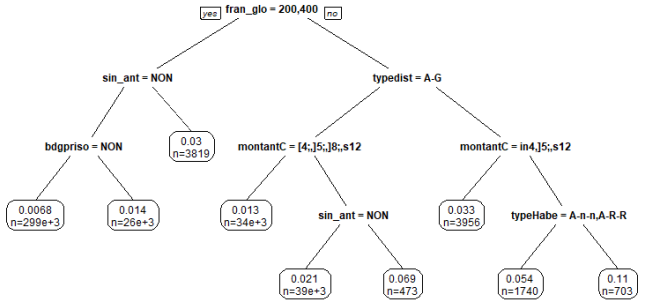

FIGURE J.2 – Arbre final pour la garantie BDG

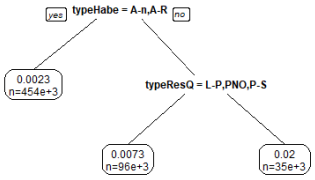

FIGURE J.3 – Arbre final pour la garantie Incendie

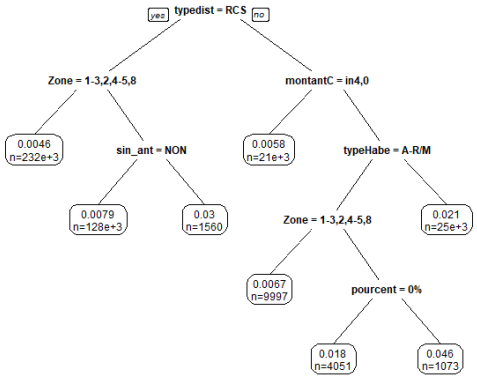

FIGURE J.4 – Arbre final pour la garantie Vol

Annexe K

Eléments Graphiques *Random Forest* fréquence

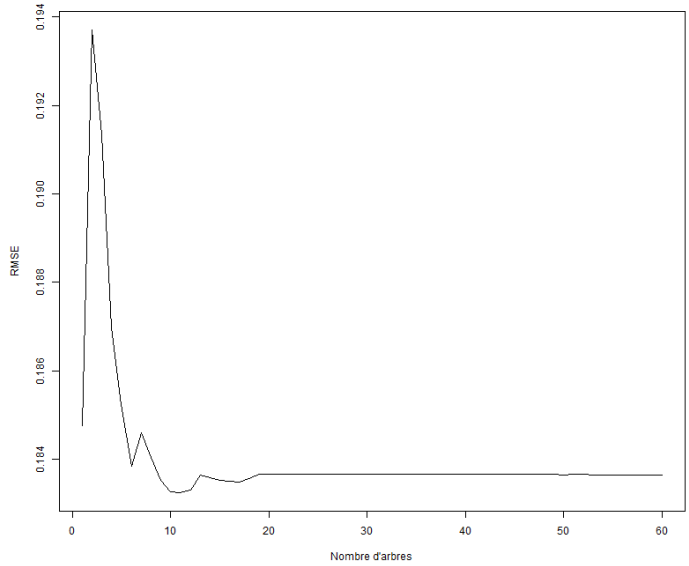

FIGURE K.1 – Evolution RMSE - RF - Fréquence - BDG

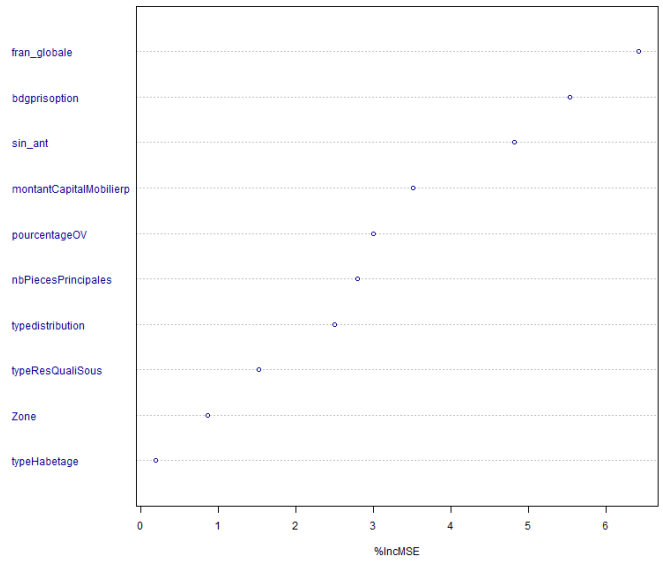

FIGURE K.2 – Importance variables - RF - Fréquence - BDG

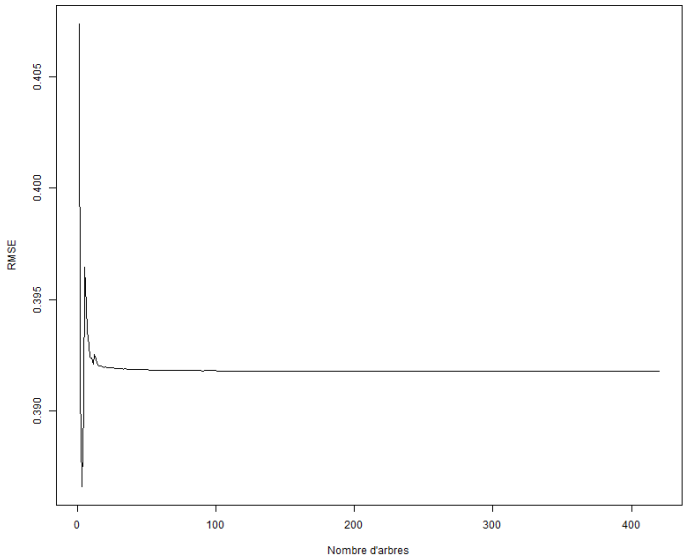

FIGURE K.3 – Evolution RMSE - RF - Fréquence - DDE

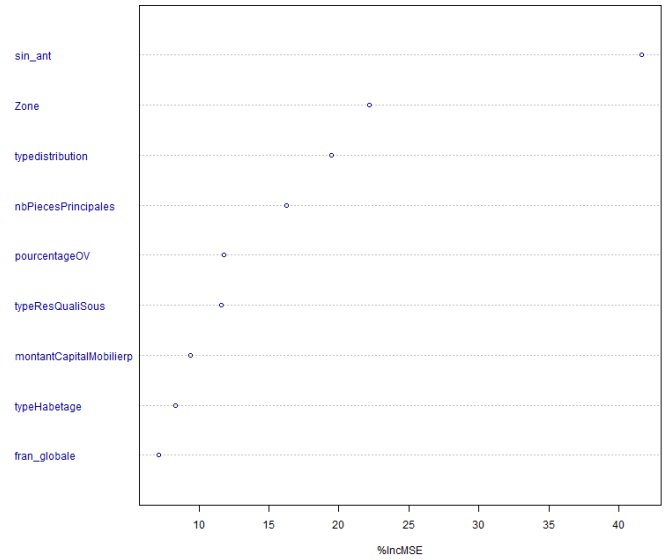

FIGURE K.4 – Importance variables - RF - Fréquence - DDE

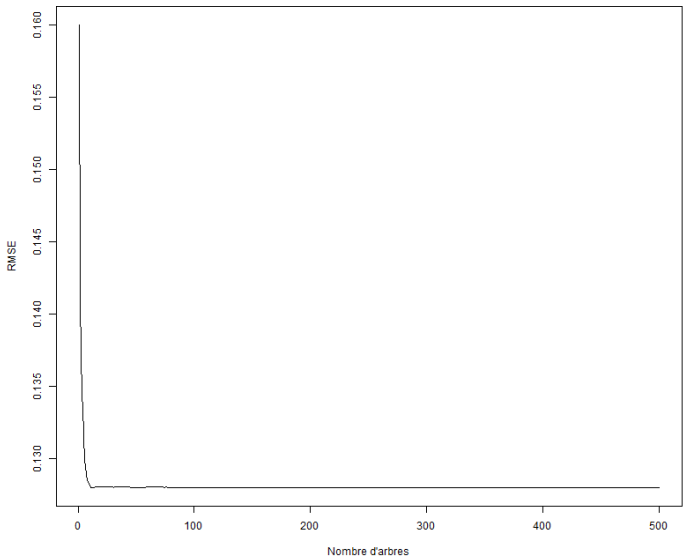

FIGURE K.5 – Evolution RMSE - RF - Fréquence - Incendie

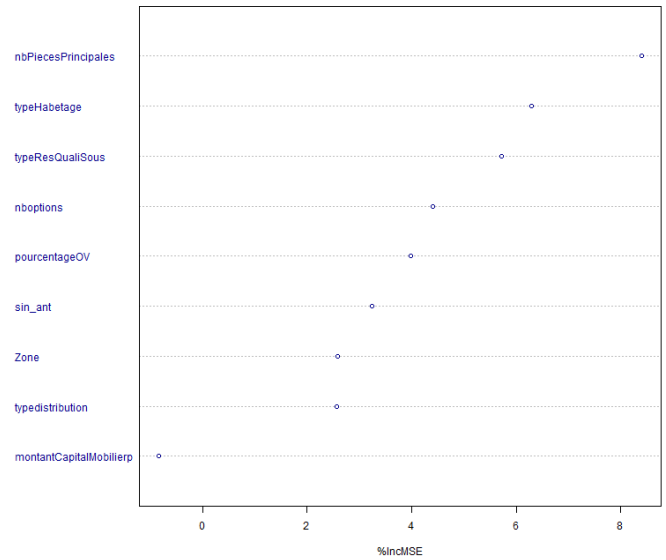

FIGURE K.6 – Importance variables - RF - Fréquence - Incendie

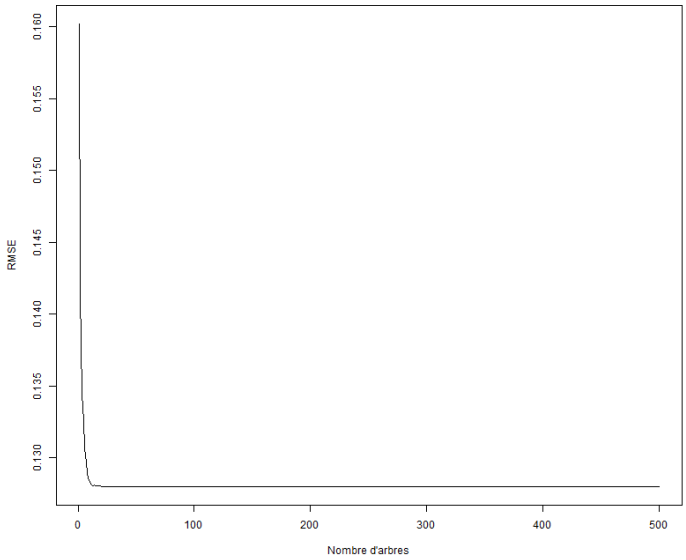

FIGURE K.7 – Evolution RMSE - RF - Fréquence - Incendie (Sans la variable avec une importance négative)

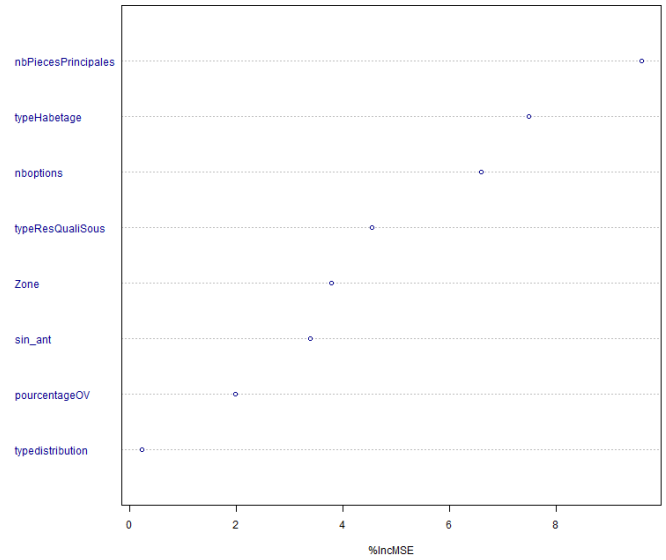

FIGURE K.8 – Importance variables - RF - Fréquence - Incendie (Sans la variable avec une importance négative)

