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

Par : **Omar SAME**

Titre du mémoire : Enrichissement des données et modélisation fine du coût des sinistres dégât des eaux habitation par le biais des rapports d'expertise

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

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

Membres présents du jury de la filière : Signature :	Entreprise : GENERALI IARD Nom Entreprise régie par le Code des Assurances 552 062 663 R.C.S. PARIS Siège Social : 2, rue Pillet-Will Signature : 75009 Paris
	Directeur de mémoire en entreprise Nom : **Ubezzi Robin** Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

Résumé :

La garantie dégât des eaux chez Generali se présente actuellement comme étant la plus sinistrée de la branche multirisque habitation. La dégradation observée ces dernières années sur cette garantie est principalement portée par les sinistres expertisés qui apparaissent stables en nombre mais occupent de plus en plus un poids prépondérant au niveau de la charge sinistre. Cette étude rentre ainsi dans le cadre de l'amélioration de la prévision de ces sinistres qui engendrent les indemnités les plus élevées sur cette garantie.

La spécificité de ces travaux a ainsi consisté à créer et mettre à disposition une base structurée détaillant poste par poste le coût des sinistres alors qu'initialement la seule information connue était le coût total du sinistre. Ce découpage de la sinistralité a ensuite permis de proposer une segmentation beaucoup plus fine dans l'explication du coût en dégât des eaux. Dans un premier temps, l'objectif de ce mémoire sera de permettre à Generali d'enrichir ses données en dommage aux biens afin d'aller à une maille beaucoup plus fine et de mieux comprendre les sinistres. Pour ce faire, l'obtention de la base de données est passée tout d'abord par la création d'un outil d'automatisation de la lecture des informations contenues dans les rapports faisant suite à l'évaluation de l'expert. Cet algorithme permet aujourd'hui à l'entreprise d'automatiser la construction d'une base structurée décomposant poste par poste le coût des dommages réglés selon le type de poste : Embellissement, Immobilier, Mobilier et Autres. Un travail important a également dû être effectué sur la qualité des données d'expertise pour aboutir à leur fiabilisation grâce à un processus que nous détaillerons dans le corps de ce mémoire. A la suite de la construction de la base d'étude regroupant les données risques, les données sinistres et les données rapports, des travaux concernant le retraitement et l'analyse des données ont été effectués. C'est dans ce sens que les données manquantes ont été complétées en partie par les données d'expertise dans la limite du possible avant d'appliquer la méthode des K plus proches voisins. Par ailleurs, dans le cadre de la classification, certaines variables présentaient beaucoup trop de modalités ou certaines catégories avaient un effectif trop faible, deux méthodes d'ajustement ont été testées et comparées : celle des KMeans et celle consistant à réaliser des regroupements à dire d'expert.

Une fois cette étape d'enrichissement, de nettoyage et de mise en cohérence de la base de données finalisée, nous proposons dans la seconde partie de ces travaux une approche de modélisation fine des sinistres dégât des eaux, en évaluant désormais le coût moyen des dommages poste par poste. La modélisation portera sur les deux principaux postes et a été réalisée dans un premier temps à l'aide du modèle linéaire généralisé correspondant à l'approche actuarielle « traditionnelle » ; elle sera dans un second temps challengée par un modèle d'apprentissage avancé, le Random Forest. Des analyses comparatives entre les deux approches de modélisation seront par la suite effectuées et présentées.

Mots-clés :

GLM, Arbres de décision, Random Forest, DAB (Dommages Aux Biens), Rapport d'expertise, MRH (Multirisque habitation), KNN, KMeans, Embellissement, Immobilier

Abstract :

Generali's water damage coverage is currently the most affected of the comprehensive home insurance branch. The deterioration observed over the last few years in this coverage is mainly due to claims that have been appraised, which appear to be stable in number, but are becoming more and more preponderant in terms of claims costs. This study is therefore part of the improvement of the forecasting of these claims which generate the highest indemnities for this coverage.

The specificity of this work consisted in creating and making available a structured database structured database detailing the cost of claims item by item, whereas initially the only information was the total cost of the claim. This breakdown of the claims experience has allowed us to propose a much finer segmentation in the explanation of the cost in water damage. Initially, the objective of this thesis will be to allow Generali to enrich its data in property damage in order to go to a much finer mesh and to better understand the claims. In order to do this, the database was first obtained by creating a tool to automate the reading of the information contained in the reports following the expert's evaluation. This algorithm now allows the company to automate the construction of a structured database that breaks down the cost of the damages paid, item by item according to the type of item : Beautification, Real Estate, Furniture and Other. Important work also had to be carried out on the quality of the expertise data in order to to make them more reliable, thanks to a process that we will detail in the detailed in the body of this report. Following the construction of the study database containing risk data, claims data and reports data, work was carried out on the reprocessing and analysis of the data. To this end, the missing data were partly completed by the expert data within the limits of what was possible before applying the K-nearest neighbor method. In addition, in the context of the classification, some variables had too many modalities or some categories had too few members. Two adjustment methods were tested and compared : the K-means method and the method of grouping by expert opinion.

Once this stage of enrichment, cleaning and consistency of the database is finalized, we propose in the second part of this work, we propose an approach of fine modeling of water damage claims, by evaluating the average cost of damage item by item. The modeling will focus on the two main items and has been carried out initially using the generalized linear model corresponding to the traditional actuarial approach; it will be challenged in a second phase by a challenged by an advanced learning model, the Random Forest. Comparative analyses comparative analyses between the two modeling approaches will then be performed and presented.

Mots-clés :

GLM, Decision trees, Random Forest, DAB (Damage to property), Expertise report, MRH (Multi-risk Home Insurance), KNN (K-Nearest Neighbor), KMeans, Embellishment, Real estate.

Note de synthèse :

Contexte et objectif :

Cette étude s'inscrit dans le cadre d'un projet plus large consistant à enrichir les données sur tout le dommage aux biens (DAB) et à réaliser différentes études actuarielles dans le cadre du pilotage de l'exercice. En effet, l'enrichissement des données s'avère être un enjeu de plus en plus majeur en assurance. Le manque d'informations détaillant la sinistralité continue d'être un point bloquant à la mise en place de plusieurs études actuarielles. C'est ainsi que la plupart des assureurs en IARD font face à cette problématique ne favorisant pas la compréhension détaillée des sinistres. Dans ce contexte, ce mémoire exploite les données contenues dans les rapports d'expertise afin d'apporter de nouvelles informations et d'offrir une vision précise des indemnisations allouées à la suite d'un sinistre. Le but de ce mémoire sera d'abord de répondre à ce manque d'informations pour enrichir les données préexistantes et apporter une vision plus détaillée concernant la sinistralité du portefeuille. Cela permettra par la suite de proposer une approche d'évaluation fine du coût d'un sinistre expertisé en dégât des eaux MRH (Multirisque Habitation). L'obtention de la base recensant les informations d'expertise a été possible en automatisant un outil qui a pour vocation l'extraction du contenu des rapports ainsi que le traitement et la fiabilisation des données. Cela permet aujourd'hui à Generali de disposer d'une base automatisée présentant plusieurs données d'expertises qui seront exploitées dans le cadre de différentes études. Mais aussi, elle nous a permis d'atteindre une segmentation très fine dans l'explication du coût d'un sinistre expertisé. L'objectif final de cette étude est ainsi d'affiner l'évaluation de ces sinistres et d'offrir une vision précise du coût des dommages indemnisés par le biais d'une approche moyenne du portefeuille. Ces études statistiques serviront ainsi à la politique de maîtrise des coûts de la compagnie d'assurance, mais aussi, associées au modèle de fréquence mis en oeuvre par le reste de l'équipe, seront utiles dans le cadre du pilotage de la charge.

Périmètre d'analyse :

L'étape d'enrichissement des données a été réalisée sur tout le DAB. Par la suite, notre étude concernera principalement les sinistres dégât des eaux multirisque habitation. Toutefois, les autres garanties suivront le même raisonnement.

Base d'étude :

La création de la base d'étude a nécessité la mise en place de plusieurs tables intermédiaires : la base sinistre, la base risque, la base rapport et la base des dépenses. Par la suite, une jointure successive a été réalisée afin de mettre en place la base finale qui sera retraitée puis utilisée dans le cadre des travaux effectués.

Création et description de la base de données rapports :

Une étape déterminante était la création de la base d'étude et qui servira à la modélisation. La création de cette base a nécessité la collecte des données d'expertise. Ainsi, l'étude a été faite sur une profondeur historique de 9 ans et concerne les rapports d'expertise du cabinet Saretec.

Ces rapports ont été collectés majoritairement sur les produits MRH, MRI et MRC. Dans le cadre de nos travaux, la lecture d'une centaine de rapports a été nécessaire afin de bien comprendre ces documents et d'identifier les différents éléments à extraire. Cette étape permettait également de se faire une idée sur la façon de créer un programme qui généralisera au mieux l'extraction des données sur l'ensemble des rapports.

Analyse et retraitement des données :

La préparation des données est une étape importante dans le processus de modélisation à ne surtout pas négliger. Ainsi, un travail important a été réalisé sur la qualité des données. Après l'obtention de la base via l'exécution du script créé sur l'ensemble des rapports, une attention particulière a été accordée au traitement des données obtenues notamment :

- **L'identification des rapports n'ayant pas fait l'objet d'une réelle indemnisation par Generali** : En effet, certains rapports peuvent présenter des montants sans pour autant avoir fait l'objet d'un règlement par Generali. C'est dans ce sens que dans le cadre de la fiabilisation de notre base, nous étions amenés à automatiser l'identification de ces rapports afin de conserver dans l'étude que ceux reflétant la réalité des prestations indemnisées par la compagnie.
- **Identification de la dernière vision des rapports :** Certains sinistres présentent plusieurs rapports du fait qu'un sinistre peut évoluer dans le temps. Il s'est donc avéré nécessaire de conserver la vision la plus récente de l'évaluation.