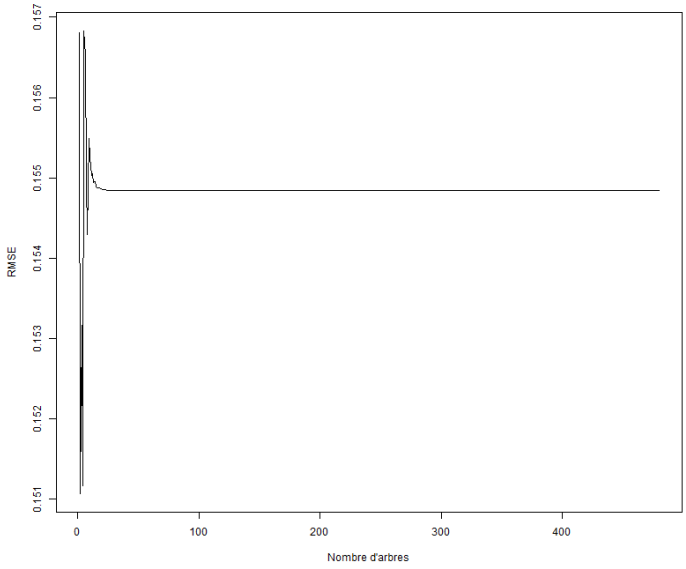

FIGURE K.9 – Evolution RMSE - RF - Fréquence - Vol

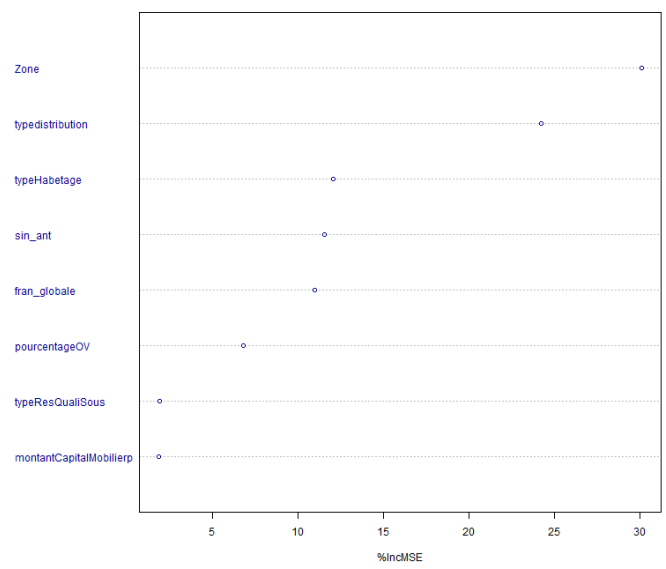

FIGURE K.10 – Importance variables - RF - Fréquence - Vol

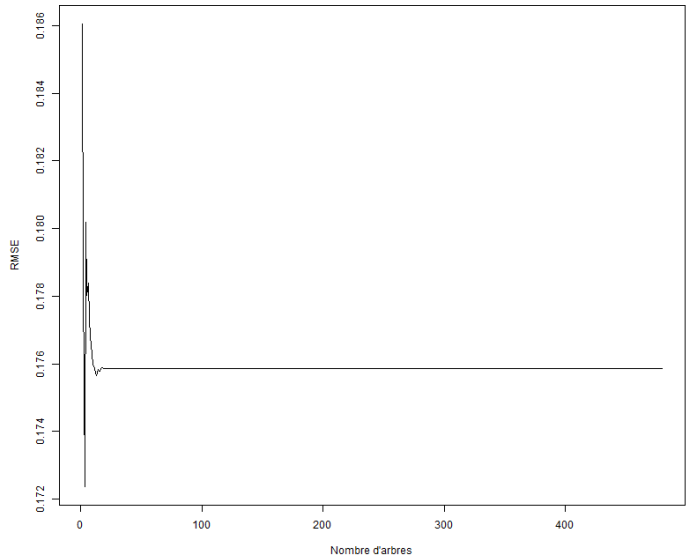

FIGURE K.11 – Evolution RMSE - RF - Fréquence - RC

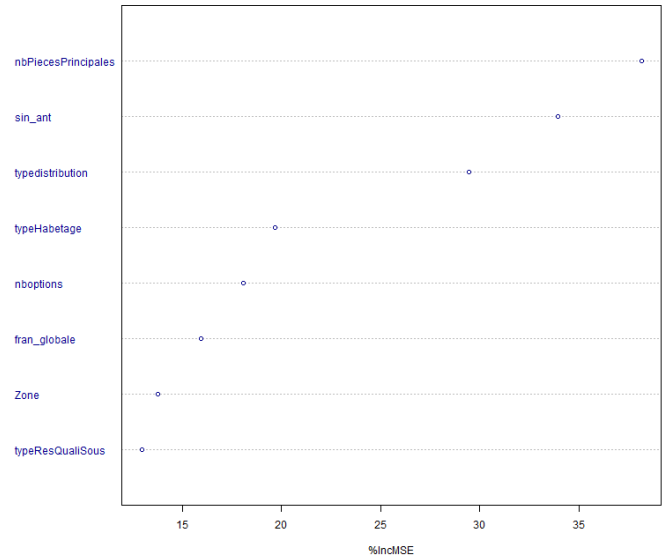

FIGURE K.12 – Importance variables - RF - Fréquence - RC

Annexe L

Eléments Graphiques *XGBoost* fréquence

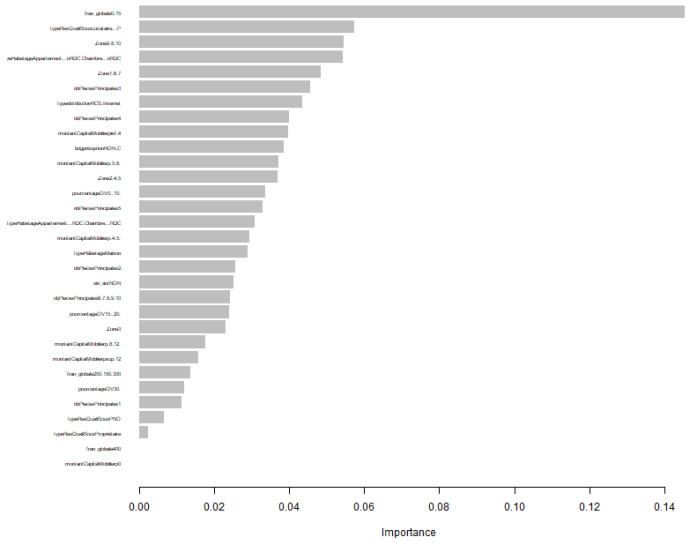

FIGURE L.1 – Importance variables - XGBoost - Fréquence - BDG

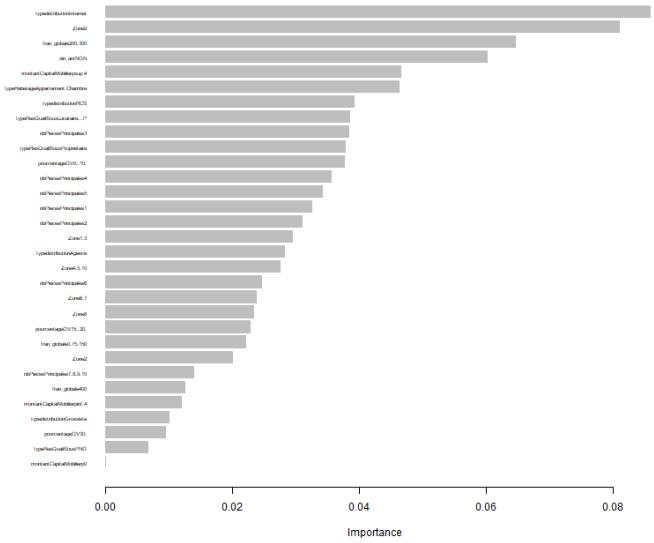

FIGURE L.2 – Importance variables - XGBoost - Fréquence - DDE

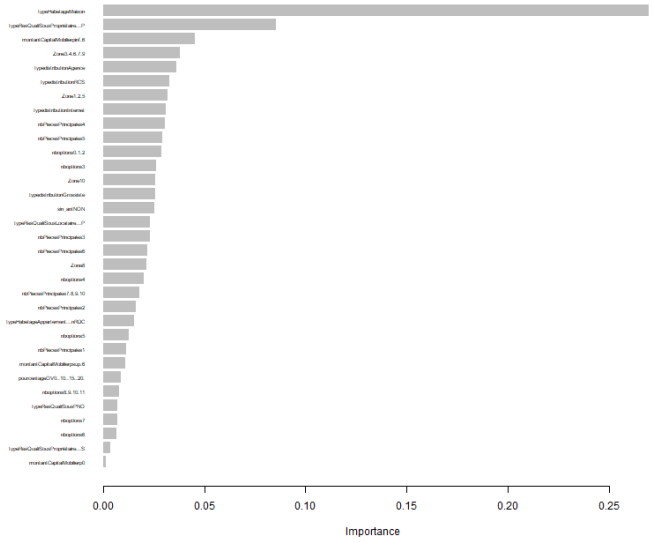

FIGURE L.3 – Importance variables - XGBoost - Fréquence - Incendie

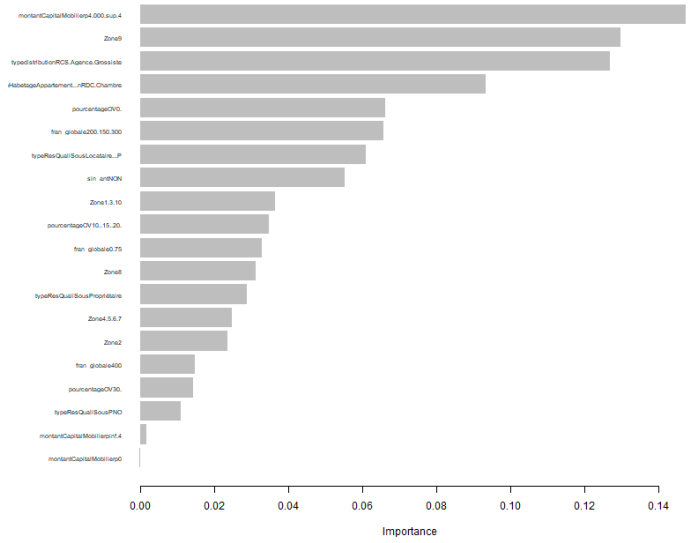

FIGURE L.4 – Importance variables - XGBoost - Fréquence - Vol

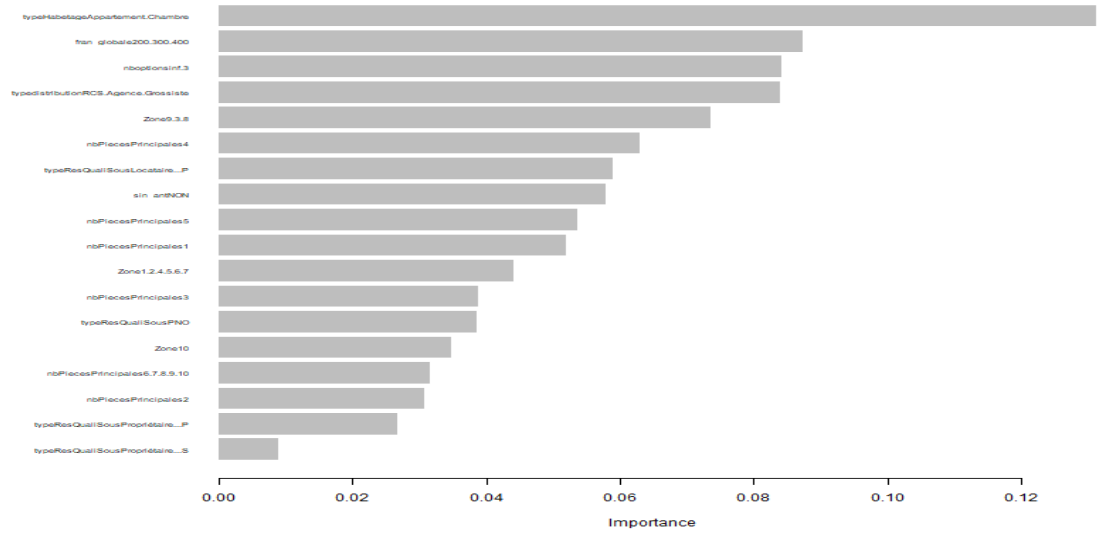

FIGURE L.5 – Importance variables - XGBoost - Fréquence - RC

Annexe M

Tableaux complémentaires pour les GLM coût

M.1 Tableaux pour adéquation des lois avant modélisation

	Gamma	Log-normale
Log-vraisemblance	-17 217.72	-17 252.06
AIC	34 439.45	34 508.11
Cramer-von-Mises	0.77192	0.57593

TABLE M.1 – Tableau des indicateurs avant modélisation pour la garantie BDG

	Gamma	Log-normale
Log-vraisemblance	-118 101.2	-117 556.9
AIC	236 206.3	235 117.7
Cramer-von-Mises	0.7845	1.0249

TABLE M.2 – Tableau des indicateurs avant modélisation pour la garantie DDE

	Gamma	Log-normale
Log-vraisemblance	-13 584.24	-13 515.23
AIC	27 172.49	27 034.45
Cramer-von-Mises	1.0104	0.68035

TABLE M.3 – Tableau des indicateurs avant modélisation pour la garantie Incendie

	Gamma	Log-normale
Log-vraisemblance	-14 864.52	-14 778.94
AIC	29 733.05	29 561.87
Cramer-von-Mises	0.88092	0.51662

TABLE M.4 – Tableau des indicateurs avant modélisation pour la garantie Vol

0. Pour toutes les garanties, les deux tests de Cramer-Von-Mises permettent de conclure à l'adéquation des deux types de lois.

	Gamma	Log-normale
Log-vraisemblance	-20 826.47	-20 480.91
AIC	41 656.94	40 965.82
Cramer-von-Mises	1.0278	0.64647

TABLE M.5 – Tableau des indicateurs avant modélisation pour la garantie RC

	Gamma	Log-normale
Log-vraisemblance	-187 974.7	-185 834.3
AIC	375 953.4	371 672.7
Cramer-von-Mises	1.1099	0.92628

TABLE M.6 – Tableau des indicateurs avant modélisation toutes garanties

M.2 Significativité des variables sélectionnées après regroupement

Variables	Df	Statistique	p-value
nbPiecesPrincipales	4	20.158	0.0004647
Zone	1	18.299	$1.889e-05$
typeHabetage	1	17.751	$2.518e-05$
typedistribution	1	5.623	0.0177251
sin_ant	1	19.245	$1.150e-05$
fran_globale	1	13.069	0.0003001
pourcentageOV	2	15.543	0.0004217
typeResQualiSous	3	299.56	$< 2.2e-16$

TABLE M.7 – Analyse de type III pour la loi Gamma et la garantie DDE

Variables	Df	Statistique	p-value
Zone	2	40.150	$1.913e-09$
typedistribution	1	9.217	0.002398
sin_ant	1	15.168	$9.837e-05$
fran_globale	2	26.997	$1.373e-06$
pourcentageOV	1	10.525	0.001178
typeResQualiSous	1	27.204	$1.831e-07$

TABLE M.8 – Analyse de type III pour la loi Gamma et la garantie BDG

Variables	Df	Statistique	p-value
nbPiecesPrincipales	1	5.603	0.0179353
Zone	2	32.018	$1.115e-07$
typeResQualiSous	1	14.845	0.0001167
typedistribution	2	27.372	$1.139e-06$
sin_ant	1	24.006	$9.604e-07$
fran_globale	1	73.894	$< 2.2e-16$
pourcentageOV	1	9.254	0.0023496

TABLE M.9 – Analyse de type III pour la loi de Log-normale et la garantie BDG

Variables	Df	Statistique	p-value
fran_globale	1	7.920	0.004888
typedistribution	1	4.260	0.039018
typeResQualiSous	1	5.480	0.019231

TABLE M.10 – Analyse de type III pour la loi Gamma et la garantie Incendie

Variables	Df	Statistique	p-value
typedistribution	1	3.316	0.06863
fran_globale	1	2.775	0.09573
pourcentageOV	1	4.457	0.03476

TABLE M.11 – Analyse de type III pour la loi de Log-normale et la garantie Incendie

Variables	Df	Statistique	p-value
nbPiecesPrincipales	2	16.216	0.0003012
pourcentageOV	1	50.409	$1.248e{-}12$
typeResQualiSous	1	7.359	0.0066727

TABLE M.12 – Analyse de type III pour la loi Gamma et la garantie Vol

Variables	Df	Statistique	p-value
typeHabetage	1	14.551	0.0001364
typeResQualiSous	3	34.608	$1.474e{-}07$
pourcentageOV	2	22.074	$1.610e{-}05$

TABLE M.13 – Analyse de type III pour la loi de Log-normale et la garantie Vol

Variables	Df	Statistique	p-value
Zone	1	7.885	0.004984
fran_globale	1	79.329	$< 2.2e-16$
typeResQualiSous	2	12.553	0.001880

TABLE M.14 – Analyse de type III pour la loi Gamma et la garantie RC

Variables	Df	Statistique	p-value
typeHabetage	1	4.277	0.03863
sin_ant	1	6.016	0.01418
fran_globale	2	149.460	$< 2.2e-16$
typeResQualiSous	2	21.785	$1.86e-05$

TABLE M.15 – Analyse de type III pour la loi de Log-normale et la garantie RC

Annexe N

Illustrations complémentaires pour les GLM coût

N.1 Garantie BDG

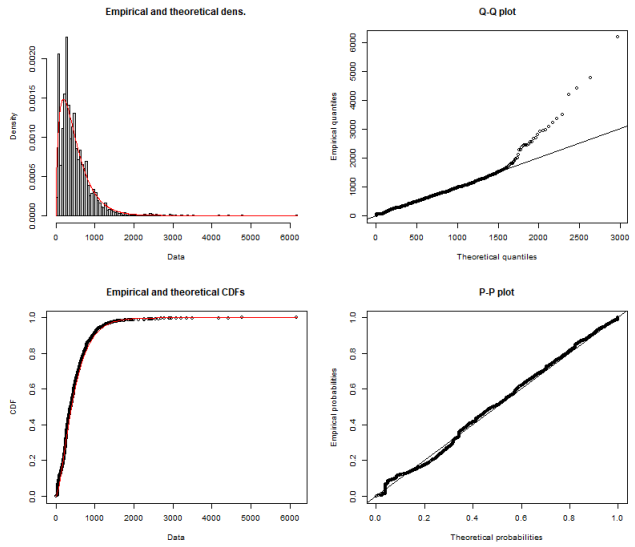

FIGURE N.1 – Analyse de la distribution du coût BDG - loi Gamma

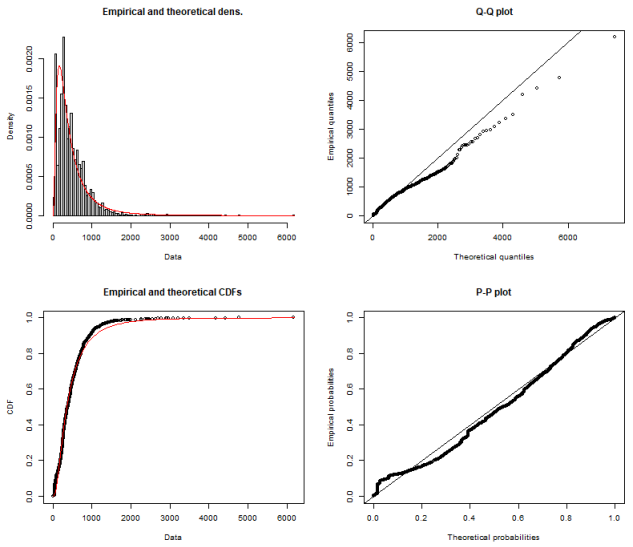

FIGURE N.2 – Analyse de la distribution du coût BDG - loi Log-normale

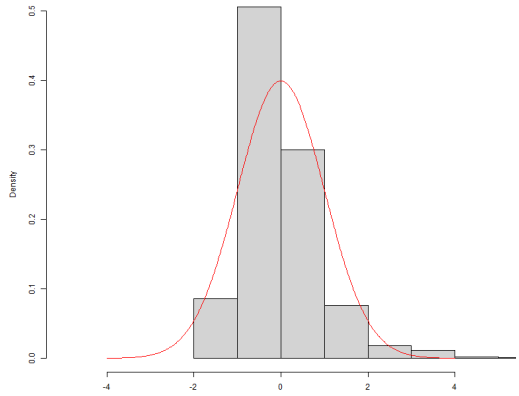

FIGURE N.3 – Histogramme des Résidus de Pearson - Gamma - BDG

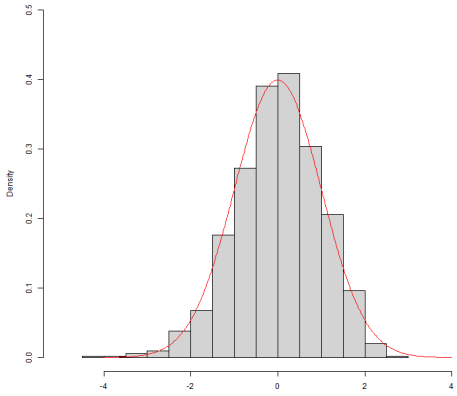

FIGURE N.4 – Histogramme des Résidus de Pearson - Log-normale - BDG

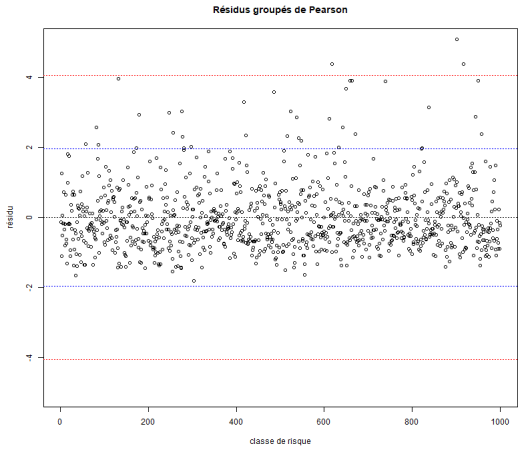

FIGURE N.5 – Résidus de Pearson en fonction de la classe de risque - Gamma - BDG

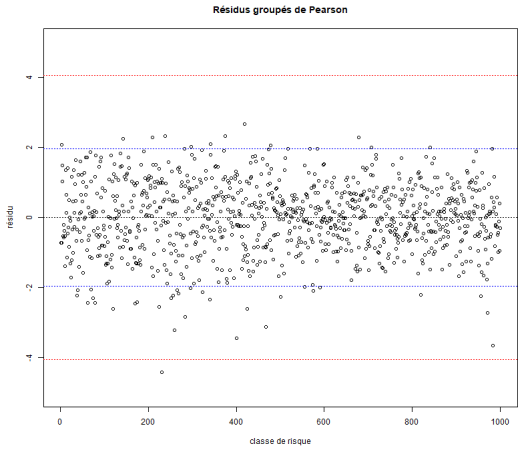

FIGURE N.6 – Résidus de Pearson en fonction de la classe de risque - Log-normale - BDG

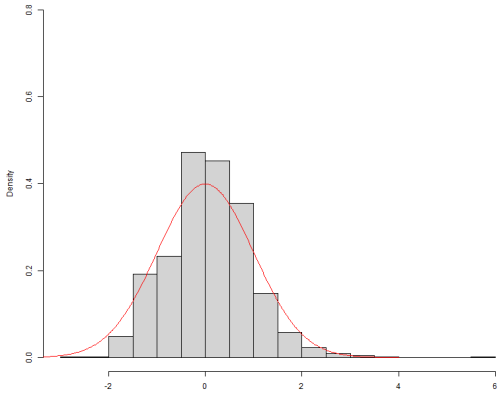

FIGURE N.7 – Histogramme des Résidus Quantiles - Gamma - BDG

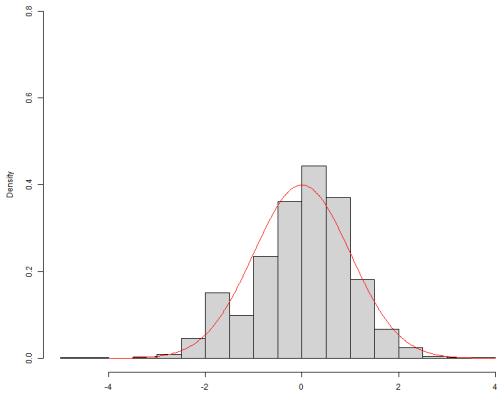

FIGURE N.8 – Histogramme des Résidus Quantiles - Log-normale - BDG

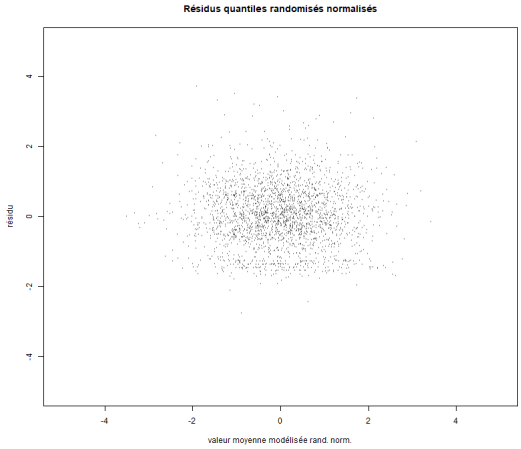

FIGURE N.9 – Résidus Quantiles - Gamma - BDG

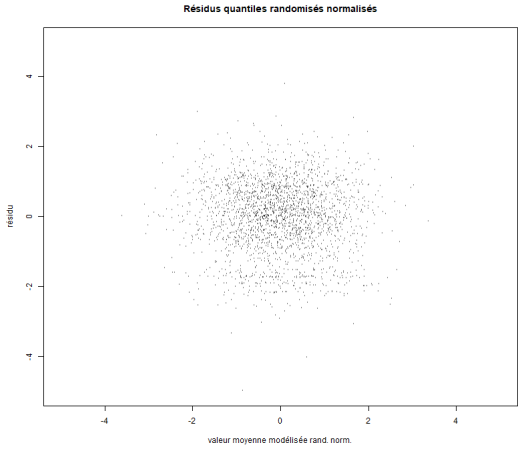

FIGURE N.10 – Résidus Quantiles - Log-normale - BDG

N.2 Garantie Incendie

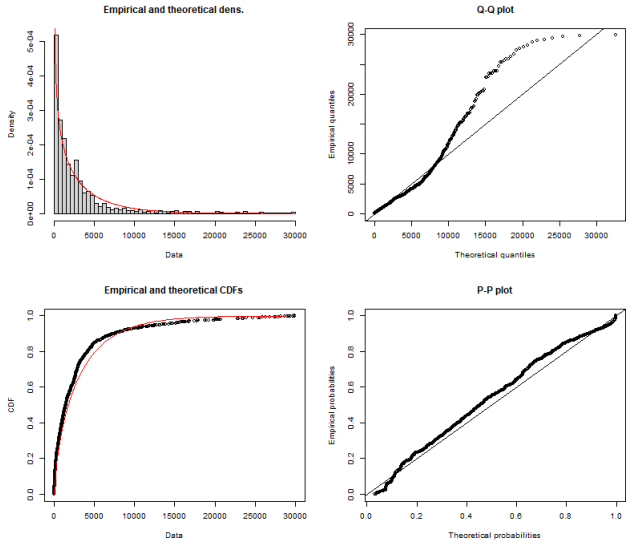

FIGURE N.11 – Analyse de la distribution du coût Incendie - loi Gamma

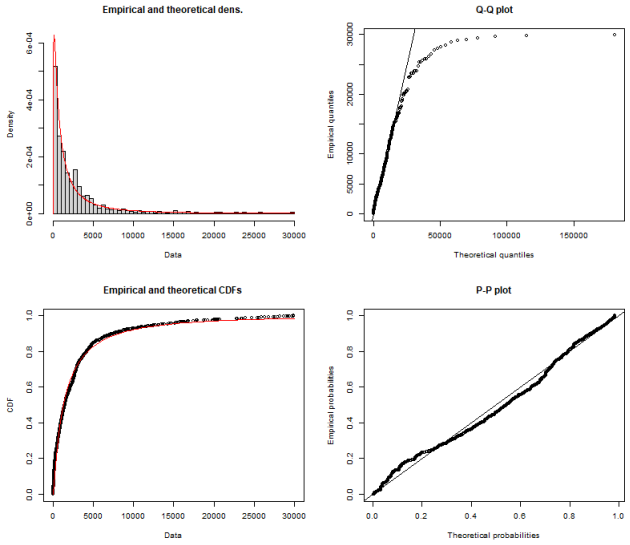

FIGURE N.12 – Analyse de la distribution du coût Incendie - loi Log-normale

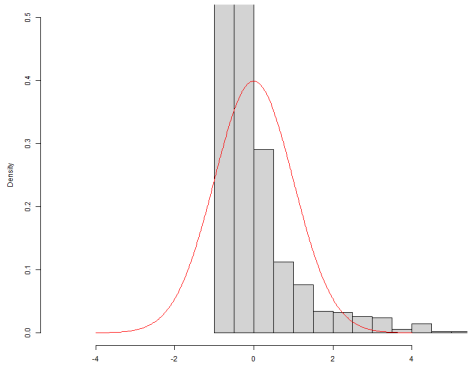

FIGURE N.13 – Histogramme des Résidus de Pearson - Gamma - Incendie

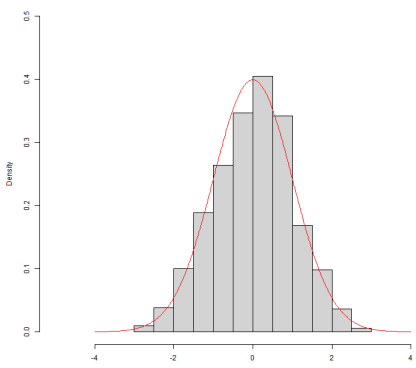

FIGURE N.14 – Histogramme des Résidus de Pearson - Log-normale - Incendie

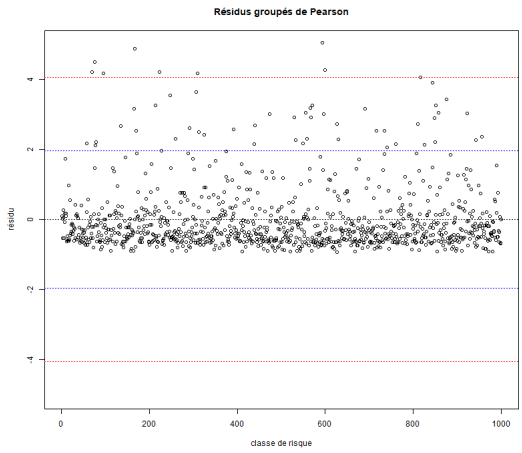

FIGURE N.15 – Résidus de Pearson en fonction de la classe de risque - Gamma - Incendie

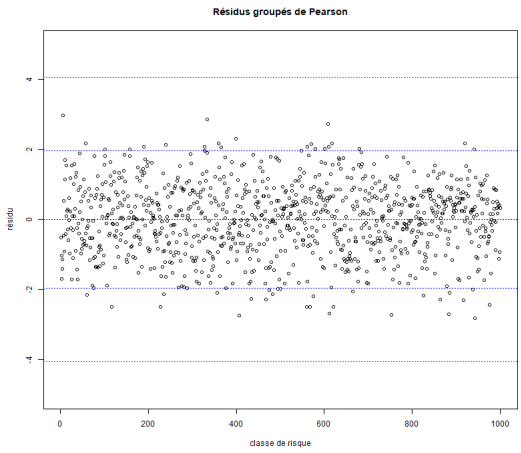

FIGURE N.16 – Résidus de Pearson en fonction de la classe de risque - Log-normale - Incendie

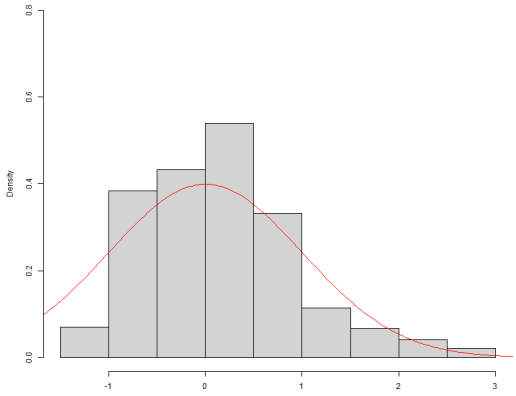

FIGURE N.17 – Histogramme des Résidus Quantiles - Gamma - Incendie

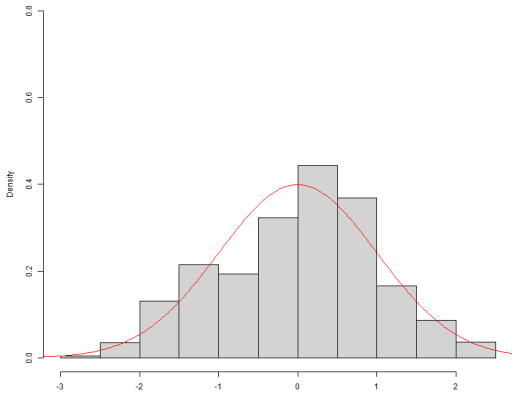

FIGURE N.18 – Histogramme des Résidus Quantiles - Log-normale - Incendie

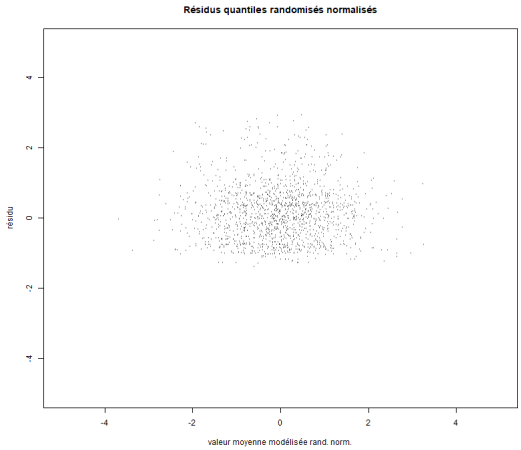

FIGURE N.19 – Résidus Quantiles - Gamma - Incendie

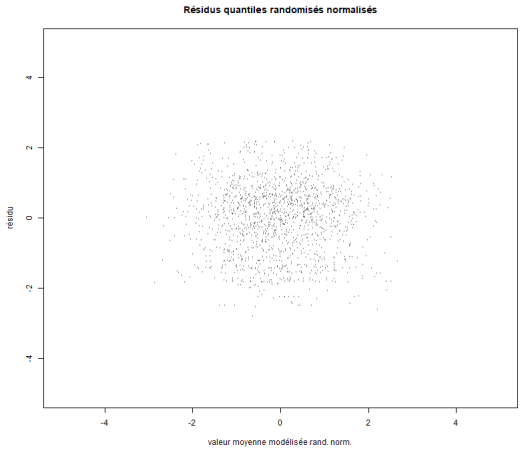

FIGURE N.20 – Résidus Quantiles - Log-normale - Incendie

N.3 Garantie RC

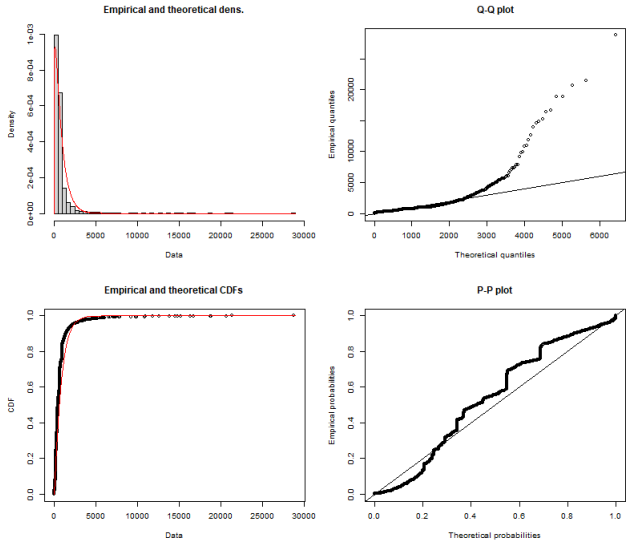

FIGURE N.21 – Analyse de la distribution du coût RC - loi Gamma

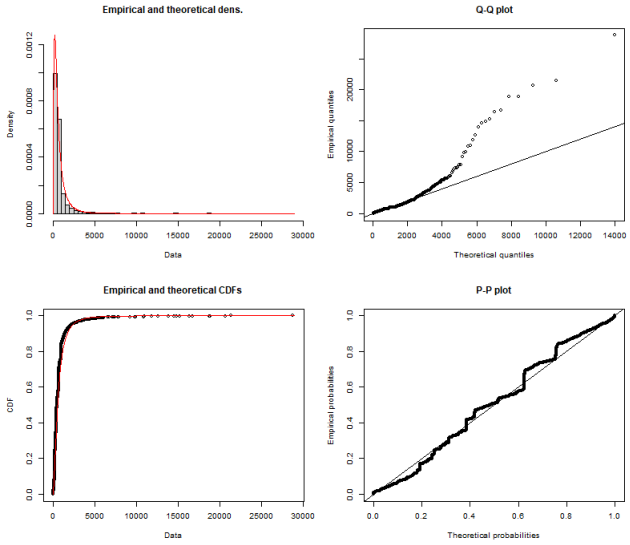

FIGURE N.22 – Analyse de la distribution du coût RC - loi Log-normale

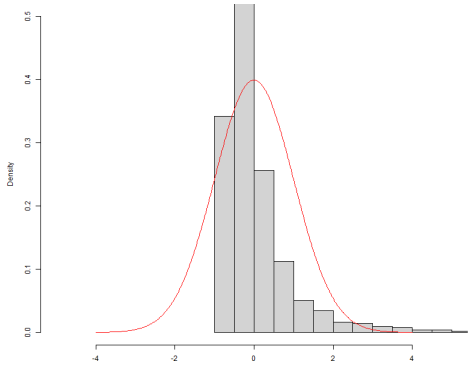

FIGURE N.23 – Histogramme des Résidus de Pearson - Gamma - RC

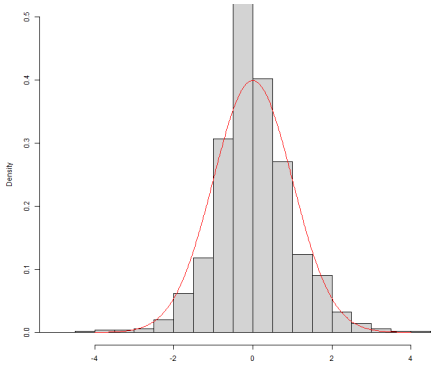

FIGURE N.24 – Histogramme des Résidus de Pearson - Log-normale - RC

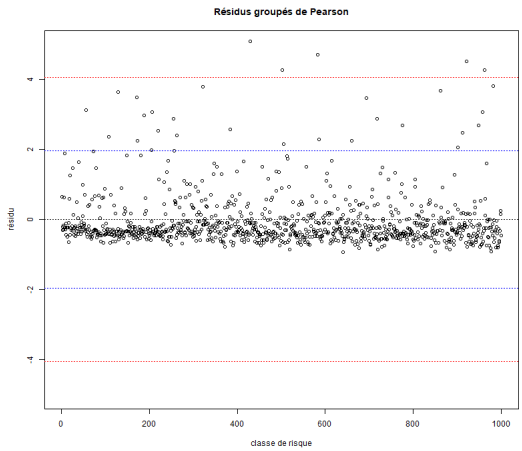

FIGURE N.25 – Résidus de Pearson en fonction de la classe de risque - Gamma - RC

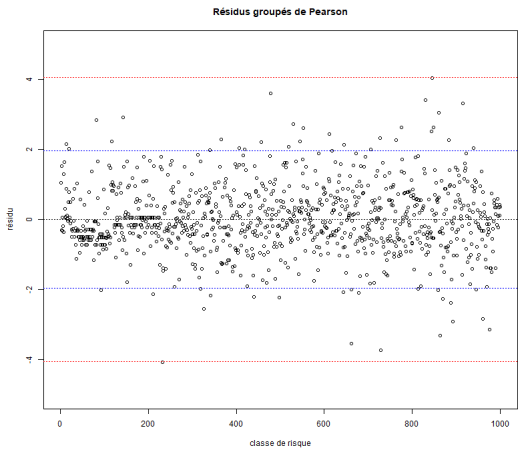

FIGURE N.26 – Résidus de Pearson en fonction de la classe de risque - Log-normale - RC

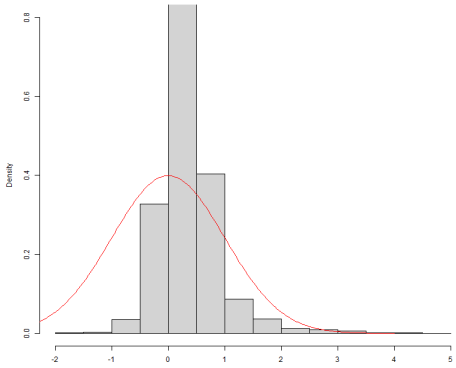

FIGURE N.27 – Histogramme des Résidus Quantiles - Gamma - RC

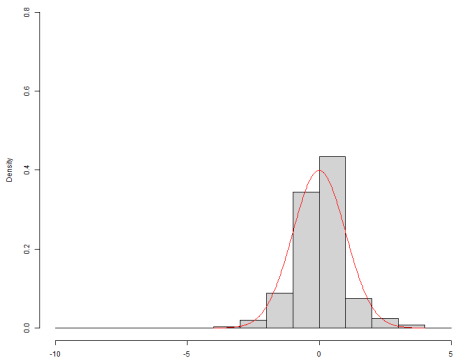

FIGURE N.28 – Histogramme des Résidus Quantiles - Log-normale - RC

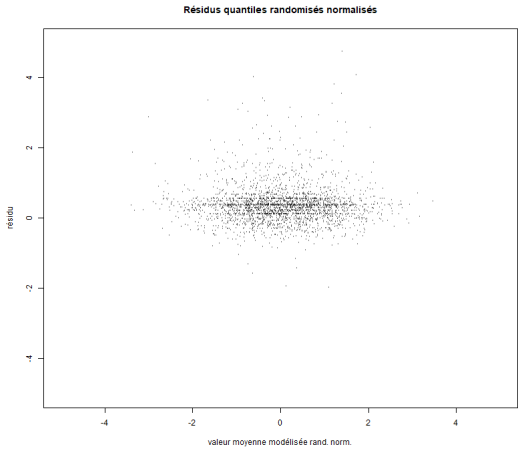

FIGURE N.29 – Résidus Quantiles - Gamma - RC

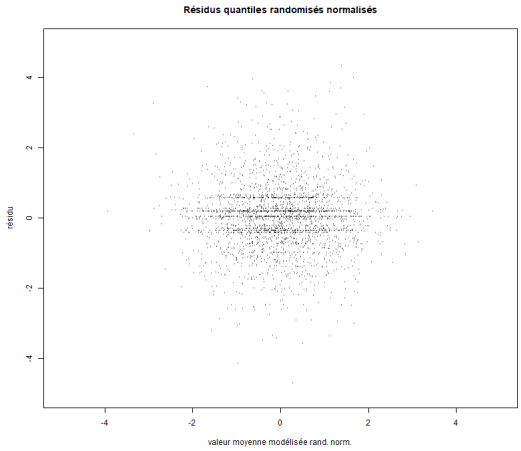

FIGURE N.30 – Résidus Quantiles - Log-normale - RC

N.4 Garantie Vol

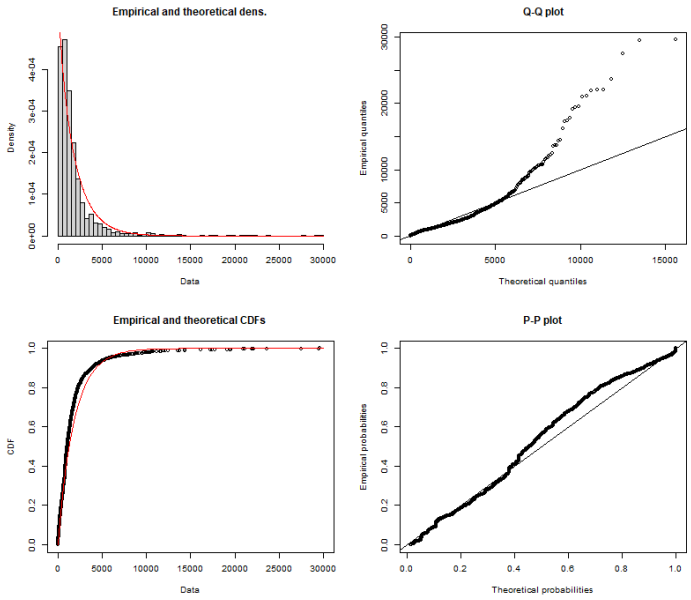

FIGURE N.31 – Analyse de la distribution du coût Vol - loi Gamma

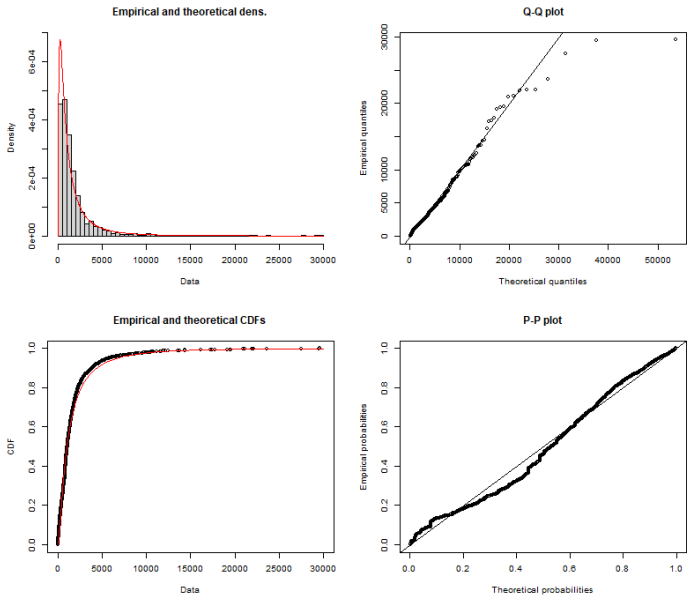

FIGURE N.32 – Analyse de la distribution du coût Vol - loi Log-normale

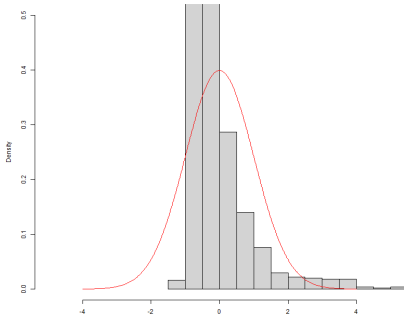

FIGURE N.33 – Histogramme des Résidus de Pearson - Gamma - Vol

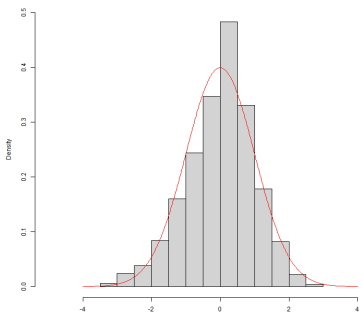

FIGURE N.34 – Histogramme des Résidus de Pearson - Log-normale - Vol

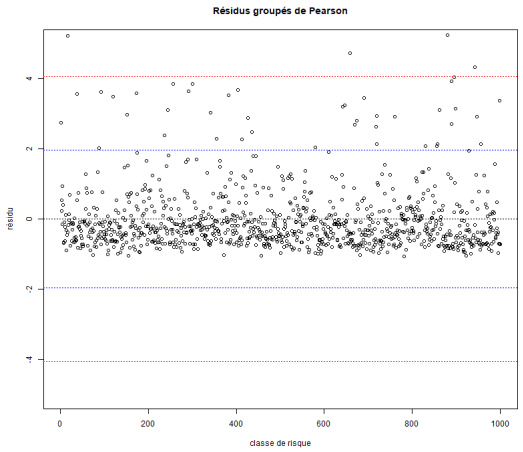

FIGURE N.35 – Résidus de Pearson en fonction de la classe de risque - Gamma - Vol