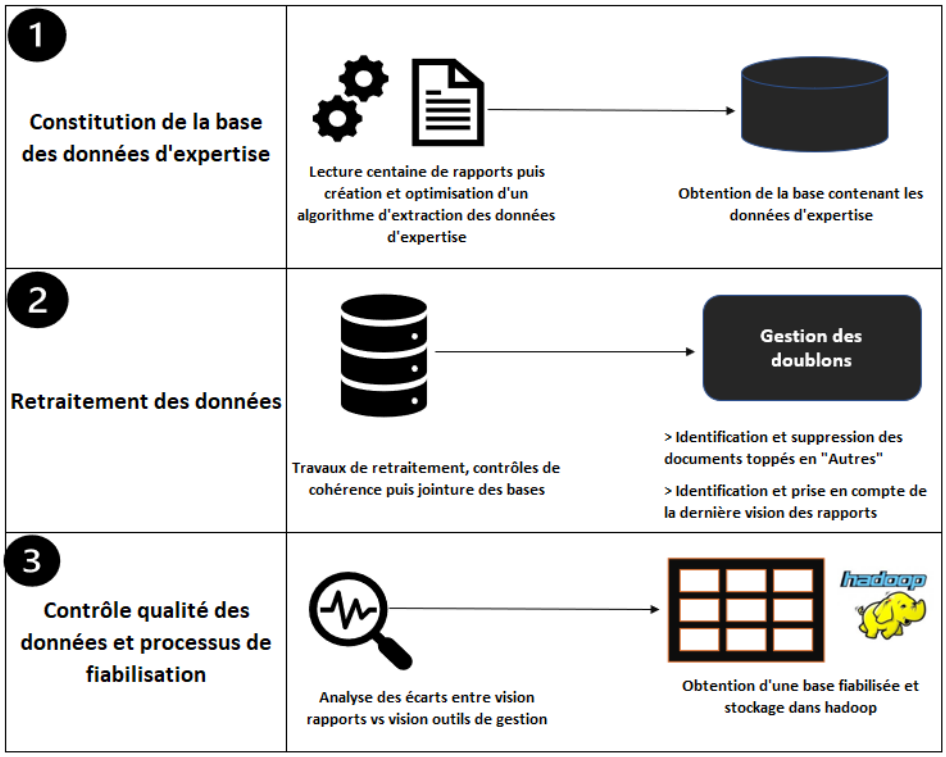
Par ailleurs, nous avons également réalisé des contrôles de cohérence. Par exemple, nous avons cherché à respecter la règle qui suit au niveau des dates :

$$date\ de\ survenance \preceq date\ du\ rapport \preceq date\ de\ règlement \preceq date\ clôture$$

Les dates de rapport antérieures à la survenance du sinistre ou se situant après le règlement ou la clôture du sinistre ont été supprimées compte tenu de leur faible matérialité.

Fiabilisation des données rapports :

Après les étapes de retraitement ci-dessus, la phase suivante consistait à fiabiliser les montants extraits des rapports. Le croisement avec la table des dépenses/recettes a permis de créer un script de fiabilisation des données. Cette étape a permis de voir que certains rapports ne reflètent pas toujours la réalité des indemnisations de Generali. En effet, après plusieurs travaux, investigations ainsi que des échanges avec les équipes de gestion, nous avons pu déterminer les éventuels cas responsables de ces non alignements avec les données présentes dans les outils au travers de leur expertise métier. Ces différents travaux ont ainsi permis d'aboutir à l'automatisation d'un outil fournissant une base structurée permettant de segmenter le coût des sinistres à partir des données d'expertise. Ci-dessous, une illustration de la démarche adoptée dans le cadre de sa mise en place :

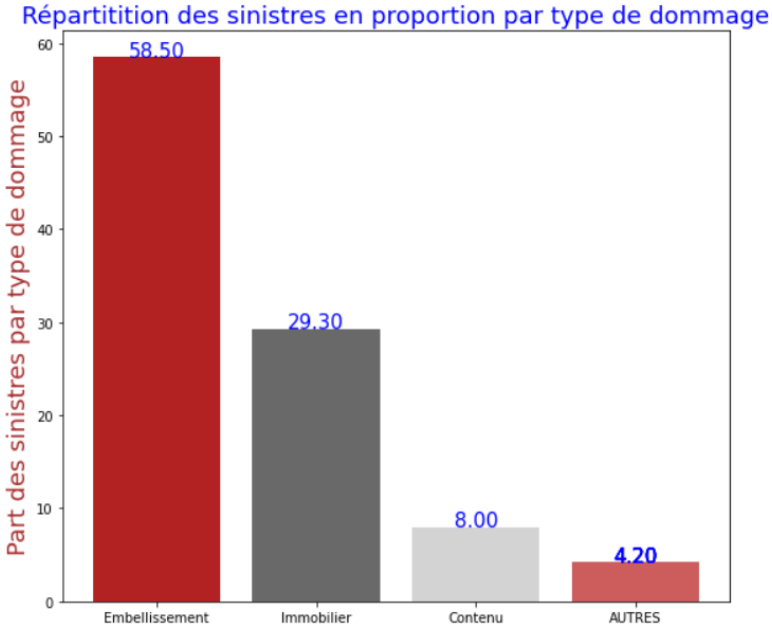


Discrétisation et traitement des données manquantes :

Un traitement des données manquantes a été également effectué dans cette étude pour limiter les biais qu'elles peuvent introduire dans l'étape de modélisation. C'est la raison pour laquelle, certaines variables explicatives ont été complétées en partie par des données rapports à l'issue d'une extraction dans la mesure du possible, avant d'appliquer la méthode des K-plus proches voisins (KNN) pour assurer la complétion totale. Par ailleurs, dans le cadre de la classification, certaines variables présentaient beaucoup trop de modalités ou certaines catégories avaient un effectif trop faible. C'est dans ce sens que nous avons testé deux méthodes d'ajustement : la méthode des KMeans et celle consistant à réaliser des regroupements à dire d'expert.

Modélisation :

Les types de dommage étudiés seront l'embellissement et l'immobilier. Il s'agit des deux postes de dommage les plus sinistrés en terme de charge et de nombre reflétant 94% des dommages indemnisés. Voici la répartition de la sinistralité en nombre par poste :



Les deux derniers dommages sont très peu sinistrés pour cette garantie et présentent une faible volumétrie dans notre base d'étude. Ils correspondent moins de 6% des coûts indemnisés sur l'ensemble de la base d'étude. Dans le cadre de la modélisation, deux approches seront mises en oeuvre : le GLM correspondant à l'approche actuarielle traditionnelle et le Random forest correspondant à une approche de modélisation avancée.

1 - La mise en place du modèle GLM s'articule en 3 étapes :

- Choix de la loi paramétrique
- Processus de sélection des variables
- Analyse des résultats et validation du modèle

Dans le cadre de la sélection des variables, une étude pré-sélective a été faite grâce à une analyse univariée. Par la suite, une approche de sélection automatique a été réalisée (la méthode stepwise) combinée à une analyse de minimisation de la déviance à l'ajout successif de variables. En parallèle, des tests de type III sont effectués pour les variables les moins significatives pour le modèle afin de tester l'hypothèse d'apport d'information de ces variables.

2 - Dans le cadre de la modélisation par la méthode du Random Forest nous nous sommes focalisés sur 3 principaux paramètres :

- Le nombre d'arbres de la forêt
- Le nombre de variables testées
- La profondeur des arbres

Pour l'estimation des paramètres, une méthode de validation croisée a été utilisée (k-fold cross validation) ainsi que la technique GridSearch pour leur optimisation.

Pour étudier les modèles, les indicateurs qui seront utilisés sont le RMSE, le MAE et l'indice de Gini. Nous ferons également, une comparaison entre valeurs prédites et observées dans le cadre de l'évaluation de la qualité des prédictions.

1- Modélisation du poste embellissement :

La modélisation du poste embellissement a été réalisé avec un GLM et un Random Forest. Les résultats suivants sont obtenus :

Méthode	RMSE test	MAE test	Gini test
GLM	728.25	518	25.02%
RF	725.58	526	24.83%

La performance des deux approches de modélisation a été évaluée grâce aux critères tels que le RMSE, le MAE et l'indice de Gini. L'analyse comparative montre que les deux méthodes présentent des performances comparables. Au sens du RMSE, les résultats laissent à croire que le GLM aura tendance à être légèrement moins efficace que le Random Forest pour les valeurs extrêmes. Toutefois, au sens du MAE qui est un indicateur mesurant les écarts par observation, le GLM performe mieux. De même, au sens de l'indice de Gini, la segmentation du risque est sensiblement de meilleure qualité pour ce dernier. Pour cette raison, il a été choisi pour la modélisation du coût moyen de l'embellissement. Par ailleurs, l'analyse comparative entre valeurs prédites et observées a permis de valider la capacité du modèle à garder tout son sens de généralisation avec des prédictions cohérentes avec les observations. Les variables qui discriminent plus le modèle sont le nombre de pièces et le zonier dégât des eaux.

2-Modélisation du poste immobilier :

De la même façon que l'embellissement, le dommage à l'immobilier a été modélisé suivant les deux approches précédentes : GLM et Random Forest. Les résultats se présentent comme suit :

Méthode	RMSE test	MAE test	Gini test
GLM	928	787	19.29%
RF	914	768	19.38%

Une fois de plus, les deux modèles présentent des performances comparables. Toutefois, contrairement à l'embellissement le Random Forest performe mieux que le GLM pour la modélisation du poste immobilier sur le jeu de test. L'analyse de la prédiction entre valeurs prédites et observées montre que les modèles (notamment le GLM) ont parfois tendance à un peu sur-évaluer les modalités avec peu d'observations. Ce phénomène de sur-estimation locale peut s'expliquer par le fait que le modèle ne capture pas totalement tous les effets risques de ces classes. Néanmoins, plus d'observations devraient permettre d'affiner la prédiction. Par ailleurs, les facteurs contribuant le plus à l'explication du coût moyen du dommage immobilier sont le statut de l'assuré et le nombre de pièces dans l'habitation.

Executive summary :

Context and objective :

This study is part of a larger project consisting in enriching the data on all property damage (DAB) and in carrying out various actuarial studies within the framework of the exercise's management. Indeed, the enrichment of data is proving to be an increasingly major issue in insurance. The lack of information detailing the loss experience continues to be a blocking point in the implementation of several actuarial studies. This is why most PC insurers are faced with this problem, which does not favor a detailed understanding of claims. In this context, this thesis exploits the data contained in the expert reports in order to provide new information and to offer a precise vision of the indemnifications allocated following a loss. The goal of this thesis will be first to answer this lack of information to enrich the pre-existing data and provide a more detailed vision of the claims experience of the portfolio. This will then allow us to propose an approach for the fine evaluation of the cost of a water damage claim appraised by an MH (Multirisque Habitation) expert. Obtaining the database containing the expertise information was made possible by automating a tool whose purpose is to extract the content of the reports as well as the processing and reliability of the data. Today, this allows Generali to have an automated database presenting several expert data that will be used in the framework of different studies. But also, it allowed us to reach a very fine segmentation in the explanation of the cost of an appraised claim. The final objective of this study is thus to refine the evaluation of these and to offer a precise vision of the cost of the compensated damages through an average average approach to the portfolio. These statistical studies will thus serve the cost control policy of the cost control policy of the insurance company, but also, in combination with the frequency model implemented by the rest of the team, will be useful in the framework of load management.