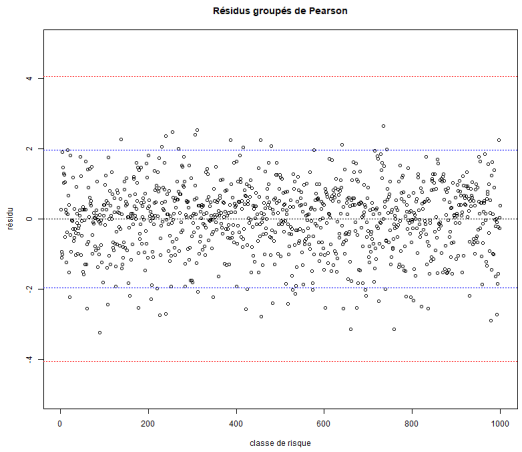

FIGURE N.36 – Résidus de Pearson en fonction de la classe de risque - Log-normale - Vol

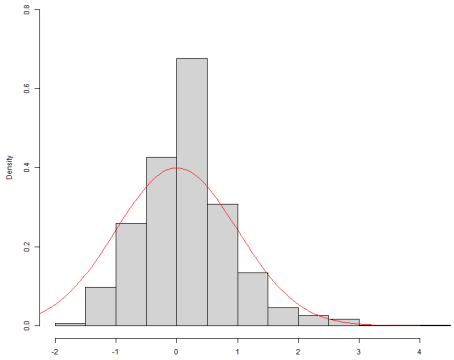

FIGURE N.37 – Histogramme des Résidus Quantiles - Gamma - Vol

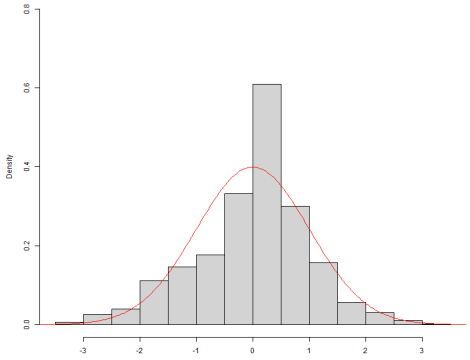

FIGURE N.38 – Histogramme des Résidus Quantiles - Log-normale - Vol

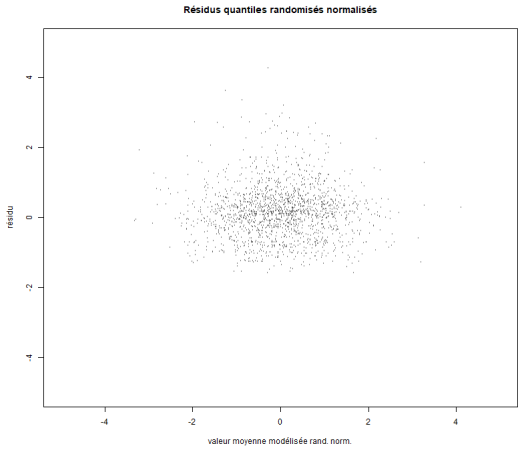

FIGURE N.39 – Résidus Quantiles - Gamma - Vol

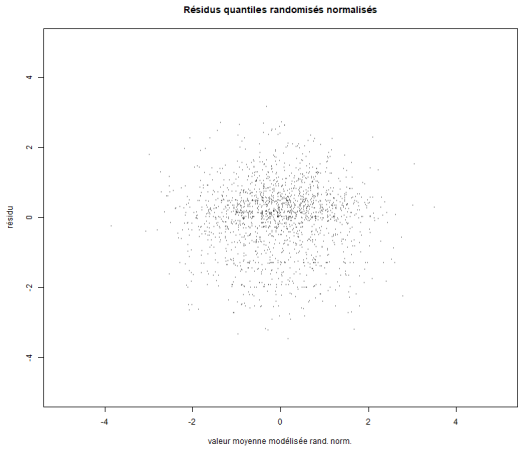

FIGURE N.40 – Résidus Quantiles - Log-normale - Vol

N.5 Toutes Garanties (Global)

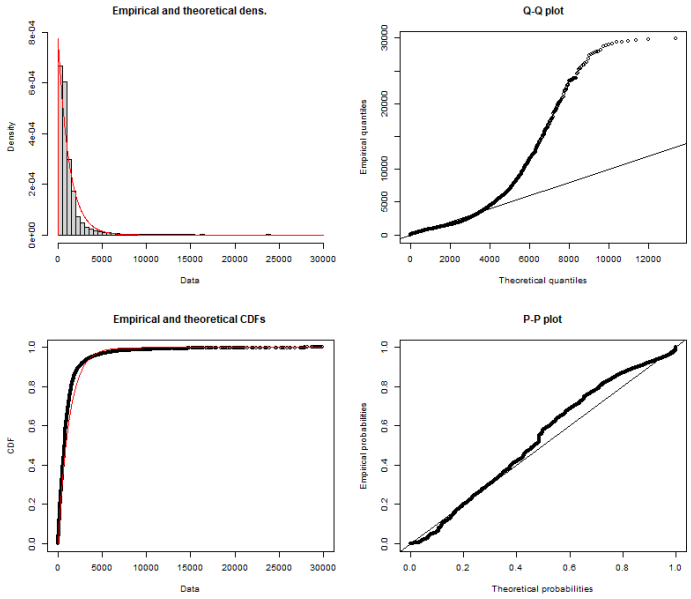

FIGURE N.41 – Analyse de la distribution du coût toutes garanties - loi Gamma

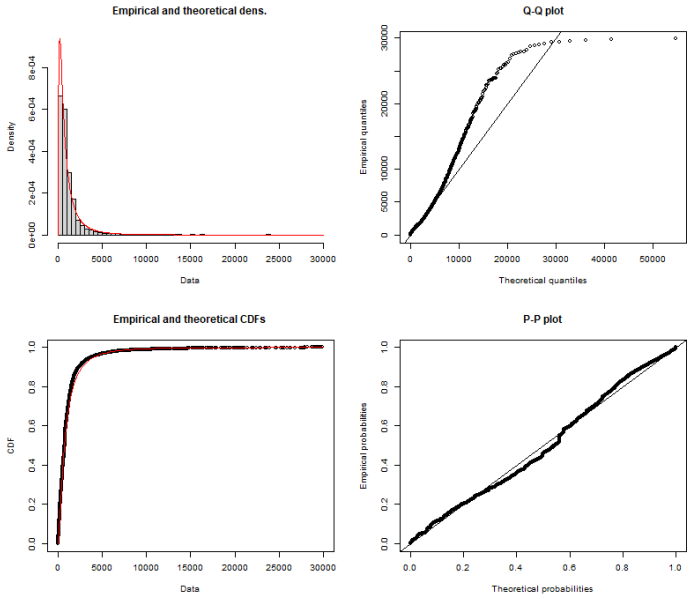

FIGURE N.42 – Analyse de la distribution du coût toutes garanties - loi Log-normale

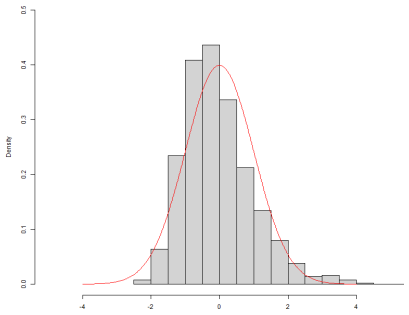

FIGURE N.43 – Histogramme des Résidus de Pearson - Gamma - toutes garanties

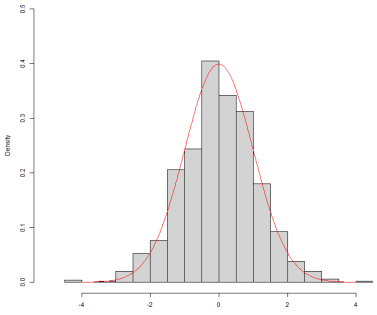

FIGURE N.44 – Histogramme des Résidus de Pearson - Log-normale - toutes garanties

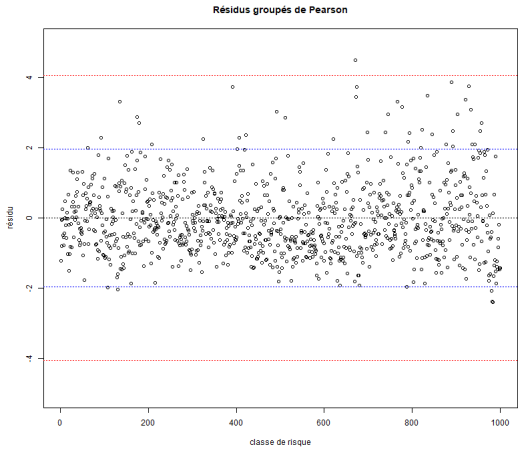

FIGURE N.45 – Résidus de Pearson en fonction de la classe de risque - Gamma - toutes garanties

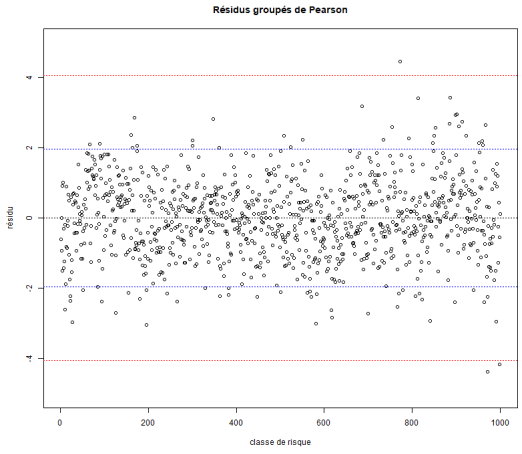

FIGURE N.46 – Résidus de Pearson en fonction de la classe de risque - Log-normale - toutes garanties

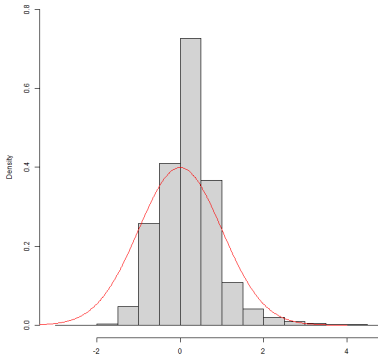

FIGURE N.47 – Histogramme des Résidus Quantiles - Gamma - toutes garanties

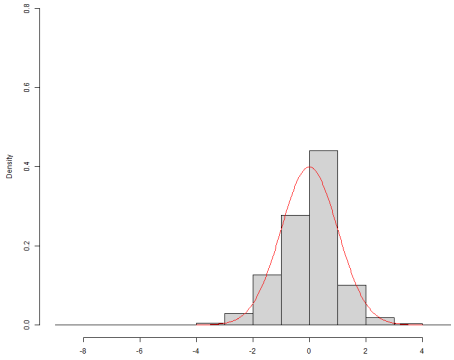

FIGURE N.48 – Histogramme des Résidus Quantiles - Log-normale - toutes garanties

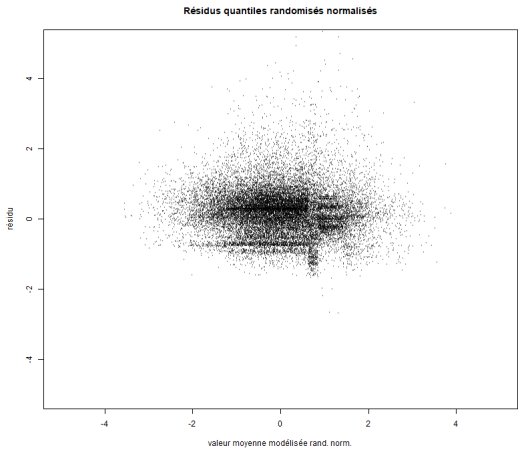

FIGURE N.49 – Résidus Quantiles - Gamma - toutes garanties

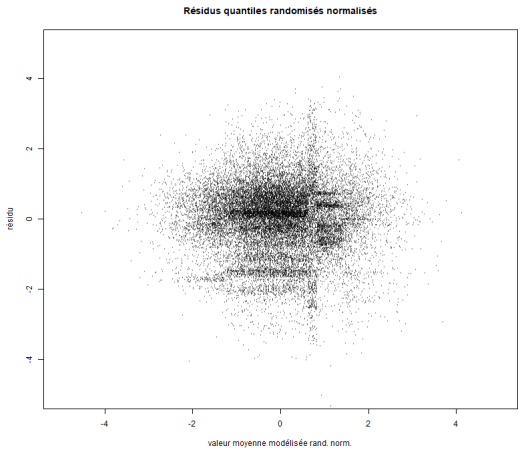

FIGURE N.50 – Résidus Quantiles - Log-normale - toutes garanties

Annexe O

Eléments graphiques complémentaires pour les modèles CART cout

Dans cette partie, les différents arbres CART pour le coût sont affichés.

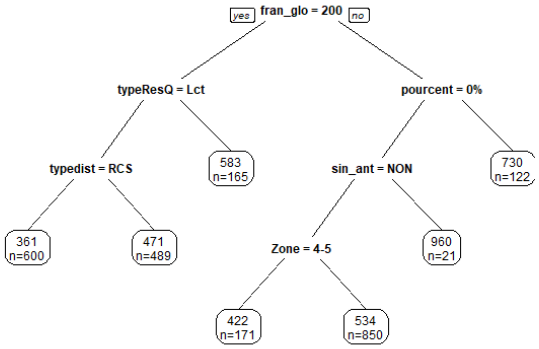

FIGURE O.1 – Arbre final de coût pour la garantie BDG - Approche 1

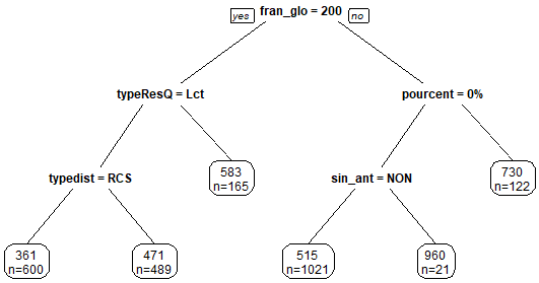

FIGURE O.2 – Arbre final de coût pour la garantie BDG - Approche 2

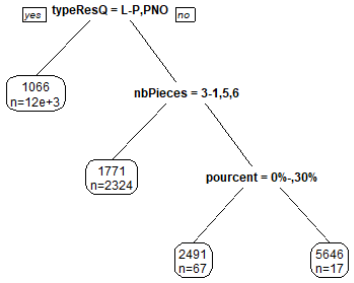

FIGURE O.3 – Arbre final de coût pour la garantie DDE - Approche 1

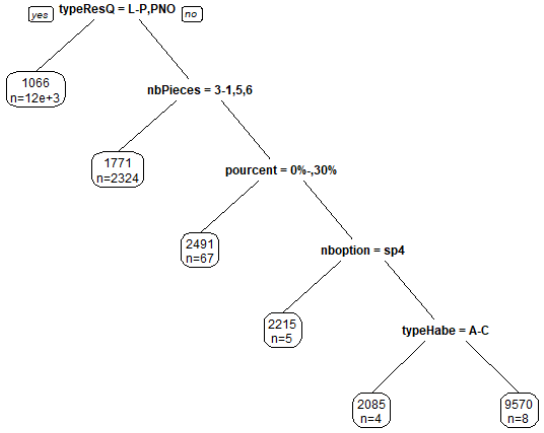

FIGURE O.4 – Arbre final de coût pour la garantie DDE - Approche 2

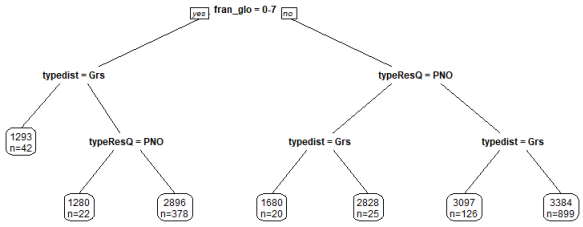

FIGURE O.5 – Arbre final de coût pour la garantie Incendie - Approche 1

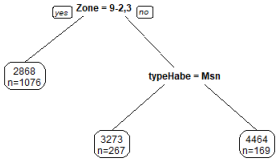

FIGURE O.6 – Arbre final de coût pour la garantie Incendie - Approche 2

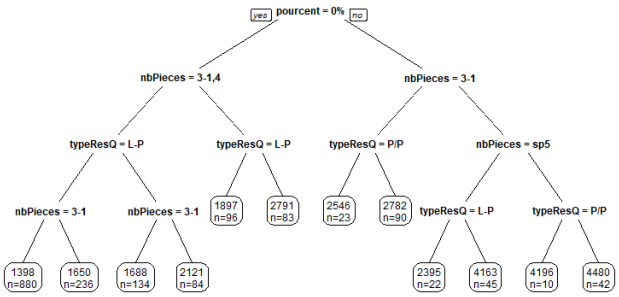

FIGURE O.7 – Arbre final de coût pour la garantie Vol - Approche 1

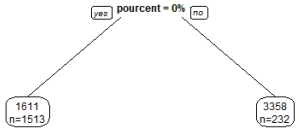

FIGURE O.8 – Arbre final de coût pour la garantie Vol - Approche 2

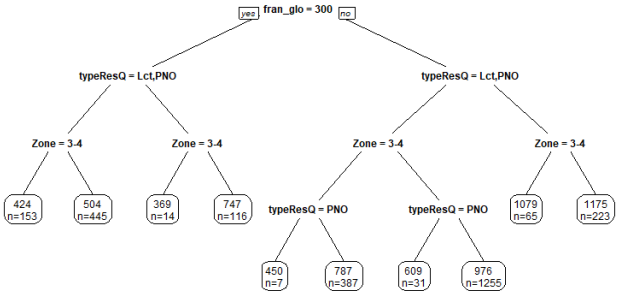

FIGURE O.9 – Arbre final de coût pour la garantie RC - Approche 1

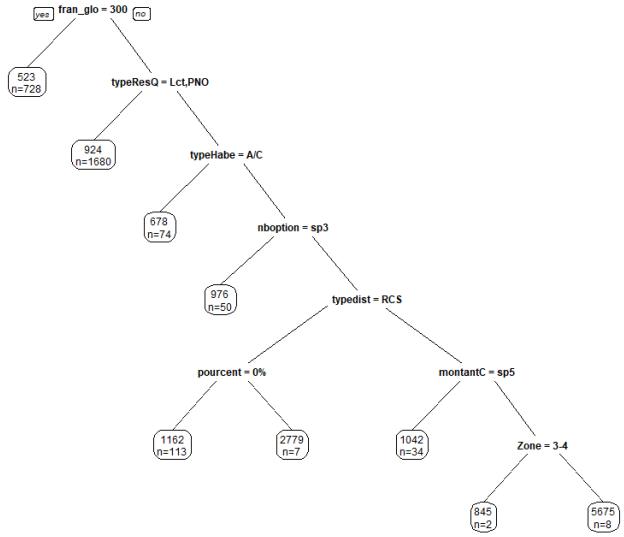

FIGURE O.10 – Arbre final de coût pour la garantie RC - Approche 2

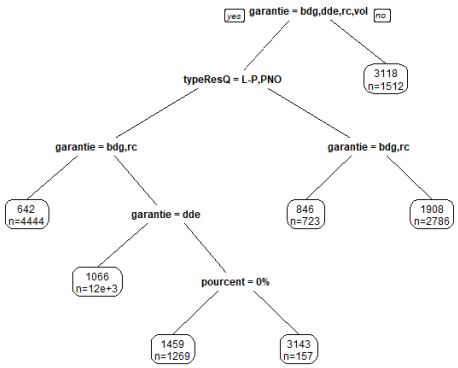

FIGURE O.11 – Arbre final de coût toutes garanties - Approche 1

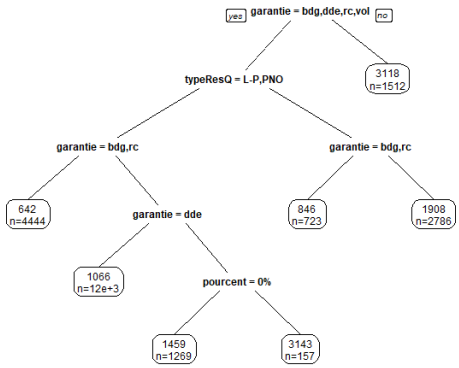

FIGURE O.12 – Arbre final de coût toutes garanties - Approche 2

Annexe P

Eléments Graphiques *Random Forest* coût

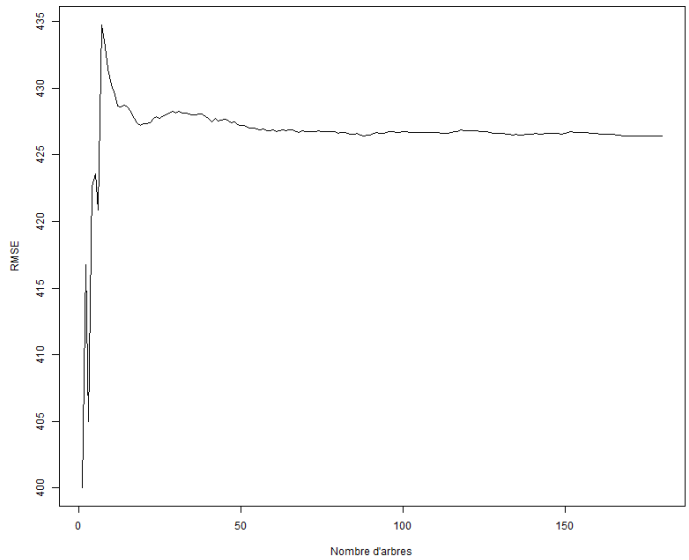

FIGURE P.1 – Evolution RMSE - RF Approche 1 - Coût - BDG

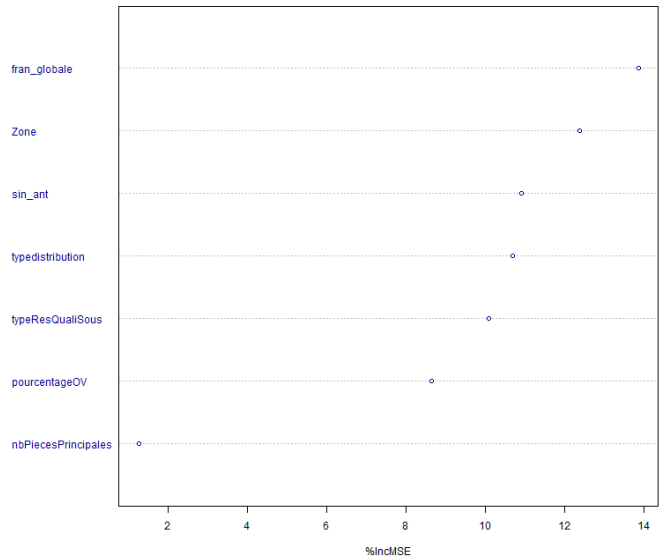

FIGURE P.2 – Importance variables - RF Approche 1 - coût - BDG

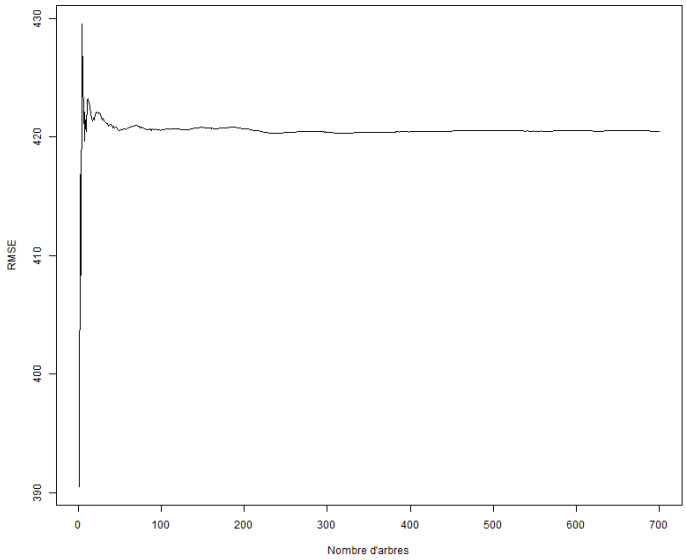

FIGURE P.3 – Evolution RMSE - RF Approche 2 - Coût - BDG

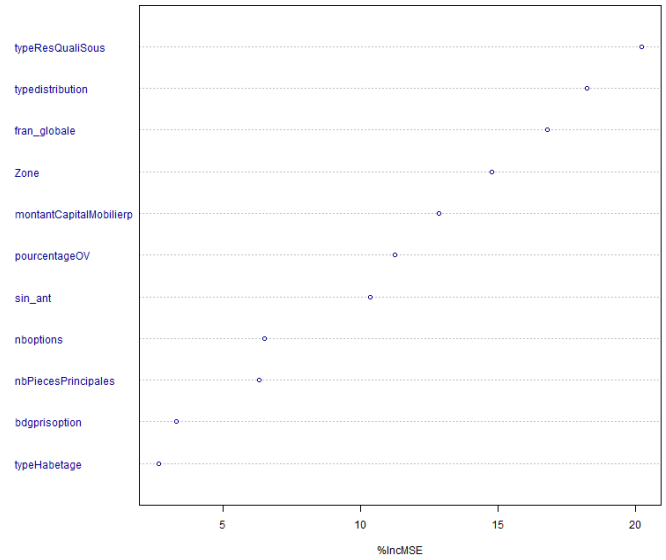

FIGURE P.4 – Importance variables - RF Approche 2 - coût - BDG

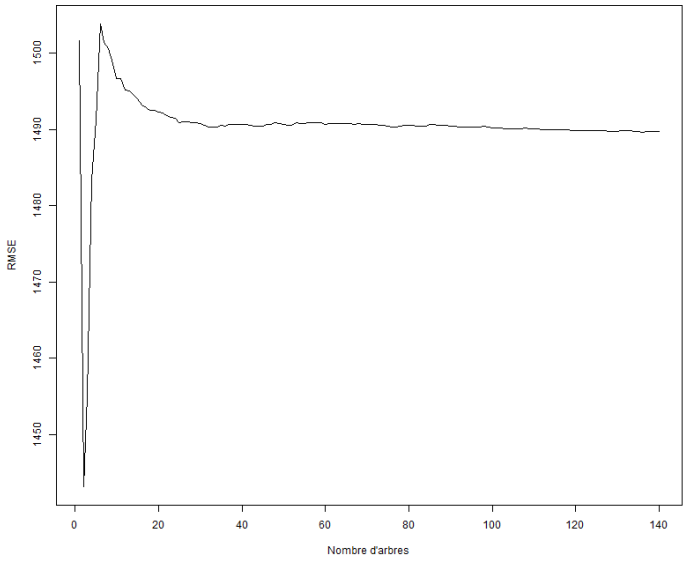

FIGURE P.5 – Evolution RMSE - RF A1 - Coût - DDE

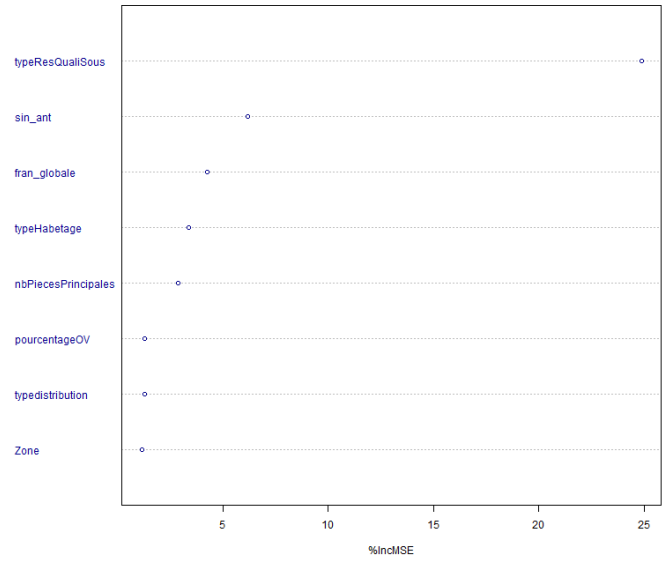

FIGURE P.6 – Importance variables - RF A1 - coût - DDE

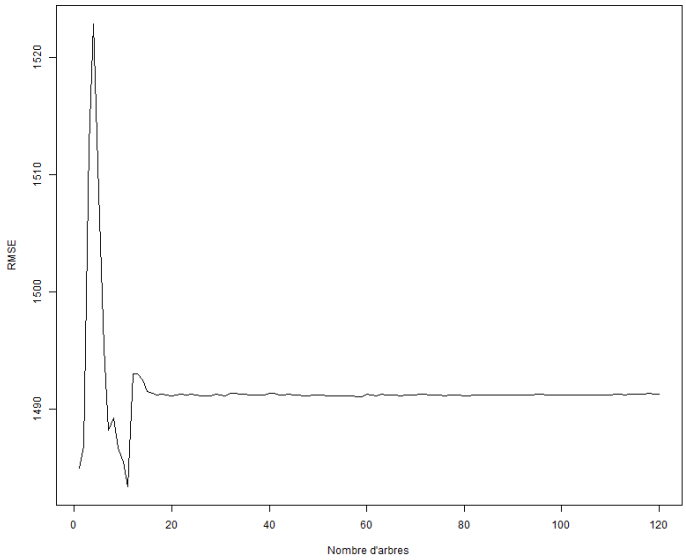

FIGURE P.7 – Evolution RMSE - RF A2 - Coût - DDE

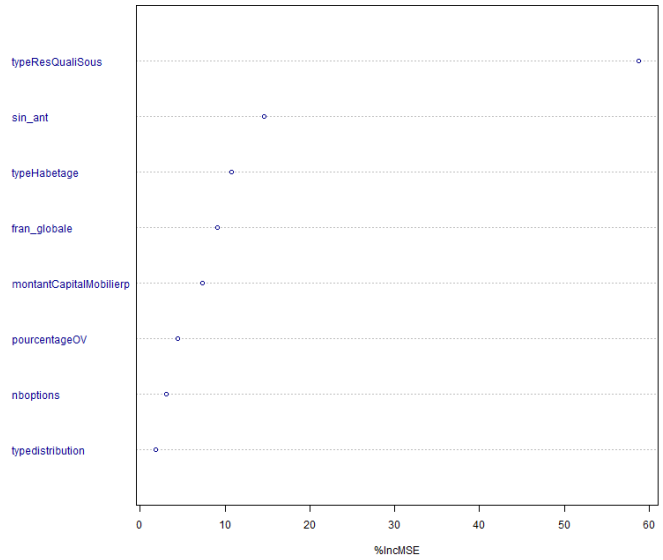

FIGURE P.8 – Importance variables - RF A2 - coût - DDE

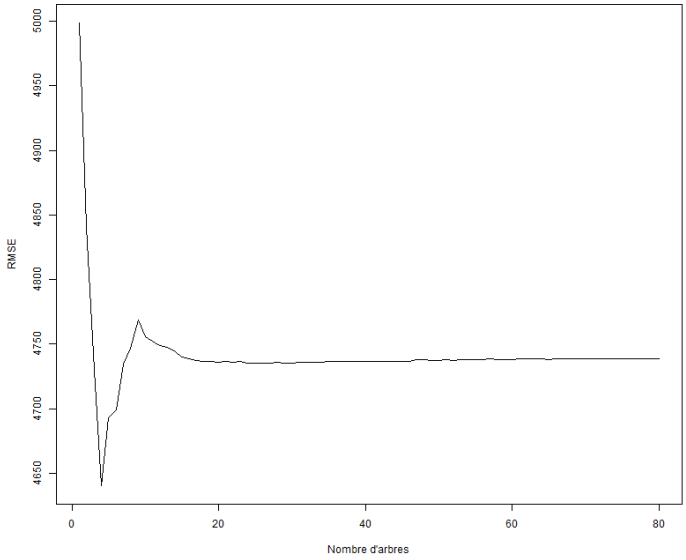

FIGURE P.9 – Evolution RMSE - RF A1 - Coût - Incendie

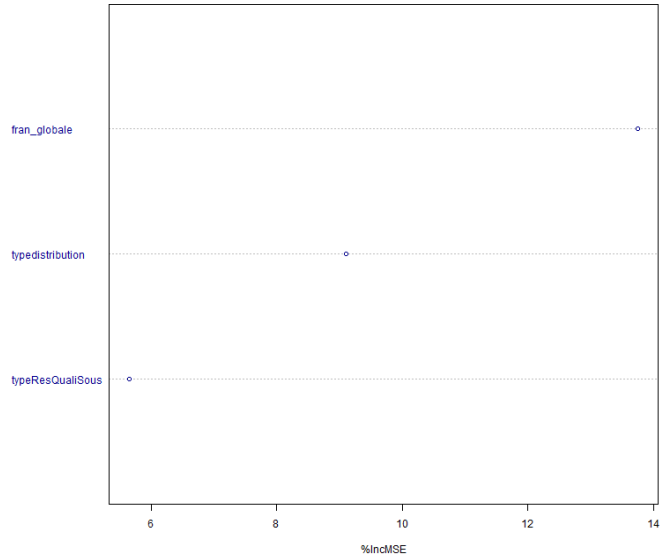

FIGURE P.10 – Importance variables - RF A1 - coût - Incendie

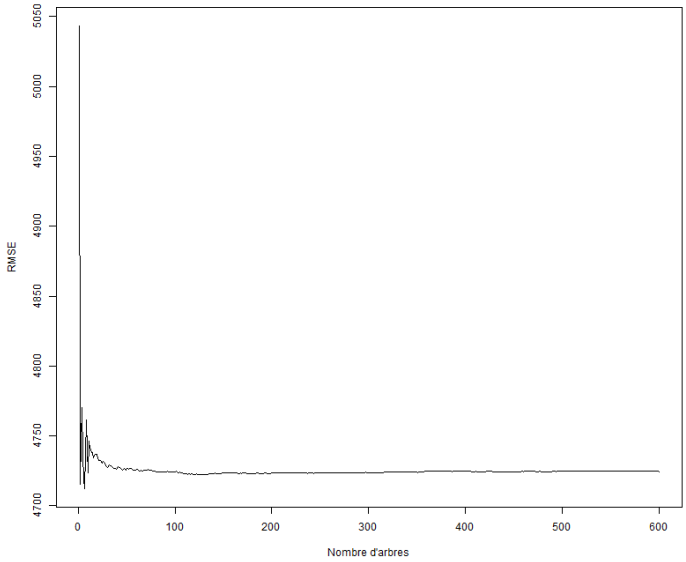

FIGURE P.11 – Evolution RMSE - RF A2 - Coût - Incendie

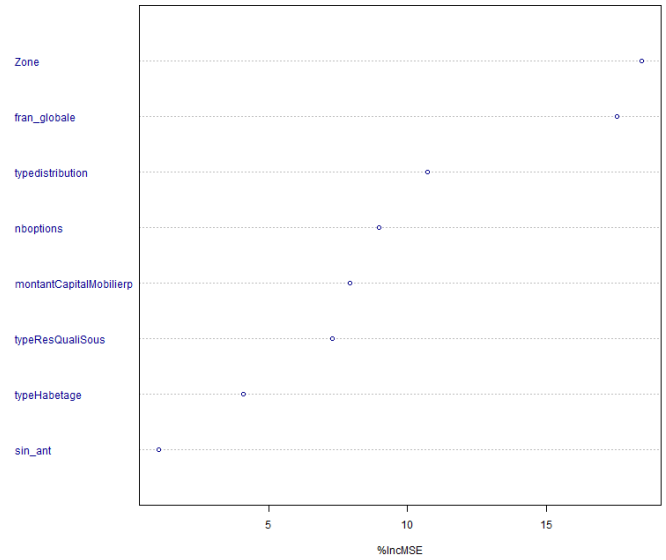

FIGURE P.12 – Importance variables - RF A2 - coût - Incendie

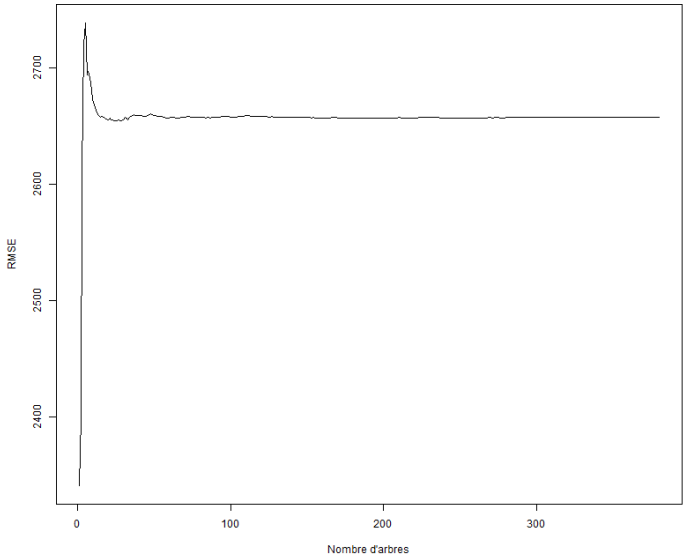

FIGURE P.13 – Evolution RMSE - RF A1 - Coût - Vol

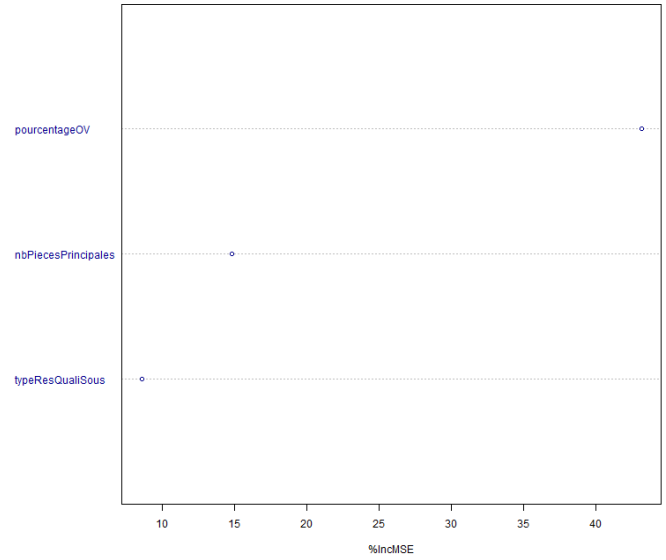

FIGURE P.14 – Importance variables - RF A1 - coût - Vol

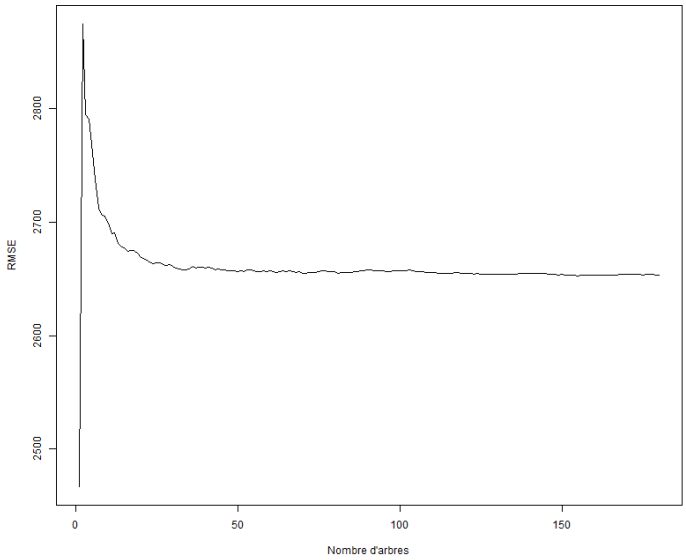

FIGURE P.15 – Evolution RMSE - RF A2 - Coût - Vol

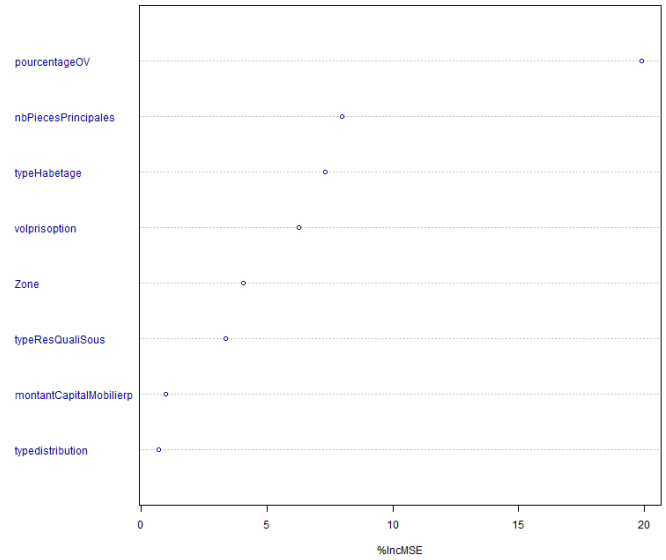

FIGURE P.16 – Importance variables - RF A2 - coût - Vol

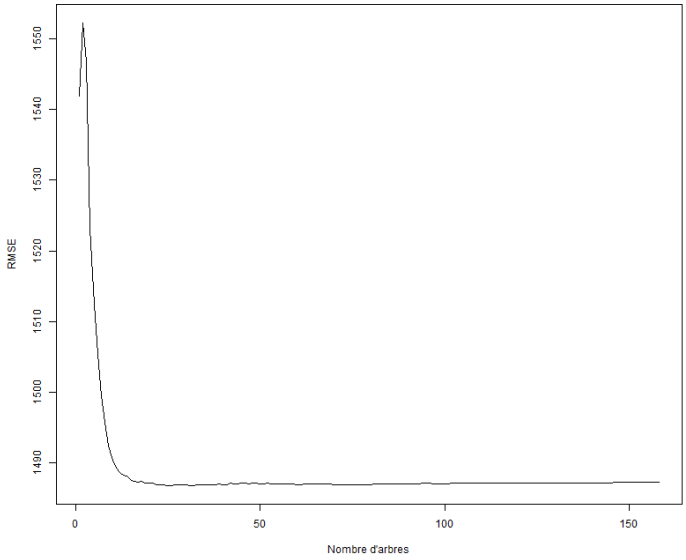

FIGURE P.17 – Evolution RMSE - RF A1 - Coût - RC

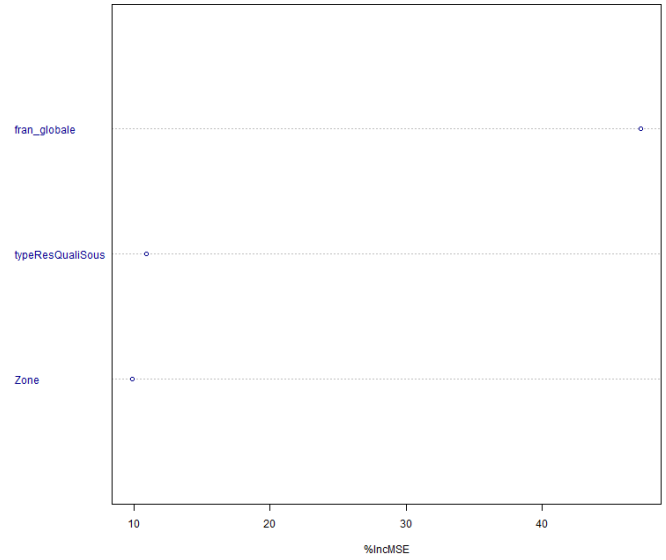

FIGURE P.18 – Importance variables - RF A1 - coût - RC

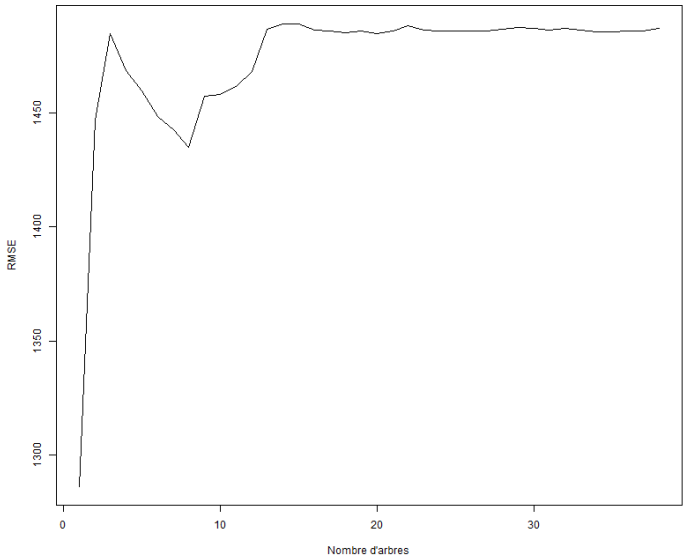

FIGURE P.19 – Evolution RMSE - RF A2 - Coût - RC

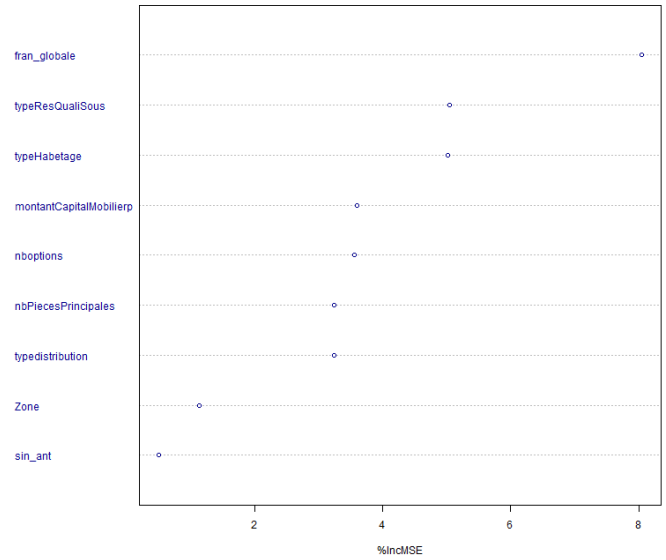

FIGURE P.20 – Importance variables - RF A2 - coût - RC

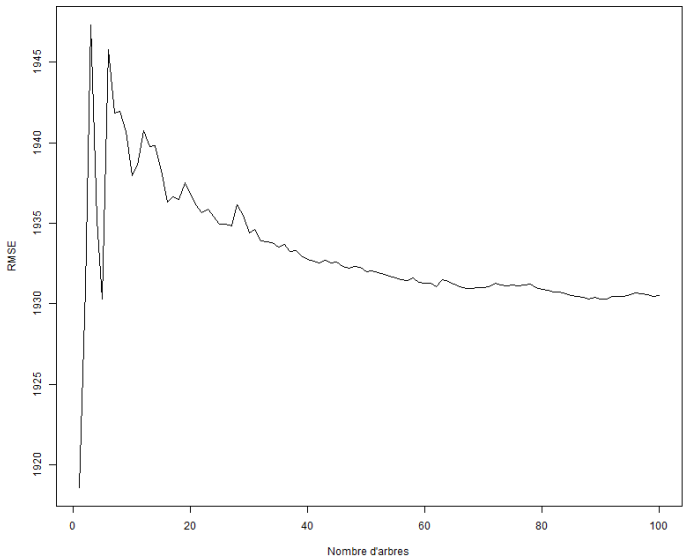

FIGURE P.21 – Evolution RMSE - RF A1 - Coût - Toutes Garanties

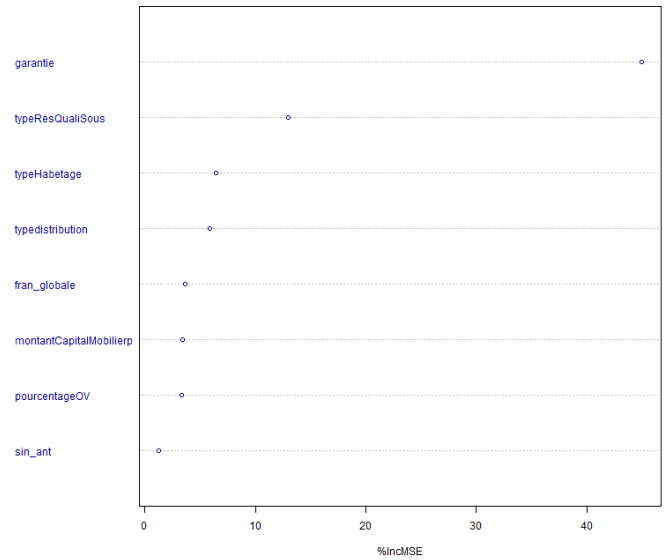

FIGURE P.22 – Importance variables - RF A1 - coût - Toutes Garanties

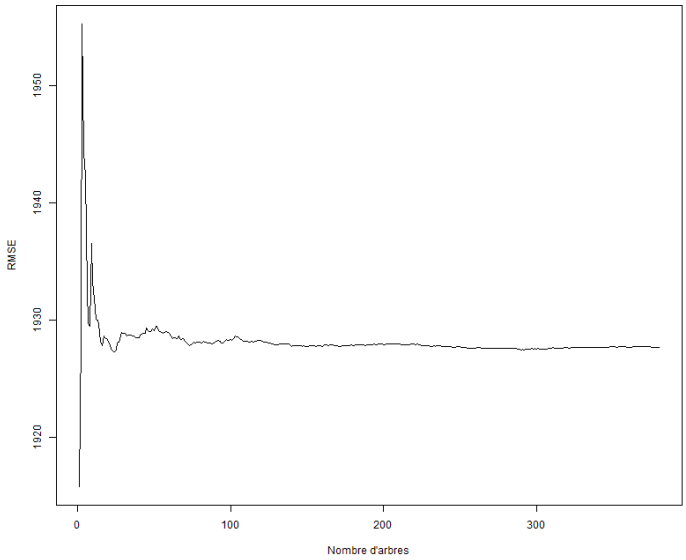

FIGURE P.23 – Evolution RMSE - RF A2 - Coût - Toutes Garanties

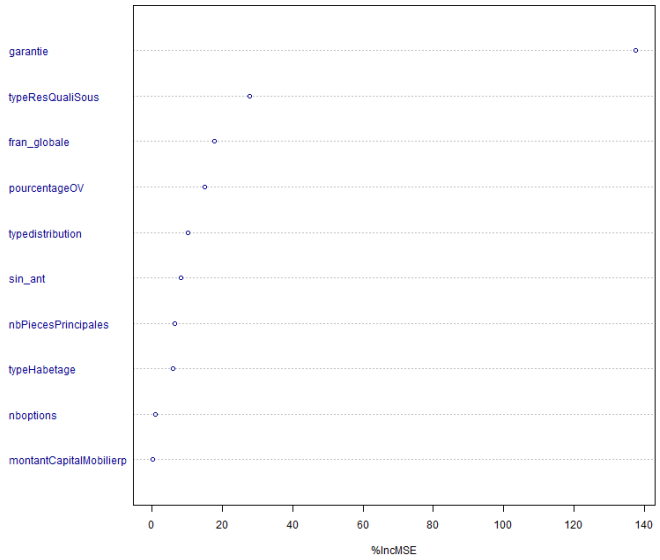

FIGURE P.24 – Importance variables - RF A2 - coût - Toutes Garanties

Annexe Q

Eléments Graphiques *XGBoost* coût

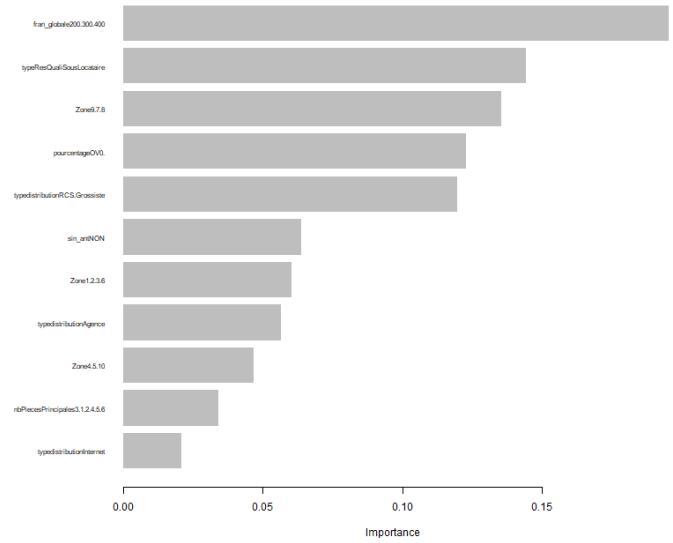

FIGURE Q.1 – Importance variables - XGBoost A1 - Coût - BDG

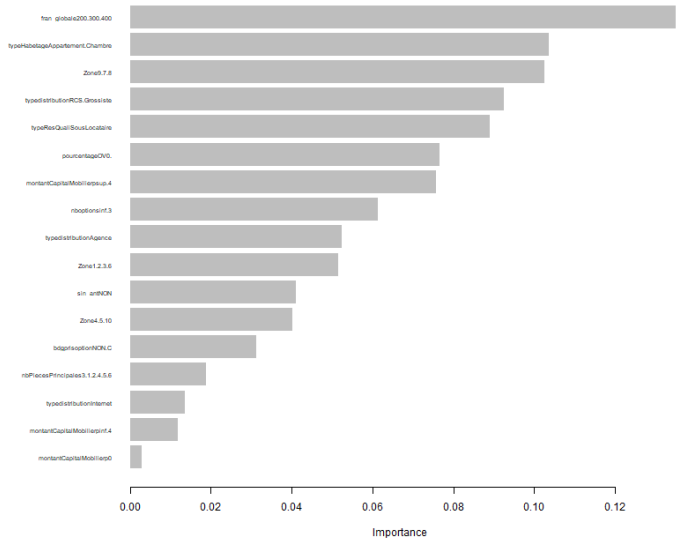

FIGURE Q.2 – Importance variables - XGBoost A2 - Coût - BDG

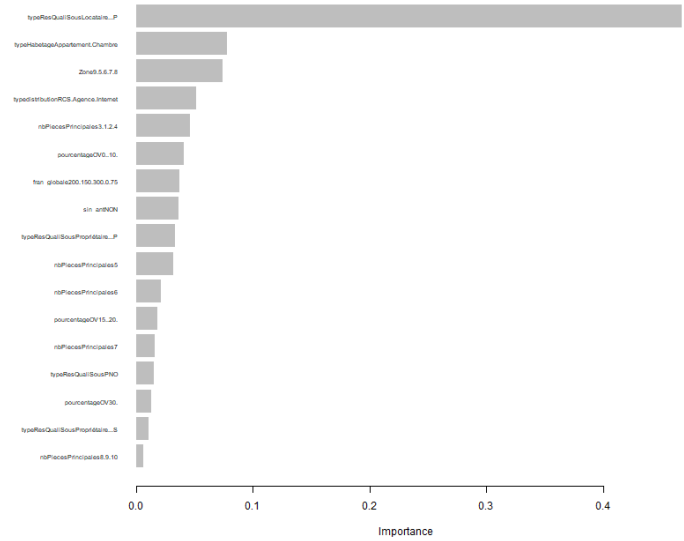

FIGURE Q.3 – Importance variables - XGBoost A1 - Coût - DDE

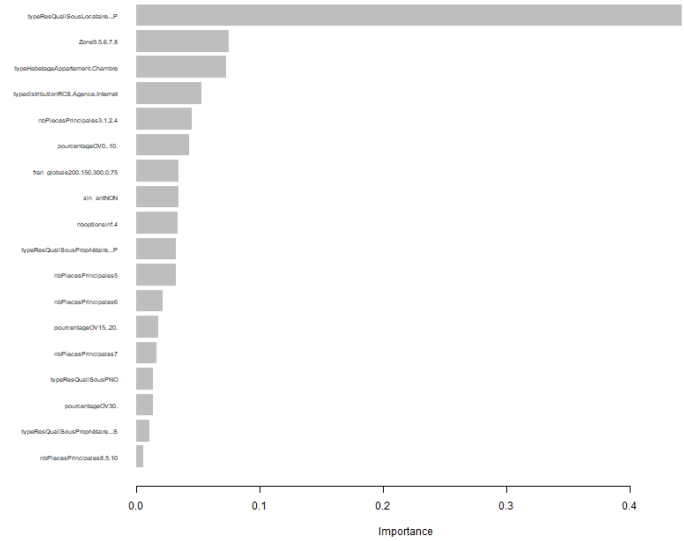

FIGURE Q.4 – Importance variables - XGBoost A2 - Coût - DDE

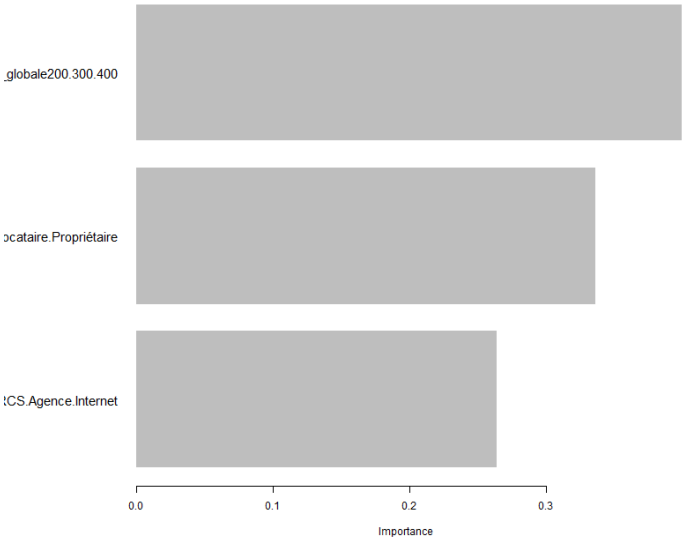

FIGURE Q.5 – Importance variables - XGBoost A1 - Coût - Incendie

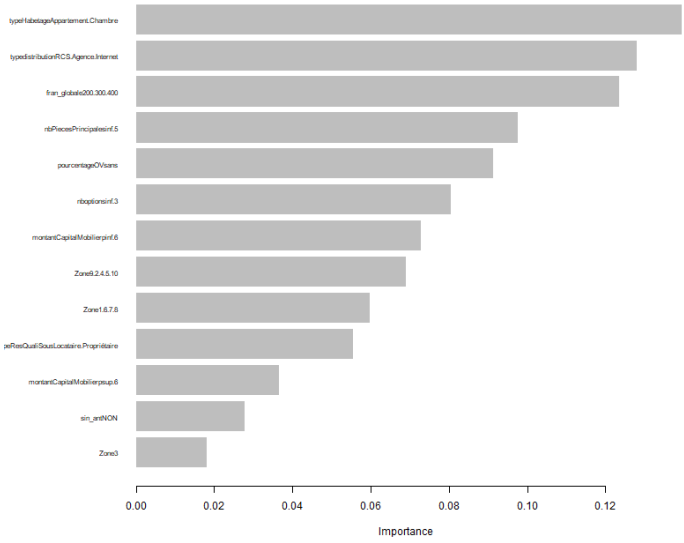

FIGURE Q.6 – Importance variables - XGBoost A2 - Coût - Incendie

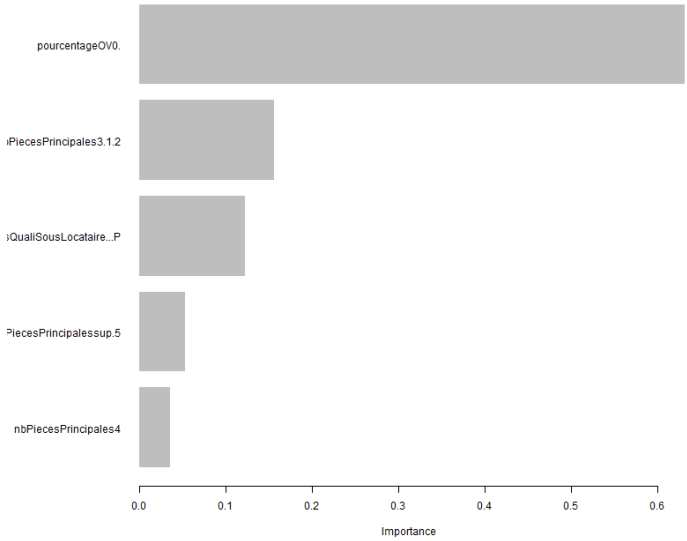

FIGURE Q.7 – Importance variables - XGBoost A1 - Coût - Vol

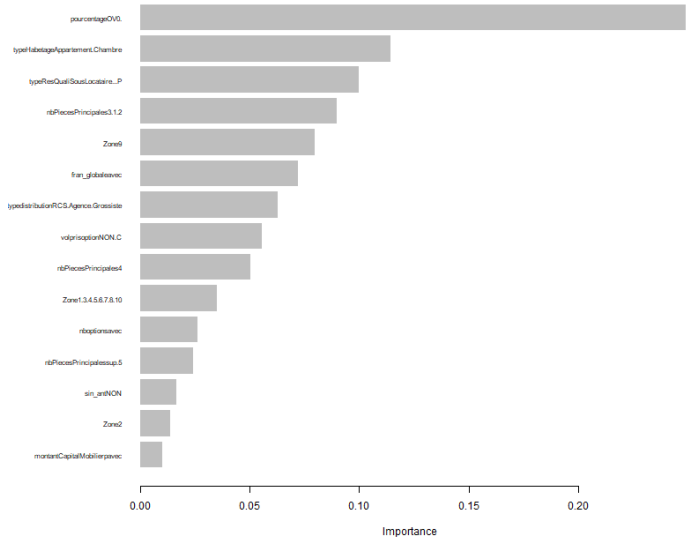

FIGURE Q.8 – Importance variables - XGBoost A2 - Coût - Vol

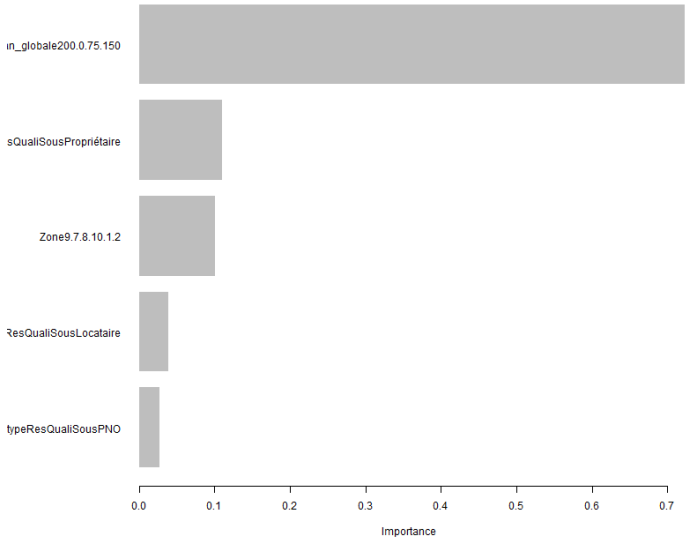

FIGURE Q.9 – Importance variables - XGBoost A1 - Coût - RC

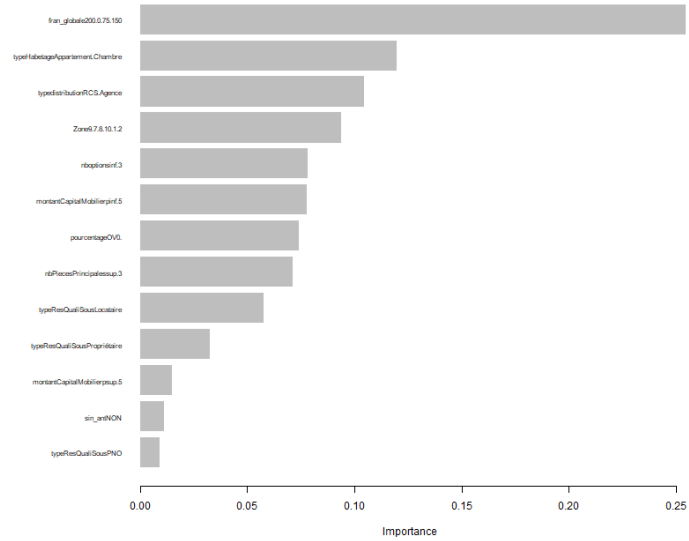

FIGURE Q.10 – Importance variables - XGBoost A2 - Coût - RC

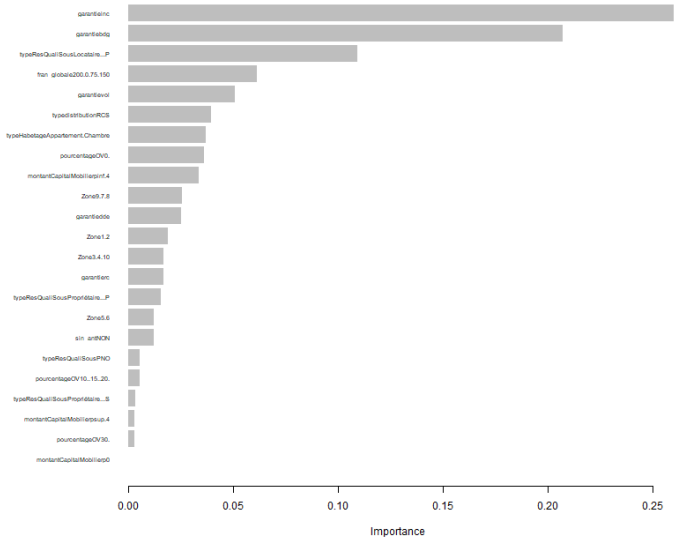

FIGURE Q.11 – Importance variables - XGBoost A1 - Coût - Toutes Garanties

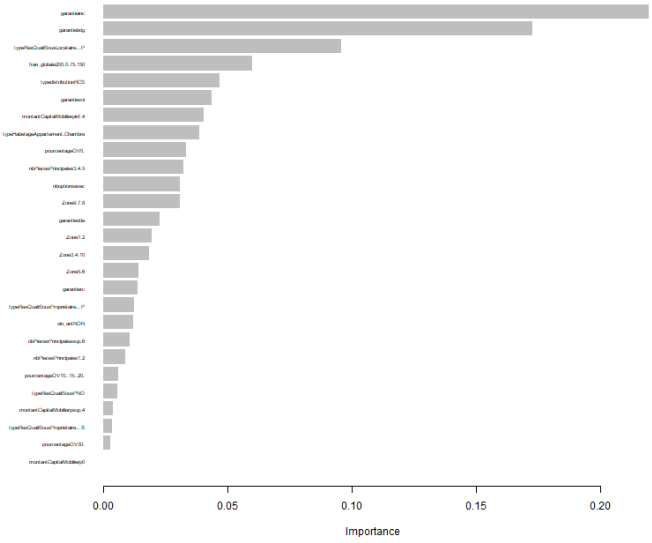

FIGURE Q.12 – Importance variables - XGBoost A2 - Coût - Toutes Garanties

Table des figures