Perimeter of analysis :

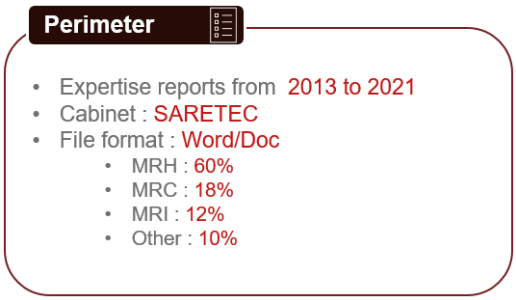
The data enrichment stage was carried out on the entire DAB. In the following, our study will mainly concern water damage claims for multi-risk housing. However, the other guarantees will follow the same reasoning.

The study data :

The creation of the study base required the setting up of several intermediary tables tables: the claim base, the risk base, the report base and the expense base. Subsequently, a successive the final database that will be reprocessed and used in the work carried out used in the work carried out.

Creation and description of the reports database :

A key step was the creation of the study database that will be used for modeling. The creation of this base required the collection of expertise data. Thus, the study was made on a historical depth of 9 years and concerns the expertise reports of the firm Saretec.



The reports were collected mainly on HRM, IRM and CRM products. In the course of our work, it was necessary to read about 100 reports in order to fully understand these documents and to identify the different elements to be extracted. This step also allowed us to get an idea of how to create a program that will best generalize the extraction of data on all the reports.

Data analysis and reprocessing :

Data preparation is an important step in the modeling process that should not be neglected. Thus, a lot of work has been done on data quality. After obtaining the database via the execution of the script created on all the reports, particular attention was paid to the processing of the data obtained, in particular :

— **The identification of reports that have not been the subject of a real compensation by Generali** : Indeed, some reports may present amounts without having been settled by Generali. It is in this sense that in the context of the reliability of our database, we had to automate the identification of these reports in order to keep in the study only those reflecting the reality of the benefits compensated by the company.

— **Identification of the last vision of the reports :** Some claims have multiple reports because a claim can change over time. It was therefore necessary to keep the most recent view of the evaluation.

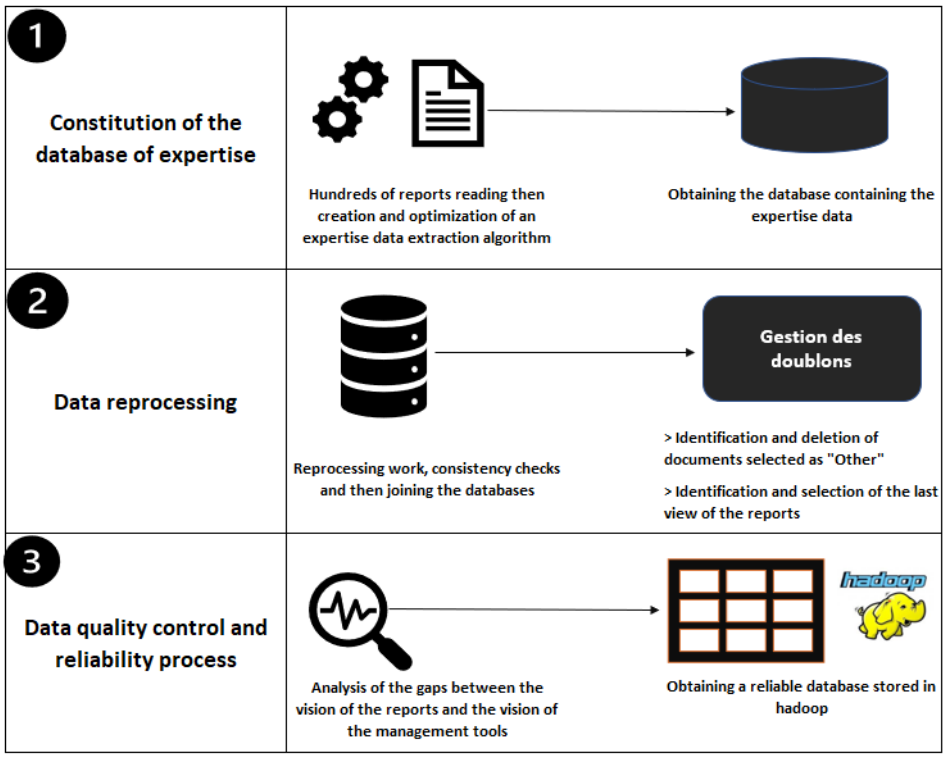
In addition, we also performed consistency checks. For example, we tried to respect the following rule for dates:

$$date\ of\ occurrence \preceq date\ of\ the\ repport \preceq date\ of\ payment \preceq closing\ date$$

Reporting dates prior to the occurrence of the claim or after the settlement or closure of the claim have been removed due to their low materiality.

Reliability treatments :

After the above reprocessing steps, the next phase consisted in making the amounts extracted from the reports reliable. The cross-referencing with the table of expenses/revenues made it possible to create a script to make the data reliable. This step allowed us to see that some reports do not always reflect the reality of Generali's indemnifications. Indeed, after several works, investigations as well as exchanges with the management teams, we were able to determine the possible cases responsible for these non alignments with the data present in the tools through their business expertise. This work led to the automation of a tool that provides a structured base for segmenting the cost of claims based on expert data. Below is an illustration of the approach adopted in the context of its implementation :

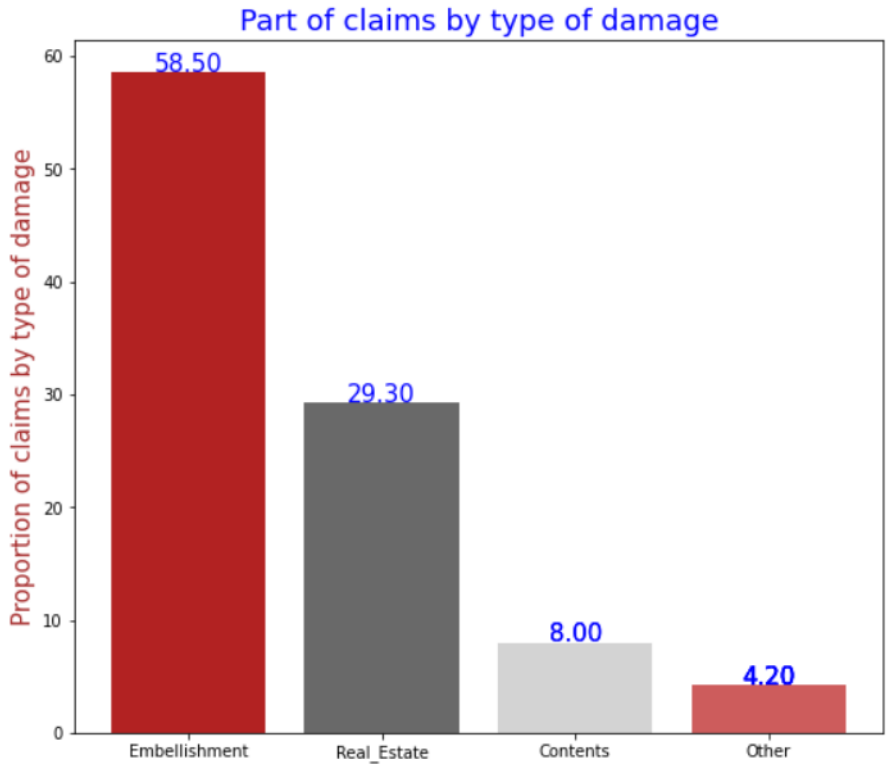


Discretization and treatment of missing data :

Missing data were also treated in this study to limit the biases that they may introduce in the modeling stage. For this reason, some explanatory variables were partially completed with data reported after extraction whenever possible, before applying the K-nearest neighbor (KNN) method to ensure total completion. completion. In addition, in the classification, some variables had too many modalities or some categories were too small in number. It is in this sense that we tested two adjustment methods We tested two adjustment methods: the KMeans method and the method of grouping by expert opinion.

Modeling :

The types of damage studied will be beautification and real estate. These are the two most affected damage items in terms of load and number reflecting 94% of the amounts allocated for damages.



Indeed, the last two damages are very little claimed for this coverage and present a low volume in our study base. They correspond to less than 6% of the indemnified costs on the whole study base. In the context of the modeling, two approaches will be used: the GLM corresponding to the traditional actuarial approach and the Random forest corresponding to an advanced modeling approach.

1 - The implementation of the GLM model consists of 3 steps :

- Choice of parametric law
- Process of selecting variables
- Analysis of the results and model validation

In the context of variable selection, a pre-selection study was carried out using a univariate analysis. Subsequently, an automatic selection approach was carried out (the stepwise method) combined with a deviance minimization analysis with the successive addition of variables. In parallel, type III tests are performed for the least significant variables in the model in order to test the hypothesis of information contribution of these variables.

2 - In the framework of the modeling by the Random Forest method we focused on 3 main parameters :

- Number of trees in the forest
- Number of variables tested
- The depth of trees

For the estimation of the parameters, a cross validation method was used (k-fold cross validation) as well as the GridSearch technique for their optimization. To study the models, the indicators that will be used are the RMSE, the MAE and the Gini index. We will also make a comparison between predicted and observed values in order to evaluate the quality of the predictions.

1- Embellishment modeling :

The modeling of the embellishment item was performed with a GLM and a Random Forest. The following results are obtained :

Method	RMSE test	MAE test	Gini test
GLM	728.25	518	25.02%
RF	725.58	526	24.83%

The performance of the two modeling approaches was evaluated using criteria such as RMSE, MAE and Gini index. The comparative analysis shows that the two methods have comparable performance. In the RMSE sense, the results suggest that the GLM will tend to be slightly less effective than the Random Forest for extreme values. However, in the sense of the MAE, which is an indicator measuring the deviations per observation, the GLM performs better. Similarly, in the sense of the Gini index, the risk segmentation is significantly better for the latter. For this reason, it was chosen for modeling the average cost of beautification. Furthermore, the comparative analysis between predicted and observed values validated the ability of the model to keep its generalization meaning with predictions consistent with observations. The variables that discriminate more the model are the number of rooms and the water damage zone.

2- Real Estate Modeling :

Similarly to beautification, property damage was modeled following the two previous approaches: GLM and Random Forest. The results are as follows :

Method	RMSE test	MAE test	Gini test
GLM	928	787	19.29%
RF	914	768	19.38%

Once again, the two models show comparable performance. However, contrary to the embellishment, the Random Forest performs better than the GLM for the modeling of the real estate item on the test set. The analysis of the prediction between predicted and observed values shows that the models (in particular the GLM) have sometimes a tendency to overestimate the modalities with few observations. This phenomenon of local over-estimation can be explained by the fact that the model does not fully capture all the risk effects of these classes. Nevertheless, more observations should allow to refine the prediction. On the other hand, the factors contributing most to the explanation of the average cost of property damage are the status of the insured and the number of rooms in the home.