1 Comparaison tarif estimé/tarif actuel pour la garantie BDG . . . 4
2 Comparaison tarif estimé/tarif actuel pour la garantie DDE . . . 4
3 Comparaison tarif estimé/tarif actuel pour la garantie Incendie . . . 4
4 Comparaison tarif estimé/tarif actuel pour la garantie Vol . . . 4
5 Comparaison tarif estimé/tarif actuel pour la garantie RC . . . 4
6 Comparison of estimated rates with current rates for Glass Breakage guarantee . . . 9
7 Comparison of estimated rates with current rates for Water Damage guarantee . . . 9
8 Comparison of estimated rates with current rates for Fire guarantee . . . 9
9 Comparison of estimated rates with current rates for Theft guarantee . . . 9
10 Comparison of estimated rates with current rates for Civil Liability guarantee . . . 9

1.1 Évolution du chiffre d'affaire annuel par branche, en Mds d'€ . . . 18
1.2 Ratios de solvabilité des assureurs français . . . 19
1.3 Parts de chiffre d'affaire entre IARD et "Vie" en pourcentage . . . 19
1.4 Part de l'assurance de biens et responsabilité (dont MRH) et part IARD complète, en mds d'€ . . . 20
1.5 Évolution annuelle du chiffre d'affaire de la MRH dans l'IARD, en pourcentage . . . 20
1.6 Variations fréquence et coût moyen MRH par garantie . . . 21
1.7 Taux moyens de résiliations et d'affaires nouvelles des contrats MRH . . . 21

2.1 Illustration simplifiée d'un arbre CART . . . 31
2.2 Schéma Explicatif du *Bagging* . . . 33
2.3 Schéma Explicatif du *Random Forest* . . . 34
2.4 Exemple de représentation d'une courbe de Lorenz . . . 39
2.5 Exemple de représentation d'une courbe de Lorenz . . . 39

3.1 Triangle de règlement pour le provisionnement . . . 45
3.2 Coefficients pour le provisionnement des sinistres . . . 45
3.3 Tableau d'analyse des sinistres "orphelins" . . . 46
3.4 Prime pure pour les sinistres "orphelins" par garantie . . . 46
3.5 Tableau d'analyse des sinistres graves . . . 46
3.6 Prime pure pour les sinistres graves, par garantie . . . 47
3.7 Étude de la fréquence et du coût moyen suivant le type d'habitation . . . 47
3.8 Étude de la fréquence et du coût moyen suivant le type de Résidence . . . 48
3.9 Étude de la fréquence et du coût moyen suivant le nombre de pièces principales . . . 48
3.10 Étude de la fréquence et du coût moyen suivant la qualité du souscripteur . . . 49
3.11 Étude de la fréquence et du coût moyen suivant le partenaire . . . 49
3.12 Corrélation entre les variables pré-sélectionnées pour la tarification . . . 51
3.13 Corrélation entre les variables pré-sélectionnées (avec variables créées) . . . 52

4.1 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - DDE . . . 56
4.2 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- Poisson - DDE . . . 56
4.3 Distribution empirique des résidus groupés de Pearson - Poisson - DDE . . . 57
4.4 Fonction de répartition empirique des résidus de Pearson - Poisson - DDE . . . 57
4.5 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - DDE . . . 57

4.6 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisés normalisés) - Poisson - DDE . . . 57
4.7 Distribution empirique des résidus quantiles randomisés normalisés - Poisson - DDE . . . 58
4.8 Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - DDE . . . 58
4.9 Comparaison des courbes de Lorenz - GLM - BDG . . . 59
4.10 Comparaison des courbes de Lorenz - GLM - DDE . . . 60
4.11 Comparaison des courbes de Lorenz - GLM - Incendie . . . 61
4.12 Comparaison des courbes de Lorenz - GLM - Vol . . . 62
4.13 Comparaison des courbes de Lorenz - GLM - RC . . . 63
4.14 Arbre final fréquence pour la garantie DDE . . . 65
4.15 Lecture de l'arbre . . . 65
4.16 Comparaison des courbes de Lorenz GLM/CART pour la garantie BDG . . . 66
4.17 Comparaison des courbes de Lorenz GLM/CART pour la garantie DDE . . . 66
4.18 Comparaison des courbes de Lorenz GLM/CART pour la garantie Incendie . . . 66
4.19 Comparaison des courbes de Lorenz GLM/CART pour la garantie Vol . . . 66
4.20 Comparaison des courbes de Lorenz GLM/CART pour la garantie RC . . . 67

5.1 Analyse de la distribution du coût DDE - loi Gamma . . . 72
5.2 Analyse de la distribution du coût DDE - loi Log-normale . . . 72
5.3 Histogramme des Résidus de Pearson - Gamma - DDE . . . 74
5.4 Histogramme des Résidus de Pearson - Log-normale - DDE . . . 74
5.5 Résidus de Pearson en fonction de la classe de risque - Gamma - DDE . . . 74
5.6 Résidus de Pearson en fonction de la classe de risque - Log-normale - DDE . . . 74
5.7 Histogramme des Résidus Quantiles - Gamma - DDE . . . 75
5.8 Histogramme des Résidus Quantiles - Log-normale - DDE . . . 75
5.9 Résidus des Quantiles - Gamma - DDE . . . 75
5.10 Résidus des Quantiles - Log-normale - DDE . . . 75
5.11 Courbes de Lorenz - coût - BDG . . . 76
5.12 Courbes de Lorenz - coût - DDE . . . 77
5.13 Courbes de Lorenz - coût - Incendie . . . 78
5.14 Courbes de Lorenz - coût - Vol . . . 79
5.15 Courbes de Lorenz - coût - RC . . . 80
5.16 Courbes de Lorenz - coût - Global . . . 82

6.1 Comparaison tarif estimé/tarif actuel sur l'ensemble des garanties . . . 93
6.2 Comparaison tarif estimé/tarif actuel pour la garantie BDG . . . 94
6.3 Comparaison tarif estimé/tarif actuel pour la garantie DDE . . . 94