Remerciements :

Je tiens tout d'abord à exprimer toute ma reconnaissance à mon tuteur entreprise Robin Ubezzi responsable d'études d'actuariat pour m'avoir accompagné tout au long de l'alternance et de ce mémoire. Mais aussi pour son soutien, sa bienveillance et ses encouragements durant la construction de cette étude.

Je tiens à remercier toute l'équipe Étude Indemnisations et tout particulièrement le manager VIEU Jean-Sébastien pour m'avoir intégré dans son équipe dans le cadre de la réalisation de mon alternance. Mais aussi pour sa gentillesse et sa bonne humeur au quotidien.

J'exprime ma gratitude à tous les membres du Bureau d'Études Techniques Non-Vie. Plus particulièrement To Vong Nguyen, Nicolas Martineau, Diallo Abdourahmane, Margaux Limbergère et Lin Océane pour la relecture, les remarques pertinentes ainsi que leurs conseils avisés dans cette étude. Mais aussi Mathieu Wolf pour l'aide qu'il m'a apporté durant ces travaux ainsi que Yassine Laghzali et Xin Zhang qui ont su m'offrir de leur temps afin de mieux appréhender la branche MRH à travers des échanges enrichissants.

Mes remerciements vont à l'encontre de tout ceux avec qui j'ai pu collaborer dans le cadre de ce projet plus particulièrement Ambre Le Stum et Aissaoui Ghada. Mais aussi, Christel Even pour avoir partagé avec nous sa connaissance des rapports d'expertise.

J'aimerais aussi exprimer ma reconnaissance envers mon tuteur académique Thomas Debais pour sa relecture, ses différents conseils et remarques. Un grand merci également à Olivier Lopez pour ses précieux conseils et le temps qu'il a su m'accorder durant ce mémoire ainsi que tout le corps professoral de l'ISUP.

Pour finir, je tiens à remercier mes parents et toute ma famille pour leurs encouragements tout au long de ces années d'études ainsi que leur soutien indéfectible.

Table of Contents

Introduction :

Les compagnies d'assurance IARD font face à un contexte de manque d'informations sur leur sinistralité rendant ainsi difficile la compréhension et le suivi détaillé des sinistres. Ainsi, à la suite d'un sinistre, la plupart des assureurs n'ont pas une grande visibilité sur le découpage de la sinistralité et sur ce qui est réellement indemnisé. Ce mémoire s'attardera principalement sur les sinistres expertisés de la garantie dégâts des eaux multirisque habitation. En effet, le coût de ces sinistres expertisés est de plus en plus prépondérant, impactant de plus en plus la charge pour cette garantie. Pour atteindre cet objectif, nous avons d'abord créé un outil permettant d'aller récupérer les données rapports. Cela nous a permis de mettre à disposition une base de données structurée incluant de nouvelles variables servant à enrichir les bases de données préexistantes et détaillant poste par poste la décomposition des sinistres. Cette base permet aujourd'hui d'atteindre une granularité très fine et d'exploiter des données jusque-là pas réellement utilisées. Dans un second temps, nous proposons une approche de modélisation fine (par poste) dans le cadre de l'explication du coût des sinistres ayant fait l'objet d'une expertise.

La première partie de ce mémoire sera axée sur la présentation de la garantie dégât des eaux du produit multirisque habitation, la contextualisation du sujet, le périmètre et les enjeux de l'étude. Par ailleurs nous décrirons la chronologie d'un sinistre expertisé et présenterons à l'occasion les informations contenues dans un rapport d'expertise.

La deuxième partie sera consacrée à la présentation des différents types de dommage identifiés suite à l'extraction des données rapports. Cette partie détaillera aussi la mise en place de la base de données d'étude (une étape cruciale dans le cadre de nos travaux) et présentera les différentes jointures et retraitements ainsi que les statistiques concernant la collecte des données rapports. Nous détaillerons les travaux et contrôles opérés menant à la fiabilisation des données utilisées. Mais également, présenterons les différentes statistiques descriptives réalisées ainsi que le retraitement des données effectué par le biais de techniques de machine learning telles que les KMeans ou encore les K plus proches voisins.

Ensuite, la troisième partie quant à elle, expliquera le coût moyen des sinistres relatif à l'embellissement et à l'immobilier via l'usage de modèles tels que les modèles linéaires généralisés et le Random Forest. Et enfin, nous conclurons cette partie en mettant l'accent sur le bilan des résultats, les limites des travaux ainsi que les axes d'amélioration.

Part I

Périmètre d'étude et problématisation

1 GENERALI FRANCE

1.1 Présentation de Generali

Le groupe Generali est une entreprise d'origine Italienne spécialisée dans les assurances multirisques fondée en 1931. L'entreprise a su se démarquer au fil des années par la fusion et création de ses nombreuses entitées. Parmi celles-ci, on retrouve notamment la fondation d'Europ Assistance en 1963, qui a révolutionne le secteur de l'assurance. Ce service est par ailleurs présent aujourd'hui dans pas moins de 30 pays. De plus, en 1990, Generali innove encore en fondant Genertel, l'une des premières compagnies spécialisées dans la vente de produits d'assurances à distance.

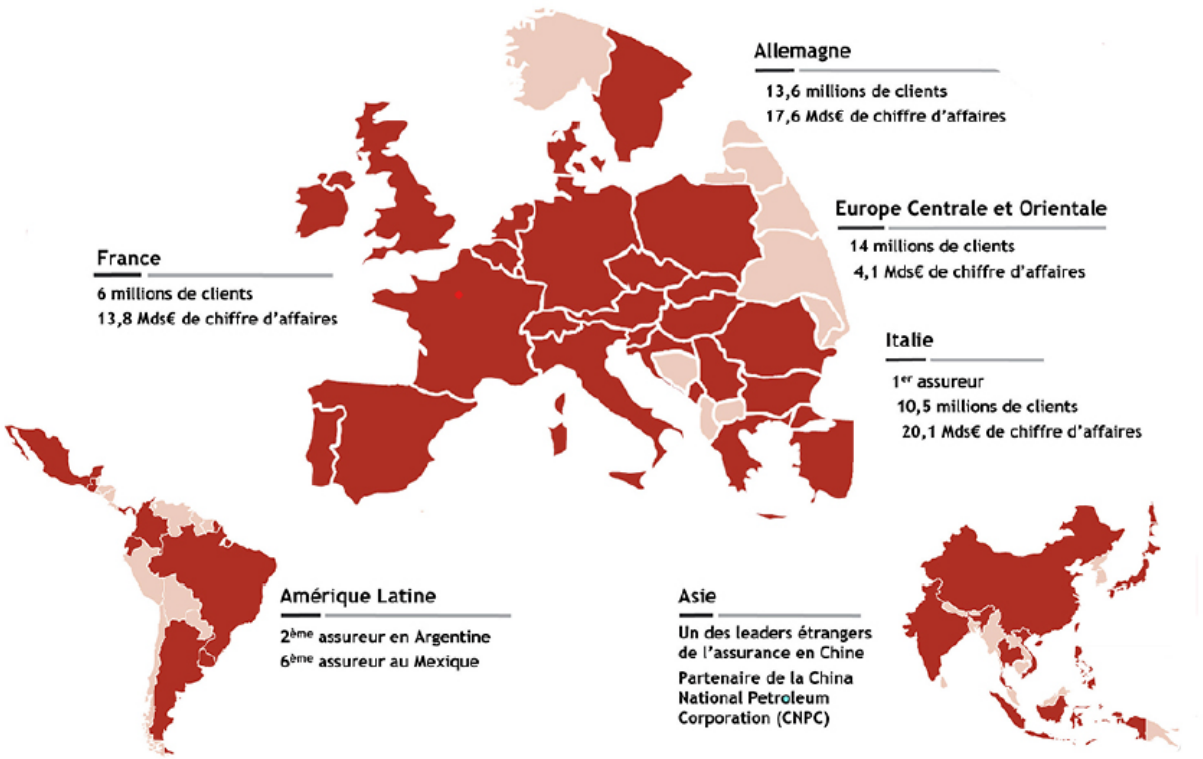


Figure 1 – Generali dans le monde

Aujourd'hui, Generali est présent dans plus de 50 pays avec plus de 61 millions de clients et 71 000 collaborateurs. L'entreprise couvre un large panel de produits d'assurances vie (épargne, retraire...) et non-vie (voiture, bâtiment...). C'est à ce jour est un des leader mondial dans le monde de l'assurance.

1.2 Présentation du service - Études Indemnisations

L'indemnisation est une somme d'argent versée par l'assureur à la suite d'un sinistre pour dédommager l'assuré touché en échange des cotisations payées par ce dernier.
Le service Études Indemnisation se trouve au sein de la direction « Technique Assurance IARD ». Cette direction a pour principal but de tarifer les contrats de toutes les branches IARD. Pour tarifer au prix le plus juste il faut comprendre au mieux son coût de sinistre. C'est le rôle du service Études Indemnisation. L'équipe pilote la charge sinistre IARD notamment en assurant un suivi de la sinistralité, des coûts moyens, du provisionnement dossier par dossier des sinistres à enjeux ainsi qu'un suivi et une projection de l'impact des évènements climatiques. Il est donc nécessaire pour l'équipe de mettre en oeuvre des analyses permettant de mieux comprendre les évolutions observées et de créer des indicateurs sur les différentes branches IARD.
L'équipe travaille en collaboration avec de nombreux services:

- Les autres équipes afin de mieux appréhender la charge sinistre par branche
- Les équipes du BET IARD afin de mieux appréhender la charge sinistre par branche
- L'équipe indemnisation IARD afin de faire le lien entre la vision chiffrée et la vision opérationnelle

2 Présentation du secteur de l'assurance et plus particulièrement l'assurance multirisque habitation :

Le contrat d'assurance est une convention par laquelle l'assureur s'engage à verser à l'assuré une somme d'argent pour réparer un préjudice subi en cas de survenance d'un sinistre en échange du paiement d'une somme versée. Ainsi, on distingue les assurances de personnes, regroupant les contrats d'assurance santé et les assurances vie. Mais aussi les assurances IARD, regroupant les assurances de biens et responsabilités.

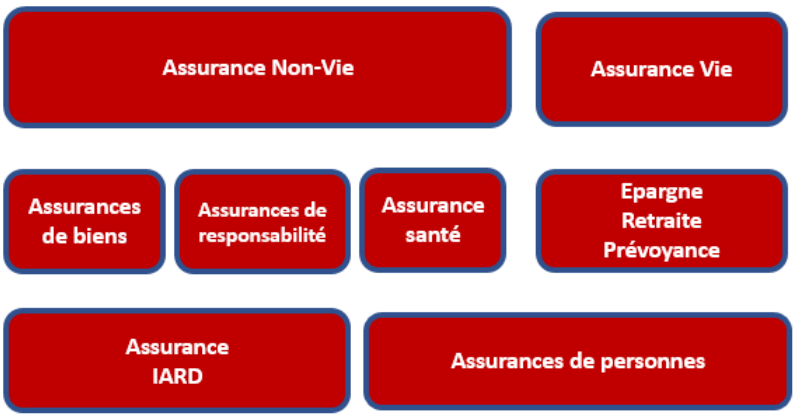


Figure 2 – Les catégories d'assurance

2.1 Le produit multirisque habitation :

Parmi les garanties existantes en assurance de dommage, il existe la garantie multirisque habitation qui fait partie des assurances indispensables à la vie quotidienne. Cette offre permet de protéger le logement, ses occupants et les biens matériels contre d'éventuels accidents en contrepartie d'un versement d'une prime. L'assurance habitation est obligatoire pour les locataires de logement meublé ou non meublé qui doivent au minimum souscrire à l'assurance responsabilité civile sous peine d'expulsion. Toutefois, étant propriétaire, elle n'est pas obligatoire mais fortement conseillée pour au moins s'assurer d'une indemnisation en cas de sinistre.
Le contrat **multirisque habitation** permet ainsi à l'assuré de protéger à la fois son patrimoine aussi bien mobilier qu'immobilier.

2.2 Statistiques marché en France :

Le marché de l'assurance habitation est un marché plus que jamais convoité. Elle permet d'assurer des personnes ainsi que leurs biens et occupe de plus en plus une part importante dans le quotidien des personnes. Nous pouvons observer au niveau du graphe ci-dessous quelques statistiques marché concernant la gestion des sinistres en IARD.

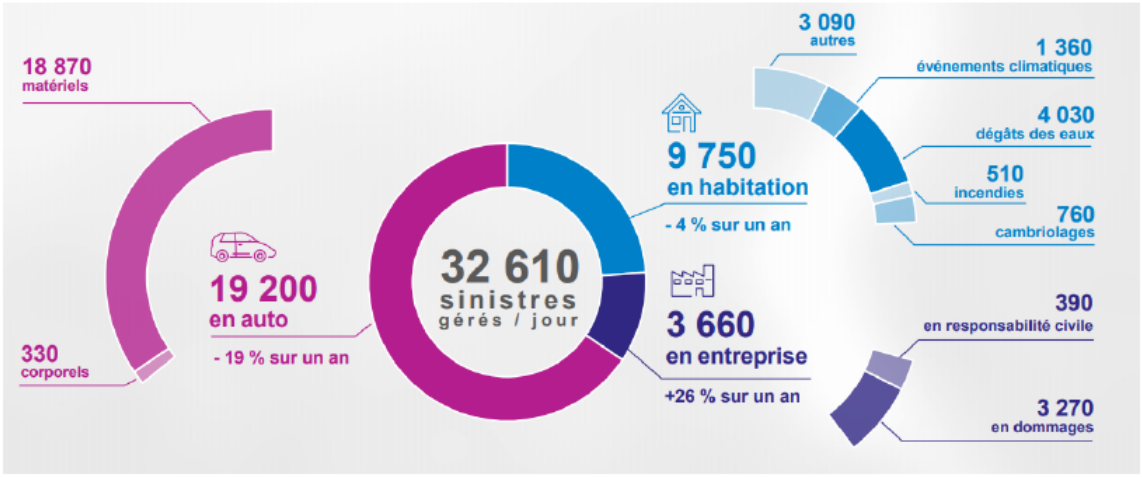


Figure 3 – Nombre de sinistres gérés par jour (source FFA 2021)

Ces quelques chiffres révèlent qu'environ 33000 sinistres étaient gérés par jour en 2021 sur la totalité du périmètre de l'assurance des biens et responsabilité. Et sur ce volume de sinistres, 9750 sont liés à des sinistres habitations correspondant à une proportion d'environ 30%. Enfin, en faisant un focus sur les sinistres dégâts des eaux, ils représentent 41,3% des sinistres habitations et 12,3% du nombre de sinistres gérés par jour.

Les garanties de base du contrat multirisque habitation :

L'offre multirisque habitation comporte plusieurs types de garanties permettant d'avoir une protection complète. Les garanties de base qui sont retrouvées le plus souvent dans les contrats d'assurance multirisque habitation sont : la responsabilité civile, le dégât des eaux, l'incendie, le vol/vandalisme, les évènements climatiques, les catastrophes naturelles et technologiques, le bris de glace ainsi que les attentats et actes de terrorisme. L'assuré, s'il le souhaite, peut décider de rajouter des garanties supplémentaires dites optionnelles adaptables en fonction du logement, du mode de vie et des besoins.
Dans le cadre de ce mémoire et tout au long des études, nous nous focaliserons que sur la garantie **dégât des eaux** du produit **multirisque habitation**.

2.3 Focus sur la garantie dégât des eaux chez Generali:

Cette garantie est presque toujours présente dans les contrats habitation. L'assuré ayant souscrit à cette garantie peut bénéficier d'une indemnisation en cas de dommages aux biens de façon accidentelle causés par l'action de l'eau. Une indemnisation qui dépendra du coût des dommages suite à la survenance du sinistre garanti, mais aussi des limites de garantie qui ont été définies par le contrat notamment de la franchise et du LCI (limite contractuelle d'indemnisation) convenue entre l'assureur et l'assuré). Ainsi, la garantie dégât des eaux permet de se protéger contre les dommages causés par :

- **Des infiltrations accidentelles** : l'infiltration par ou à travers la toiture, la façade, les balcons, les ciels vitrés, terrasses ... créant ainsi des dommages mobiliers, immobiliers ou des dommages au niveau de l'embellissement suite à une évaluation par l'assureur.
- **les écoulements d'eau accidentels** : il s'agit généralement des dommages résultants des écoulements d'appareils à effet d'eau comme une machine à laver, un aquarium ou bien de chauffage etc.
- **Débordement**

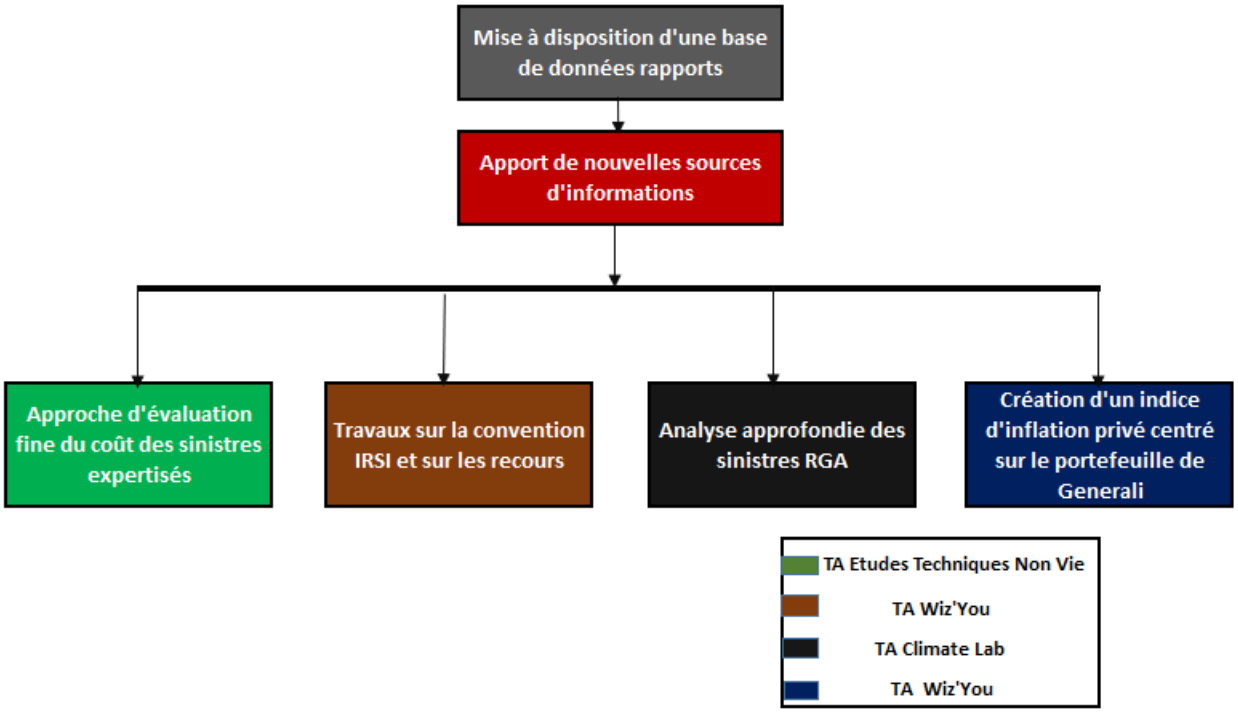
Par ailleurs, la garantie dégât des eaux présente aussi des exclusions prévues par le contrat d'assurance. En effet, l'exclusion de garantie est une clause prévue dans les contrats d'assurance et donne le droit à l'assureur de ne pas indemniser l'assuré en cas de dommages dûs à certains évènements. Ainsi, Generali, dans sa garantie dégats des eaux, exclu toute forme d'indemnisation en cas de dommages causés par :

- **La condensation**
- **L'humidité**
- **Un manque d'entretien**

3 Contexte et motivations de l'étude:

3.1 Le manque de données détaillant la sinistralité en DAB :

Dans le cadre du pilotage de son activité, un organisme assureur cherche à avoir le plus de données et d'informations concernant sa sinistralité pour maîtriser au mieux les risques auxquels il fait face au quotidien. A l'heure actuelle, Generali n'a pas une vision très détaillée concernant le découpage des sinistres notamment les prestations versées aux assurés suite à un sinistre. C'est pour pallier cette problématique de manque d'informations qu'entrent en jeu les rapports d'expertise par le biais de la richesse des données qu'ils renferment. En effet, savoir explorer et capter les informations d'expertise servira à alimenter les bases afin d'enrichir les données préexistantes et de disposer de nouvelles variables. Cela permettra ainsi de mieux comprendre les sinistres et d'améliorer sa maîtrise du risque. Un des objectifs de cette étude sera de mettre en place un outil qui automatisera l'extraction du contenu des rapports, le traitement et la mise en place d'une base regroupant les données d'expertise de nos sinistres. Cette dernière permettra aux équipes d'entreprendre différentes études en lien avec le pilotage de l'exercice dont les principaux d'entre eux se présentent de la manière suivante :



Dans ce mémoire, nous nous focaliserons sur l'analyse détaillée du coût des sinistres expertisés en dégât des eaux de la branche multirisque habitation. L'objectif sera d'introduire une approche de modélisation fine dans le cadre de l'évaluation du coût des indemnités versées aux assurés la suite du passage de l'expert. Par ailleurs, avec ces nouvelles données à disposition, différents travaux seront lancés parmi lesquels nous pouvons citer dans un premier temps l'étude concernant la convention **IRSI**.