6.4 Comparaison tarif estimé/tarif actuel pour la garantie Incendie . . . 94
6.5 Comparaison tarif estimé/tarif actuel pour la garantie Vol . . . 94
6.6 Comparaison tarif estimé/tarif actuel pour la garantie RC . . . 95
6.7 Comparaison tarif en fonction de type Habitation et Qualité du Souscripteur - moins de −50% DDE 95
6.8 Comparaison tarif en fonction de type Habitation et Qualité du Souscripteur - plus 50% DDE . . . 95
6.9 Comparaison tarif par nombre de pièces (Maison-Locataire) - moins de −50% DDE . . . 96
6.10 Comparaison tarif par franchise (Maison-Locataire-3 pièces principales) - moins de −50% DDE . . . 96
6.11 Comparaison tarif par franchise (Maison-Locataire- 4 pièces principales) - moins de −50% DDE . . . 96
6.12 Comparaison tarif par franchise (Maison-Locataire- plus 4 pièces principales) - moins de −50% DDE 97
6.13 Comparaison tarif par franchise (Maison-Locataire-plus de 2 pièces principales) - moins de −50% DDE 97
6.14 Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - sans franchise) - moins de −50% DDE . . . 98
6.15 Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - 200) - moins de −50% DDE . . . 98
6.16 Comparaison tarif par pourcentage d'OV (Maison-Locataire-plus de 2 pièces principales - 400) - moins de −50% DDE . . . 98
6.17 Comparaison tarif par nombre de pièces (Appartement-Propriétaire) - plus de 50% DDE . . . 99
6.18 Comparaison tarif par nombre de pièces (Maison-Propriétaire) - plus de 50% DDE . . . 99
6.19 Comparaison tarif par nombre de pièces (Appartement-PNO) - plus de 50% DDE . . . 99
6.20 Comparaison tarif par nombre de pièces (Maison-PNO) - plus de 50% DDE . . . 99

6.21 Comparaison tarif par franchise (Appartement-PNO) - plus de 50% DDE . . . 100
6.22 Comparaison tarif par franchise (Maison-PNO) - plus de 50% DDE . . . 100
6.23 Comparaison tarif par franchise (Appartement-Propriétaire) - plus de 50% DDE . . . 100
6.24 Comparaison tarif par pourcentage d'OV (Appartement-Propriétaire) - plus de 50% DDE . . . 101
6.25 Comparaison tarif par nombre de pièces principales - plus de 50% Incendie . . . 101
6.26 Comparaison tarif par nombre de pièces principales - moins de −50% RC . . . 102
6.27 Comparaison tarif par nombre de pièces principales - moins de −50% Vol . . . 102

C.1 Tableau des facteurs de développement individuels . . . 110
C.2 Analyse des colonnes du tableau précédent . . . 110
C.3 *C-C plots* pour le provisionnement . . . 110

D.1 Comparaison des données et de l'indice FFB sur toutes les garanties . . . 111
D.2 Comparaison des données et de l'indice FFB par garantie . . . 112
D.3 Comparaison des données et de l'indice de l'inflation sur toutes les garanties . . . 112
D.4 Comparaison des données et de l'indice de l'inflation par garantie . . . 113

I.1 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - DDE 123
I.2 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - DDE . . . 123
I.3 Distribution empirique des résidus groupés de Pearson - BN - DDE . . . 123
I.4 Fonction de répartition empirique des résidus de Pearson - BN - DDE . . . 123
I.5 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - DDE . . . 124
I.6 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - DDE . . . 124
I.7 Distribution empirique des résidus quantiles randomisés normalisés - BN - DDE . . . 124
I.8 Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - DDE . . . 124
I.9 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - BDG . . . 124
I.10 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - BDG . . . 124
I.11 Distribution empirique des résidus groupés de Pearson - Poisson - BDG . . . 125
I.12 Fonction de répartition empirique des résidus de Pearson - Poisson - BDG . . . 125
I.13 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - BDG . . . 125
I.14 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - BDG . . . 125
I.15 Distribution empirique des résidus quantiles randomisés normalisés - Poisson - BDG . . . 125
I.16 Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - BDG . . . 125
I.17 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - BDG 126
I.18 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - BDG . . . 126
I.19 Distribution empirique des résidus groupés de Pearson - BN - BDG . . . 126
I.20 Fonction de répartition empirique des résidus de Pearson - BN - BDG . . . 126
I.21 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - BDG . . . 126
I.22 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - BDG . . . 126
I.23 Distribution empirique des résidus quantiles randomisés normalisés - BN - BDG . . . 127
I.24 Fonction de répartition empirique des résidus quantiles randomisées normalisés - BN - BDG . . . 127
I.25 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Incendie . . . 127
I.26 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Incendie . . . 127
I.27 Distribution empirique des résidus groupés de Pearson - Poisson - Incendie . . . 127
I.28 Fonction de répartition empirique des résidus de Pearson - Poisson - Incendie . . . 127
I.29 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Incendie . . . 128
I.30 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - Incendie . . . 128