Le but de cette étude sur l'IRSI est de réussir à identifier les sinistres rattachés à cette convention afin de calculer les taux de recours exercés/subis ainsi que notre taux de RC (responsabilité civile). Dans un second temps, il est possible de citer également l'étude orientée sur une analyse approfondie des sinistres de type **RGA** (retrait-gonflement des argiles) ou encore la mise en place de différents reportings de suivi de la tendance du coût moyen à une maille plus fine (la maille type de dommage indemnisé). Pour finir, des travaux seront réalisés sur l'inflation, l'enjeu de ce projet étant de mettre en place un indice d'inflation privé qui sera beaucoup plus centré sur les spécificités du portefeuille de Generali et de ses clients et qui sera sans doute challengé avec l'indice d'inflation public.

3.2 Les statistiques portefeuille chez Generali :

Dans cette partie, nous allons présenter le contexte portefeuille. Notamment par le biais de quelques statistiques en charge et en nombre concernant la sinistralité.

3.2.1 Répartition de la charge et des sinistres par garantie :

Il s'agit de la garantie qui présente la charge et la fréquence la plus élevée dans les contrats multirisque habitation chez Generali. Ces typologies de sinistres peuvent représenter un risque de taille et un enjeu financier considérable pour un organisme assureur. Dans le cadre de cette étude, nous excluons tout sinistre dégât des eaux engendré par des inondations généralement couvert par une autre garantie (la garantie évènements climatiques). Les graphiques ci-dessous permettent d'apprécier les statistiques concernant la sinistralité par garantie du portefeuille en multirisque habitation:

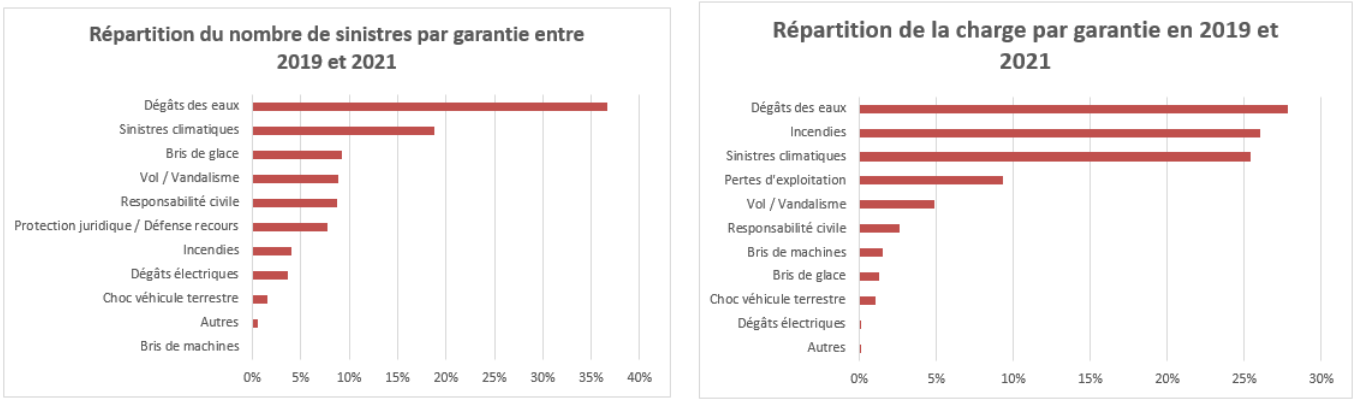


Figure 4 – Répartition des sinistres en nombre et en montant par garantie en MRH

L'analyse du graphique ci-dessus montre que sur la période d'étude, 36 % des sinistres en multirisque habitation sont liés à des phénomènes de dégâts des eaux. De même, nous pouvons également remarquer que 28% de la charge en multirisque habitation est portée par cette même garantie.

Ce qui permet de conclure que cette garantie est la plus sinistrée en terme de nombre et de charge parmis toutes les garanties présentes au niveau de la branche multirisque habitation.

3.2.2 Répartition de la sinistralité par type d'expertise :

Dans le cadre de la gestion des sinistres, nous constatons les sinistres expertisés et les sinistres non expertisés. Pour rappel, un sinistre est expertisé lorsque l'enjeu financier dépasse un certains montant/seuil. Dans un tel cas de figure, l'assureur mandate son expert pour réaliser un inventaire voir un descriptif technique des pertes financières. Nous pouvons observer ci-dessous quelques statistiques concernant la décomposition de la charge entre expertisés et non expertisés :

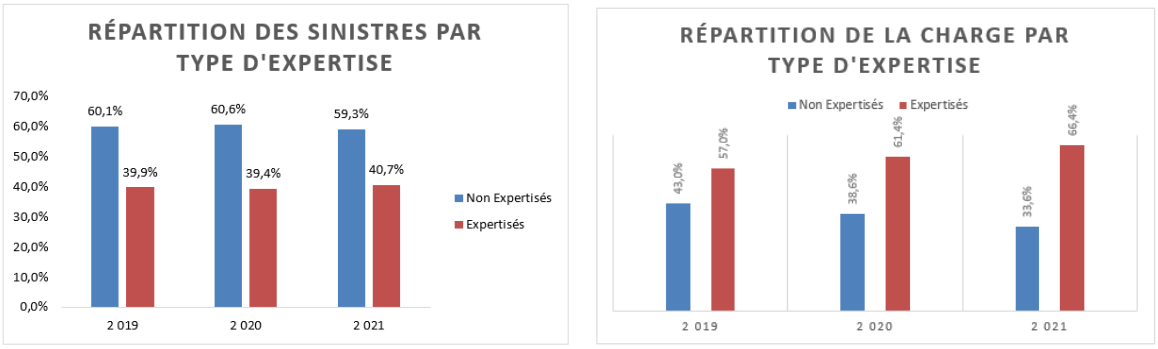


Figure 5 – Répartition de la charge des sinistres dégât des eaux par type d'expertise

A travers ces statistiques réalisés entre 2019 et 2021, nous pouvons voir clairement que la charge nette de ces sinistres expertisés suit une tendance à la hausse en passant de 57% en 2019 à 66.4% en 2021 malgré la stabilité observée en terme de nombre. Le poids de ces sinistres expertisés est de plus en plus prépondérant au fil des années alourdissant de plus en plus la charge pour cette garantie. C'est ainsi que dans le cadre de ce projet, notre étude consistera dans un premier temps à enrichir les données et de mettre en place une base permettant d'atteindre une segmentation plus fine du coût de nos sinistres. Par la suite, notre objectif sera d'introduire une approche de modélisation fine dans le cadre de l'estimation du coût moyen des dommages indemnisés relatifs aux sinistres ayant fait l'objet d'une expertise.

3.3 L'expertise en assurance :

L'expertise en assurance-dommage se définit comme étant l'opération par laquelle les compagnies d'assurance mandatent après la survenance d'un sinistre des experts dans le cadre de l'inventaire et de l'évaluation des dommages. L'objectif étant d'identifier l'origine du sinistre, d'évaluer les enjeux économiques du dossier, de prendre des mesures conservatoires afin que les frais ne soient pas trop conséquents mais aussi d'identifier les responsabilités afin que les recours puissent être récupérés auprès des éventuels responsables adverses. Cependant, il est important de souligner que tous les sinistres ne sont pas expertisés. Il n'est pas toujours nécessaire d'avoir recours à une expertise. En règle générale, un expert sera mandaté si l'évaluation des dommages dépasse un certain seuil ou si l'assureur à des doutes concernant les déclarations de l'assuré. Ce seuil est souvent évalué à 3000 euros, mais peut être revu à la baisse notamment en cas de sinistre du type de dégâts des eaux où un expert est généralement mandaté si les préjudices atteignent les 1600 euros. Ainsi, en deçà du seuil, si le préjudice financier n'est pas très important et que l'assureur arrive à se fier aux déclarations de son assuré, il peut décider de ne pas mandater d'expert. Différents types d'expertises se présentent:

— **L'expertise a l'amiable :** L'expertise amiable à pour objectif d'essayer de solutionner un litige lorsque le gré à gré n'a pas pu aboutir afin d'éviter d'entamer des procédure judiciaires. Ainsi, l'expertise amiable peut être **unilatérale**, c'est à dire l'expert est désigné par l'assureur uniquement. Elle est qualifiée de **contradictoire** en cas de présence d'un tiers responsable où chaque partie désigne son propre expert.
Dans tous les cas, il sera toujours possible de contester les conclusions.

— **La contre expertise :** Elle fait suite au premier passage d'un expert mandaté par l'assureur. En effet, dans le cas d'un désaccord avec les conclusions de l'expert, il est possible de rédiger une lettre de contestation qui sera suivi par la contre-expertise. Ainsi, dans ce cas là, l'expert sera chargé de réaliser la même opération d'expertise en accordant une attention particulière aux points de désaccord.

— **L'expertise judiciaire:** Elle fait suite à un désaccord persistant entre les parties prenantes menant ainsi à des procédures judiciaires, même suite à une expertise amiable.

Dans un premier temps nous allons présenter le déroulement d'un sinistre, afin de mieux comprendre à quel moment intervient l'expert, quel est son rôle. Mais aussi, sa façon de rédiger un rapport d'expertise à travers une présentation de la structure et du contenu de ce dernier.

3.4 Étapes et gestion d'un sinistre en multirisque habitation :

Le sinistre est défini comme la survenance d'un évènement prévu par les contrats d'assurance. Autrement dit, la réalisation du risque garantie par l'assureur. Ainsi, après la survenance d'un sinistre habitation, l'étape préliminaire sera de vérifier auprès de son assureur que l'on est bien assuré pour le sinistre pour lequel on veut être pris en charge et de le déclarer ensuite. Toutefois, il peut y avoir des spécificités en fonction des sinistres. Par exemple en cas de dégât des eaux, l'idéal sera de déterminer la source de la fuite si possible et réparer le problème à l'aide d'un professionnel afin d'éviter d'amplifier la situation. La déclaration par l'assuré permettra aux gestionnaires de l'organisme assureur de pouvoir procéder à l'enregistrement du sinistre et de mandater un expert si nécessaire. En effet, dans certains cas un passage d'expert peut être obligatoire alors que dans d'autres cas l'assureur a le choix de mandater ou pas. Un choix qui dépendra de l'importance des dommages. Dans le cas où un expert est mandaté, il sera chargé d'évaluer et de recueillir des éléments qui lui permettront de rédiger son rapport d'expertise. Ainsi, cela permettra de quantifier les coûts des dommages, de permettre à l'assureur de définir un montant d'indemnisation et de pouvoir procéder au règlement.

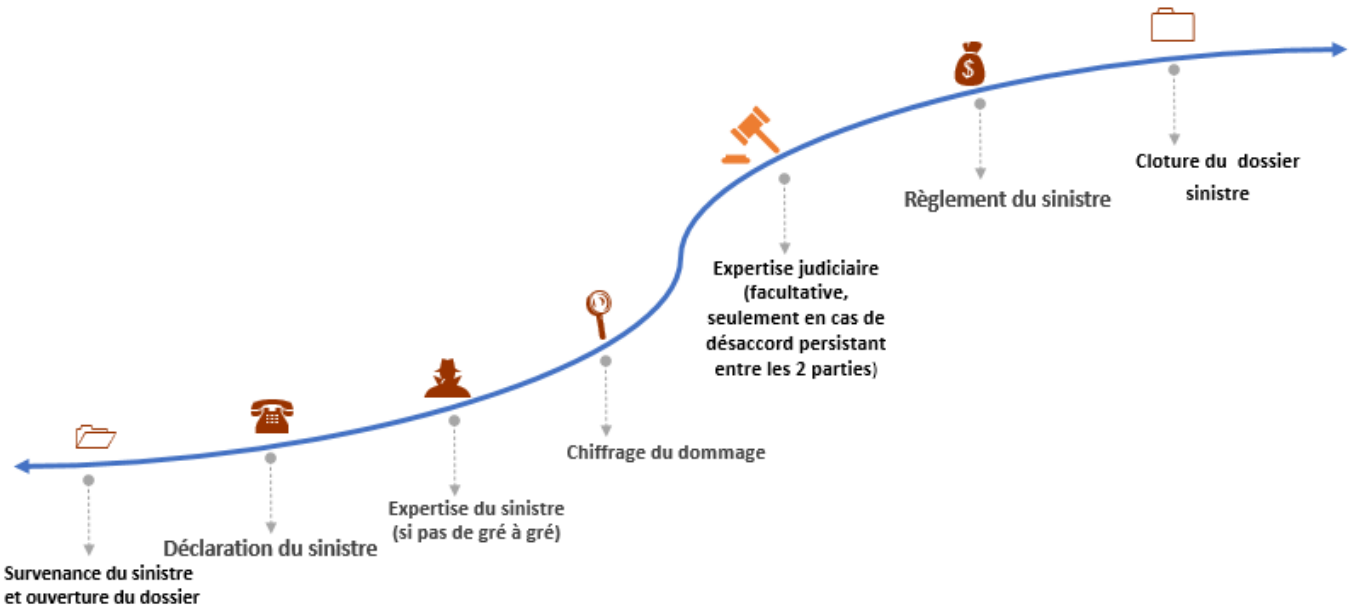


Figure 6 – Processus de gestion des sinistres expertisés MRH

Présentation du cabinet Saretec:

Saretec est un cabinet qui propose des solutions d'expertise aux organismes assureurs. Dans le cadre de ses expertises, la société fournit des prestations dans les branches telles que:

- La responsabilité civile
- La construction
- Le dommage aux biens
- La protection juridique

Les offres d'expertises proposées par Saretec concernent tous les risques prévus par les contrats d'assurances mis à part les risques liés à l'assurance automobile. L'opération d'expertise avec saretec peut être réalisée à distance (téléphonique/visio-expertise) ou en présentiel (sur site).

Ainsi dans le cadre des missions d'expertise de Generali, Saretec fait partie des cabinets mandatés par l'entreprise afin de réaliser des missions d'expertises. A noter que, parmi l'ensemble des sinistres survenus en DAB à Generali, à peu prés **40%** sont expertisés et Saretec expertise une plus grande partie de ces sinistres. En effet, Saretec expertise la plus grande part de cette proportion, suivi de son concurrent **Texa**.

A ce jour, la part de marché de Saretec sur l'ensemble des sinistres expertisés par Generali est évaluée aujourd'hui à environ **47%**.

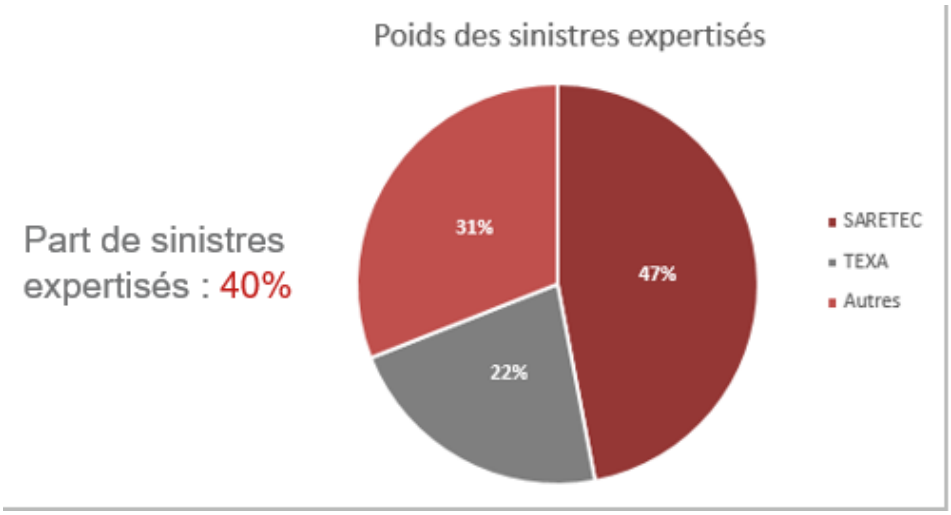


Figure 7 – La part de marché Saretec chez Generali

Présentation d'un rapport d'expertise multirisque habitation :

Les rapports sont des documents rédigés par des experts mandatés par l'assureur dans le but de réaliser un constat des dommages. Cela permettra par la suite à l'assureur de pouvoir vérifier si le sinistre est bien couvert par le contrat et de proposer le montant d'indemnisation relatif à des réparations ou à des remplacements de nature mobilière/immobilière dans les limites contractuelles. Le rapport d'expertise est destiné à faire un descriptif technique du sinistre, un constat des dégâts et des types de dommages, une analyse plus ou moins détaillée des causes du sinistre ainsi qu'une quantification du montant d'indemnisation à proposer à l'assuré. Ainsi, il se trouve que ces rapports d'expertise présentent beaucoup d'informations importantes. Leur extraction puis leur industrialisation permettra à Generali d'enrichir les données déjà existantes et de pouvoir aller dans une maille beaucoup plus fine dans la compréhension de la sinistralité afin d'en assurer un meilleur suivi. La consultation des rapports d'expertises destinés à l'étude montre la présence de plusieurs formats de rapports. En effet, les formats des rapports d'expertise de Saretec peuvent changer d'une année à une autre, mais aussi il peut y avoir différents types de rapports au sein d'une même année. Ci-dessous les grandes lignes du rapport le plus courant afin de se faire une idée sur les types d'informations qu'il peut contenir :

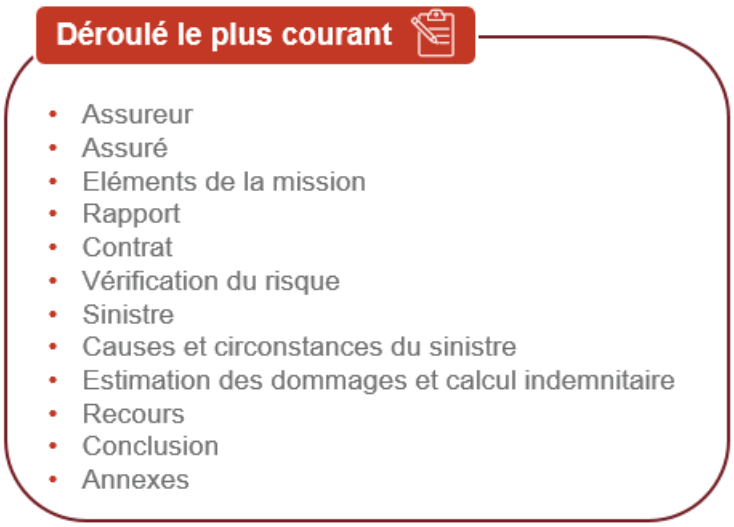


Figure 8 – Déroulé du rapport le plus courant de l'échantillon

Il est possible de visualiser un exemple de rapport d'expertise (**Cf** . [ici]).

Part II

Mise en place de l'étude

4 Présentation des données d'étude

Dans le cadre de cette étude, les données utilisées viennent de l'historique des données contrats et sinistres présentes dans les bases de données de Generali.
Mais, ces données sont complétées par les données extraites des rapports d'expertise afin de pouvoir aller dans une maille beaucoup plus fine de la compréhension de la sinistralité en DAB. Le périmètre d'étude concernera les sinistres qui ont été expertisés entre 2013 et 2021 par le cabinet Saretec.

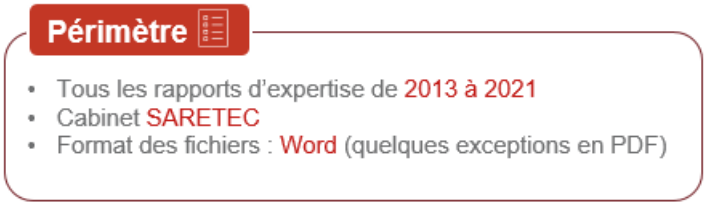


Figure 9 – Le périmètre d'étude

4.1 Présentation et création de la base des données d'expertise :

Dans le cadre de cette étude, un travail préliminaire a été effectué avant de mettre en place la base de données issue des rapports d'expertise. En effet, dans un premier temps, l'historique des rapports concernant les sinistres expertisés entre 2013 et 2021 a été récupéré et stocké sous une plate-forme (HDFS).

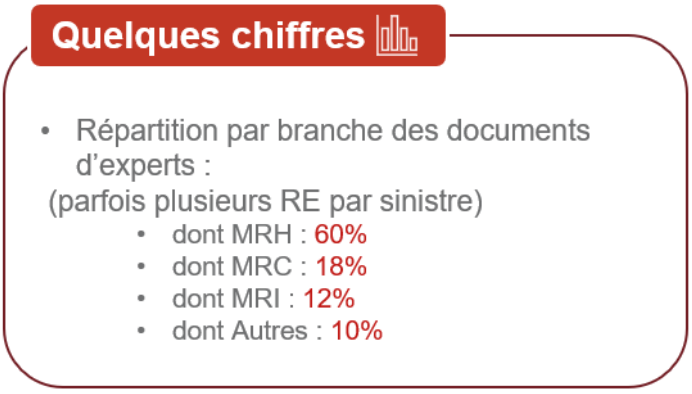


Figure 10 – Quelques chiffres clés

Environ 200.000 rapports ont été collectés majoritairement sur le produit MRH (Multirisque Habitation). Le but de ce projet sera de réussir à industrialiser la construction d'une base rapport qui servira à alimenter les base et à enrichir les données déjà existantes.