I.31 Distribution empirique des résidus quantiles randomisés normalisés - Poisson - Incendie 128
I.32 Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - Incendie . 128
I.33 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Incendie128
I.34 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Incendie 128
I.35 Distribution empirique des résidus groupés de Pearson - BN - Incendie 129
I.36 Fonction de répartition empirique des résidus de Pearson - BN - Incendie 129
I.37 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Incendie 129
I.38 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - Incendie 129
I.39 Distribution empirique des résidus quantiles randomisées normalisés - BN - Incendie 129
I.40 Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - Incendie 129
I.41 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Vol130
I.42 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Vol 130
I.43 Distribution empirique des résidus groupés de Pearson - Poisson - Vol 130
I.44 Fonction de répartition empirique des résidus de Pearson - Poisson - Vol 130
I.45 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - Vol 130
I.46 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - Vol 130
I.47 Distribution empirique des résidus quantiles randomisées normalisés - Poisson - Vol 131
I.48 Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - Vol 131
I.49 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Vol . 131
I.50 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - Vol 131
I.51 Distribution empirique des résidus groupés de Pearson - BN - Vol 131
I.52 Fonction de répartition empirique des résidus de Pearson - BN - Vol 131
I.53 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - Vol 132
I.54 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - Vol 132
I.55 Distribution empirique des résidus quantiles randomisés normalisés - BN - Vol 132
I.56 Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - Vol 132
I.57 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - RC132
I.58 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- Poisson - RC 132
I.59 Distribution empirique des résidus groupés de Pearson - Poisson - RC 133
I.60 Fonction de répartition empirique des résidus de Pearson - Poisson - RC 133
I.61 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - Poisson - RC 133
I.62 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - Poisson - RC 133
I.63 Distribution empirique des résidus quantiles randomisés normalisés - Poisson - RC 133
I.64 Fonction de répartition empirique des résidus quantiles randomisés normalisés - Poisson - RC 133
I.65 Résidus de Pearson en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - RC . 134
I.66 Résidus de Pearson en fonction de la classe de risque (1 000 classes)- BN - RC 134
I.67 Distribution empirique des résidus groupés de Pearson - BN - RC 134
I.68 Fonction de répartition empirique des résidus de Pearson - BN - RC 134
I.69 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (échelle logarithmique) - BN - RC 134
I.70 Résidus quantiles randomisés normalisés en fonction de la valeur moyenne modélisée (randomisée normalisée) - BN - RC 134
I.71 Distribution empirique des résidus quantiles randomisés normalisés - BN - RC 135
I.72 Fonction de répartition empirique des résidus quantiles randomisés normalisés - BN - RC 135

J.1 Arbre final pour la garantie RC 136
J.2 Arbre final pour la garantie BDG 136
J.3 Arbre final pour la garantie Incendie 137
J.4 Arbre final pour la garantie Vol 137

K.1 Evolution RMSE - RF - Fréquence - BDG . . . 138
K.2 Importance variables - RF - Fréquence - BDG . . . 138
K.3 Evolution RMSE - RF - Fréquence - DDE . . . 139
K.4 Importance variables - RF - Fréquence - DDE . . . 139
K.5 Evolution RMSE - RF - Fréquence - Incendie . . . 139
K.6 Importance variables - RF - Fréquence - Incendie . . . 139
K.7 Evolution RMSE - RF - Fréquence - Incendie (Sans la variable avec une importance négative) . . . 140
K.8 Importance variables - RF - Fréquence - Incendie (Sans la variable avec une importance négative) . . 140
K.9 Evolution RMSE - RF - Fréquence - Vol . . . 140
K.10 Importance variables - RF - Fréquence - Vol . . . 140
K.11 Evolution RMSE - RF - Fréquence - RC . . . 141
K.12 Importance variables - RF - Fréquence - RC . . . 141

L.1 Importance variables - XGBoost - Fréquence - BDG . . . 142
L.2 Importance variables - XGBoost - Fréquence - DDE . . . 142
L.3 Importance variables - XGBoost - Fréquence - Incendie . . . 143
L.4 Importance variables - XGBoost - Fréquence - Vol . . . 143
L.5 Importance variables - XGBoost - Fréquence - RC . . . 143

N.1 Analyse de la distribution du coût BDG - loi Gamma . . . 147
N.2 Analyse de la distribution du coût BDG - loi Log-normale . . . 147
N.3 Histogramme des Résidus de Pearson - Gamma - BDG . . . 147
N.4 Histogramme des Résidus de Pearson - Log-normale - BDG . . . 147
N.5 Résidus de Pearson en fonction de la classe de risque - Gamma - BDG . . . 148
N.6 Résidus de Pearson en fonction de la classe de risque - Log-normale - BDG . . . 148
N.7 Histogramme des Résidus Quantiles - Gamma - BDG . . . 148
N.8 Histogramme des Résidus Quantiles - Log-normale - BDG . . . 148
N.9 Résidus Quantiles - Gamma - BDG . . . 148
N.10 Résidus Quantiles - Log-normale - BDG . . . 148
N.11 Analyse de la distribution du coût Incendie - loi Gamma . . . 149
N.12 Analyse de la distribution du coût Incendie - loi Log-normale . . . 149
N.13 Histogramme des Résidus de Pearson - Gamma - Incendie . . . 149
N.14 Histogramme des Résidus de Pearson - Log-normale - Incendie . . . 149
N.15 Résidus de Pearson en fonction de la classe de risque - Gamma - Incendie . . . 149
N.16 Résidus de Pearson en fonction de la classe de risque - Log-normale - Incendie . . . 149
N.17 Histogramme des Résidus Quantiles - Gamma - Incendie . . . 150
N.18 Histogramme des Résidus Quantiles - Log-normale - Incendie . . . 150
N.19 Résidus Quantiles - Gamma - Incendie . . . 150
N.20 Résidus Quantiles - Log-normale - Incendie . . . 150
N.21 Analyse de la distribution du coût RC - loi Gamma . . . 151
N.22 Analyse de la distribution du coût RC - loi Log-normale . . . 151
N.23 Histogramme des Résidus de Pearson - Gamma - RC . . . 151
N.24 Histogramme des Résidus de Pearson - Log-normale - RC . . . 151
N.25 Résidus de Pearson en fonction de la classe de risque - Gamma - RC . . . 151
N.26 Résidus de Pearson en fonction de la classe de risque - Log-normale - RC . . . 151
N.27 Histogramme des Résidus Quantiles - Gamma - RC . . . 152
N.28 Histogramme des Résidus Quantiles - Log-normale - RC . . . 152
N.29 Résidus Quantiles - Gamma - RC . . . 152
N.30 Résidus Quantiles - Log-normale - RC . . . 152
N.31 Analyse de la distribution du coût Vol - loi Gamma . . . 153
N.32 Analyse de la distribution du coût Vol - loi Log-normale . . . 153
N.33 Histogramme des Résidus de Pearson - Gamma - Vol . . . 153
N.34 Histogramme des Résidus de Pearson - Log-normale - Vol . . . 153
N.35 Résidus de Pearson en fonction de la classe de risque - Gamma - Vol . . . 153
N.36 Résidus de Pearson en fonction de la classe de risque - Log-normale - Vol . . . 153
N.37 Histogramme des Résidus Quantiles - Gamma - Vol . . . 154
N.38 Histogramme des Résidus Quantiles - Log-normale - Vol . . . 154

N.39 Résidus Quantiles - Gamma - Vol . . . 154
N.40 Résidus Quantiles - Log-normale - Vol . . . 154
N.41 Analyse de la distribution du coût toutes garanties - loi Gamma . . . 155
N.42 Analyse de la distribution du coût toutes garanties - loi Log-normale . . . 155
N.43 Histogramme des Résidus de Pearson - Gamma - toutes garanties . . . 155
N.44 Histogramme des Résidus de Pearson - Log-normale - toutes garanties . . . 155
N.45 Résidus de Pearson en fonction de la classe de risque - Gamma - toutes garanties . . . 155
N.46 Résidus de Pearson en fonction de la classe de risque - Log-normale - toutes garanties . . . 155
N.47 Histogramme des Résidus Quantiles - Gamma - toutes garanties . . . 156
N.48 Histogramme des Résidus Quantiles - Log-normale - toutes garanties . . . 156
N.49 Résidus Quantiles - Gamma - toutes garanties . . . 156
N.50 Résidus Quantiles - Log-normale - toutes garanties . . . 156

O.1 Arbre final de coût pour la garantie BDG - Approche 1 . . . 157
O.2 Arbre final de coût pour la garantie BDG - Approche 2 . . . 157
O.3 Arbre final de coût pour la garantie DDE - Approche 1 . . . 158
O.4 Arbre final de coût pour la garantie DDE - Approche 2 . . . 158
O.5 Arbre final de coût pour la garantie Incendie - Approche 1 . . . 158
O.6 Arbre final de coût pour la garantie Incendie - Approche 2 . . . 158
O.7 Arbre final de coût pour la garantie Vol - Approche 1 . . . 159
O.8 Arbre final de coût pour la garantie Vol - Approche 2 . . . 159
O.9 Arbre final de coût pour la garantie RC - Approche 1 . . . 159
O.10 Arbre final de coût pour la garantie RC - Approche 2 . . . 159
O.11 Arbre final de coût toutes garanties - Approche 1 . . . 160
O.12 Arbre final de coût toutes garanties - Approche 2 . . . 160

P.1 Evolution RMSE - RF Approche 1 - Coût - BDG . . . 161
P.2 Importance variables - RF Approche 1 - coût - BDG . . . 161
P.3 Evolution RMSE - RF Approche 2 - Coût - BDG . . . 162
P.4 Importance variables - RF Approche 2 - coût - BDG . . . 162
P.5 Evolution RMSE - RF A1 - Coût - DDE . . . 162
P.6 Importance variables - RF A1 - coût - DDE . . . 162
P.7 Evolution RMSE - RF A2 - Coût - DDE . . . 163
P.8 Importance variables - RF A2 - coût - DDE . . . 163
P.9 Evolution RMSE - RF A1 - Coût - Incendie . . . 163
P.10 Importance variables - RF A1 - coût - Incendie . . . 163
P.11 Evolution RMSE - RF A2 - Coût - Incendie . . . 164
P.12 Importance variables - RF A2 - coût - Incendie . . . 164
P.13 Evolution RMSE - RF A1 - Coût - Vol . . . 164
P.14 Importance variables - RF A1 - coût - Vol . . . 164
P.15 Evolution RMSE - RF A2 - Coût - Vol . . . 165
P.16 Importance variables - RF A2 - coût - Vol . . . 165
P.17 Evolution RMSE - RF A1 - Coût - RC . . . 165
P.18 Importance variables - RF A1 - coût - RC . . . 165
P.19 Evolution RMSE - RF A2 - Coût - RC . . . 166
P.20 Importance variables - RF A2 - coût - RC . . . 166
P.21 Evolution RMSE - RF A1 - Coût - Toutes Garanties . . . 166
P.22 Importance variables - RF A1 - coût - Toutes Garanties . . . 166
P.23 Evolution RMSE - RF A2 - Coût - Toutes Garanties . . . 167
P.24 Importance variables - RF A2 - coût - Toutes Garanties . . . 167

Q.1 Importance variables - XGBoost A1 - Coût - BDG . . . 168
Q.2 Importance variables - XGBoost A2 - Coût - BDG . . . 168
Q.3 Importance variables - XGBoost A1 - Coût - DDE . . . 169
Q.4 Importance variables - XGBoost A2 - Coût - DDE . . . 169
Q.5 Importance variables - XGBoost A1 - Coût - Incendie . . . 169
Q.6 Importance variables - XGBoost A2 - Coût - Incendie . . . 169

Q.7 Importance variables - XGBoost A1 - Coût - Vol 170
Q.8 Importance variables - XGBoost A2 - Coût - Vol 170
Q.9 Importance variables - XGBoost A1 - Coût - RC 170
Q.10 Importance variables - XGBoost A2 - Coût - RC 170
Q.11 Importance variables - XGBoost A1 - Coût - Toutes Garanties 171
Q.12 Importance variables - XGBoost A2 - Coût - Toutes Garanties 171

Liste des tableaux

1 Comparaison GLM fréquence par garantie . . . 2
2 Comparaison des modèles fréquences par garantie . . . 2
3 Comparaison GLM coût par garantie . . . 3
4 Comparaison des modèles de coût par garantie . . . 3
5 Récapitulatif conclusion par garantie . . . 5
6 Frequency GLM comparison by guarantee . . . 7
7 Frequency models' comparison by guarantee . . . 7
8 Cost GLM comparison by guarantee . . . 8
9 Cost models' comparison by guarantee . . . 8
10 Overview of conclusion by guarantee . . . 10

2.1 Fonction Lien pour les lois usuelles . . . 27

3.1 Sinistres Attritionnels par garantie . . . 41
3.2 Tableau des corrélations entre le nombre et la charge des sinistres par garantie . . . 53

4.1 Tableau des coefficients de sur-dispersion par garantie . . . 54
4.2 Tableau des GLM les plus adaptés par garantie (avant modélisation) . . . 55
4.3 Analyse de type III pour la loi de Poisson et la garantie DDE . . . 55
4.4 Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie BDG . 58
4.5 Tableau des indices de Gini - Comparaison GLM - BDG . . . 59
4.6 Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie DDE . 59
4.7 Tableau des indices de Gini - Comparaison GLM - DDE . . . 60
4.8 Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie Incendie 61
4.9 Tableau des indices de Gini - Comparaison GLM - Incendie . . . 61
4.10 Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie Vol . . 62
4.11 Tableau des indices de Gini - Comparaison GLM - Vol . . . 62
4.12 Performances prédictives des modèles (sur bases apprentissage et validation) pour la garantie RC . . 63
4.13 Tableau des indices de Gini - Comparaison GLM - RC . . . 64
4.14 Tableau des GLM les mieux adaptés, par garantie . . . 64
4.15 Tableau regroupant les différents cp utilisés pour l'élagage - CART . . . 64
4.16 Tableau récapitulatif des MSE sur les bases - CART . . . 65
4.17 Récapitulatif des paramètres optimaux - Random Forest - Fréquence . . . 68
4.18 Variables les plus importantes - RF - Fréquence . . . 68
4.19 Tableau récapitulatif des MSE sur les bases - RF . . . 69
4.20 Tableau de comparaison des indices de Gini (modèle GLM/modèle RF) . . . 69
4.21 Récapitulatif des paramètres optimaux - XGBoost - Fréquence . . . 70
4.22 Variables les plus importantes - XGBoost - Fréquence . . . 70
4.23 Tableau récapitulatif des MSE sur les bases - XGBoost . . . 70
4.24 Tableau de comparaison des indices de Gini (modèle GLM/modèle XGBoost) . . . 70
4.25 Modèles de fréquence sélectionnés, par garantie . . . 71

5.1 Tableau des GLM les plus adaptés par garantie (avant modélisation) . . . 73
5.2 Analyse de type III pour la loi de Log-normale et la garantie DDE . . . 73
5.3 Tableau des indices de Gini - GLM coût . . . 76
5.4 Tableau des RMSE - GLM - Coût - BDG . . . 76

5.5 Tableau des RMSE - GLM - Coût - DDE . . . 77
5.6 Tableau des RMSE - GLM - Coût - Incendie . . . 78
5.7 Tableau des RMSE - GLM - Coût - Vol . . . 79
5.8 Tableau des RMSE - GLM - Coût - RC . . . 80
5.9 Analyse de type III pour la loi de Gamma toutes garanties . . . 81
5.10 Analyse de type III pour la loi Log-normale toutes garanties . . . 81
5.11 Tableau des RMSE - GLM - Coût - Global . . . 81
5.12 Tableau de comparaison des RMSE base de validation (modèle global/spécifique) . . . 82
5.13 Tableau de comparaison des indices de Gini (modèle global/spécifique) . . . 83
5.14 Tableau des GLM les mieux adaptés, par garantie . . . 83
5.15 Tableau regroupant les différentes cp utilisées pour l'élagage - CART . . . 84
5.16 Tableau des RMSE base de validation - CART coût . . . 84
5.17 Tableau des indices de Gini - CART coût . . . 84
5.18 Tableau de comparaison des RMSE base de validation (modèle global CART/GLM) . . . 85
5.19 Tableau de comparaison des indices de Gini (modèle global CART/GLM) . . . 85
5.20 Récapitulatif des paramètres optimaux - Random Forest - Coût . . . 86
5.21 Variables les plus importantes - RF - Coût . . . 87
5.22 Tableau des RMSE base de validation - RF coût . . . 87
5.23 Tableau des indices de Gini - RF coût . . . 87
5.24 Tableau de comparaison des RMSE base de validation (modèle global RF/GLM) . . . 88
5.25 Tableau de comparaison des indices de Gini (modèle global RF/GLM) . . . 88
5.26 Récapitulatif des paramètres optimaux - XGBoost - Coût . . . 89
5.27 Variables les plus importantes - XGBoost - Coût . . . 90
5.28 Tableau des RMSE base de validation - XGBoost coût . . . 90
5.29 Tableau des indices de Gini - XGBoost coût . . . 90
5.30 Tableau de comparaison des RMSE base de validation (modèle global XGBoost/GLM) . . . 91
5.31 Tableau de comparaison des indices de Gini (modèle global XGBoost/GLM) . . . 91

6.1 Récapitulatif des profils par garantie . . . 103

H.1 Tableau des indicateurs avant modélisation pour la garantie BDG . . . 119
H.2 Tableau des indicateurs avant modélisation pour la garantie DDE . . . 119
H.3 Tableau des indicateurs avant modélisation pour la garantie Incendie . . . 119
H.4 Tableau des indicateurs avant modélisation pour la garantie Vol . . . 119
H.5 Tableau des indicateurs avant modélisation pour la garantie RC . . . 120
H.6 Analyse de type III pour la loi de Poisson et la garantie Vol . . . 120
H.7 Analyse de type III pour la loi Binomiale Négative et la garantie Vol . . . 120
H.8 Analyse de type III pour la loi de Poisson et la garantie BDG . . . 121
H.9 Analyse de type III pour la loi Binomiale Négative et la garantie BDG . . . 121
H.10 Analyse de type III pour la loi de Poisson et la garantie RC . . . 121
H.11 Analyse de type III pour la loi Binomiale Négative et la garantie RC . . . 121
H.12 Analyse de type III pour la loi de Poisson et la garantie Incendie . . . 122
H.13 Analyse de type III pour la loi Binomiale Négative et la garantie Incendie . . . 122
H.14 Analyse de type III pour la loi Binomiale Négative et la garantie DDE . . . 122

M.1 Tableau des indicateurs avant modélisation pour la garantie BDG . . . 144
M.2 Tableau des indicateurs avant modélisation pour la garantie DDE . . . 144
M.3 Tableau des indicateurs avant modélisation pour la garantie Incendie . . . 144
M.4 Tableau des indicateurs avant modélisation pour la garantie Vol . . . 144
M.5 Tableau des indicateurs avant modélisation pour la garantie RC . . . 145
M.6 Tableau des indicateurs avant modélisation toutes garanties . . . 145
M.7 Analyse de type III pour la loi Gamma et la garantie DDE . . . 145
M.8 Analyse de type III pour la loi Gamma et la garantie BDG . . . 145
M.9 Analyse de type III pour la loi de Log-normale et la garantie BDG . . . 145
M.10 Analyse de type III pour la loi Gamma et la garantie Incendie . . . 146
M.11 Analyse de type III pour la loi de Log-normale et la garantie Incendie . . . 146
M.12 Analyse de type III pour la loi Gamma et la garantie Vol . . . 146

M.13 Analyse de type III pour la loi de Log-normale et la garantie Vol 146
M.14 Analyse de type III pour la loi Gamma et la garantie RC 146
M.15 Analyse de type III pour la loi de Log-normale et la garantie RC 146