4.1.1 Travaux d'extraction et de création de la base des rapports :

Travaux préliminaires :

Dans le cadre de la construction de la base de données des rapports, il a fallu dans un premier temps procéder à la compréhension de ces documents. De ce fait, la lecture d'une centaine de rapports d'expertise était requise afin de :

— **Bonne connaissance des rapports :** Cette étape s'avère importante dans le cadre de la poursuite de l'étude. C'est dans ce sens qu'un atelier avec l'indemnisation a été organisé afin d'avoir une vision métier des rapports d'expertise afin de la confronter aux premières analyses. Mais aussi, cela a permis d'avoir de nouvelles idées de cas d'études et d'informations à extraire des rapports et de réfléchir sur la démarche à adopter dans le cadre de la généralisation d'un script qui pourra potentiellement s'adapter à l'ensemble des rapports disponibles.

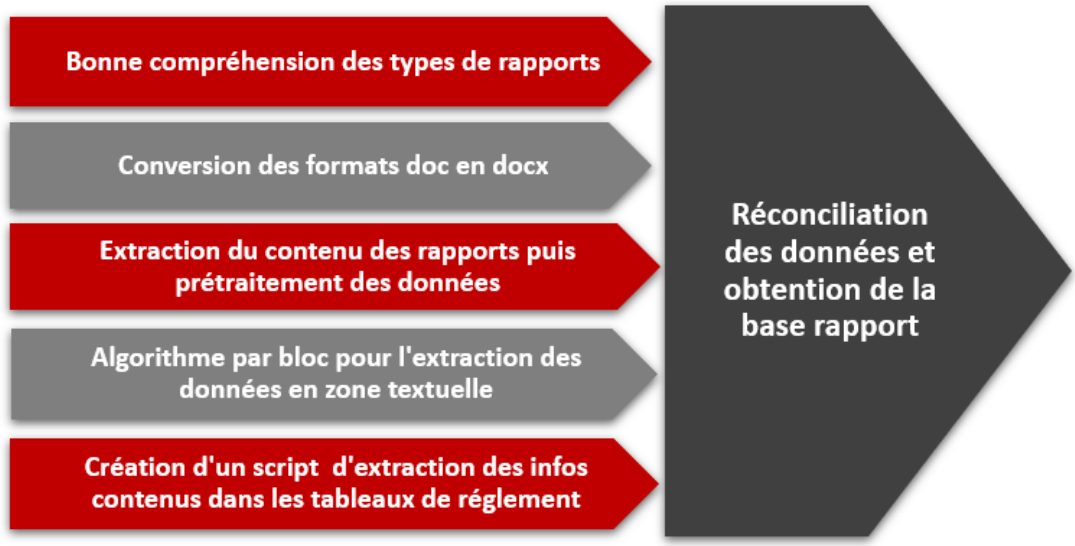
— **Identification des mots clés :** L'identification des mots clés est une étape importante dans cette procédure d'extraction. La complexité réside sur le fait qu' une variable peut parfois être rattachée à différents mots clés spécifiques à sa récupération et dépend du rapport d'expertise traité.

Remarque : Dans le cadre de l'optimisation du processus, quelques étapes préliminaires sont requises et permettent de fluidifier l'extraction et la création de la base de données (**Cf** . [ici]). En effet, il a fallu dans un premier temps transformer les majuscules en minuscules, identifier puis enlever tous les accents et caractères spéciaux. Par la suite, nous avons décidé de subdiviser les données en deux catégories : les données en zone de texte et les données identifiées au niveau des tableaux d'évaluation. La complexité du traitement diffère en fonction de la zone de localisation de l'élément à extraire.

Récupération des données (Approche générale) :

Compte tenu du volume important de rapports, ces documents sont stockés par année dans des répertoires disponibles sous la plate-forme hdfs de Generali qui est connectée à hadoop. La solution trouvée tout en évitant de saturer le mémoire était de diviser les rapports par lots de 1000 rapports puis de réaliser une jointure verticale par année de rapport.
Ensuite, pour obtenir le texte des rapports, il suffit de charger pour chaque fichier en format .docx sa représentation binaire dans une variable intermédiaire que l'on retransforme

en texte. Cela permettra de pouvoir l'utiliser lors de l'appel à des modules python qui ne travaillent qu'avec des fichiers locaux. La phase de pré-traitement du contenu des rapports d'expertise est suivie de la création d'un script permettant d'effectuer l'extraction des informations. Pour les données localisées en zone textuelle, le principe général était de jouer sur les mots clés identifiés en amont et qui sont relatifs à la variable cible cherchée. Cela a permis par la suite de créer une fonction spécifique à la récupération (algorithme par bloc avec un bloc par information à extraire). Concernant les données relatives aux règlements et situées au niveau des tableaux d'évaluation, le principe est de créer dans un premier temps un dictionnaire recensant l'ensemble des dommages identifiés dans les rapports d'expertise à la suite du passage de l'expert. L'algorithme se chargera de récupérer le dommage qui a été indemnisé qu'il rattachera à l'évaluation qui a été proposée par l'expert. Par ailleurs, l'usage du module **RE** de python permet de parcourir le texte du rapport afin d'identifier et d'extraire les éléments recherchés dans le cadre de l'étude.

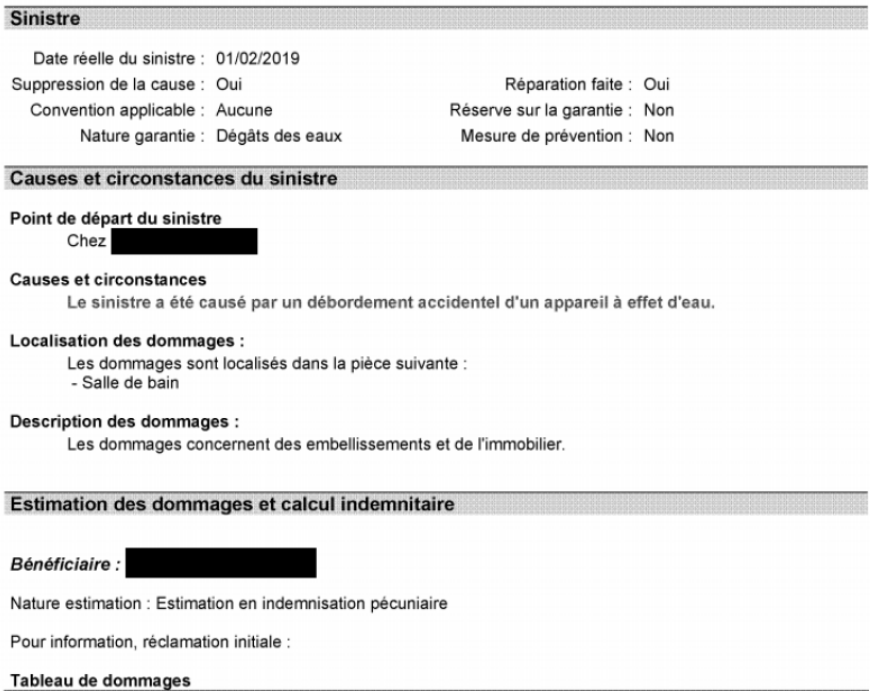


La base finale présentant l'ensemble des informations sera obtenue en créant une fonction finale consolidant l'ensemble des blocs de récupération prédéfinis.

Approche de localisation des éléments en zone de texte :

Dans le cadre de l'enrichissement des données, nous étions amenés à extraire différentes variables à partir des documents d'expertise . C'est ici que le travail sur la typographie des rapports prend tout son sens : en observant où se situe les éléments recherchés dans les rapports, on peut aller chercher l'information là où elle se situe. Si une information est dans une section précise, il suffit de localiser la section et d'extraire l'information de la section en question. Bien entendu, il faudra passer par une étape de nettoyage de l'information extraite pour seulement conserver la partie désirée et la formater au format souhaité.
Par ailleurs, certaines variables en zone de texte ont fait l'objet d'une démarche d'extraction spécifique. C'est le cas par exemple de la variable recensant la localisation des dommages.

Cette information est généralement localisée au niveau du descriptif des causes et circonstances du sinistres. Ci-dessous un exemple en guise d'illustration :



Sinistre

Date réelle du sinistre : 01/02/2019
Suppression de la cause : Oui — Réparation faite : Oui
Convention applicable : Aucune — Réserve sur la garantie : Non
Nature garantie : Dégâts des eaux — Mesure de prévention : Non

Causes et circonstances du sinistre

Point de départ du sinistre
Chez [redacted]

Causes et circonstances
Le sinistre a été causé par un débordement accidentel d'un appareil à effet d'eau.

Localisation des dommages :
Les dommages sont localisés dans la pièce suivante :
- Salle de bain

Description des dommages :
Les dommages concernent des embellissements et de l'immobilier.

Estimation des dommages et calcul indemnitaire

Bénéficiaire : [redacted]

Nature estimation : Estimation en indemnisation pécuniaire

Pour information, réclamation initiale :

Tableau de dommages

Figure 11 – Localisation des informations d'expertise

Ainsi, au delà du fait d'avoir réalisé quelques retraitements préliminaires, notamment la transformation des majuscules en minuscules ainsi que la suppression des accents et caractères spéciaux. L'idée ici était d'arriver à retrouver la localisation des dommages à la maille pièces. C'est dans ce sens que nous avions mis en place et alimenter un dictionnaire de référence avec une liste de pièces (chambre, cuisine, salle de bain, ...). Afin d'optimiser la procédure d'extraction, nous avons utilisé la forme **stemmer** des mots consistant à enlever la fin d'un mot dans le but de conserver qu'une forme très simple de ce dernier. Ci-dessous une illustration de la procédure adoptée :

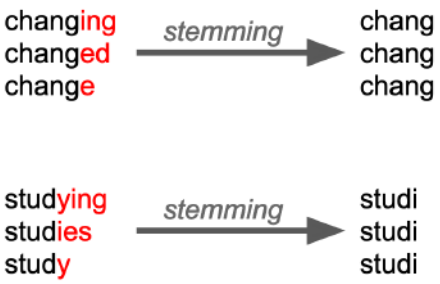
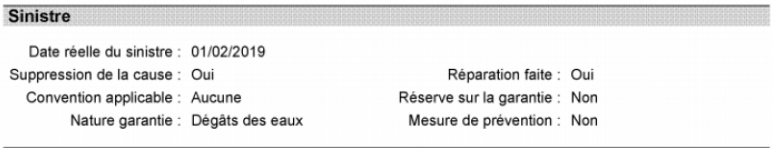


Figure 12 – Fonctionnement stemming

Cette approche permet de bien extraire la localisation des pièces et sinistres. Concernant la plupart des autres variables en corps de texte telles que la convention applicable par

exemple qui est très codifiée, les vérifications du risque, les mesures de prévention etc... il n'est pas nécessaire de recourir à l'option stemming mais plutôt d'extraire le mot ou groupe de mots situé juste après la cible recherchée.



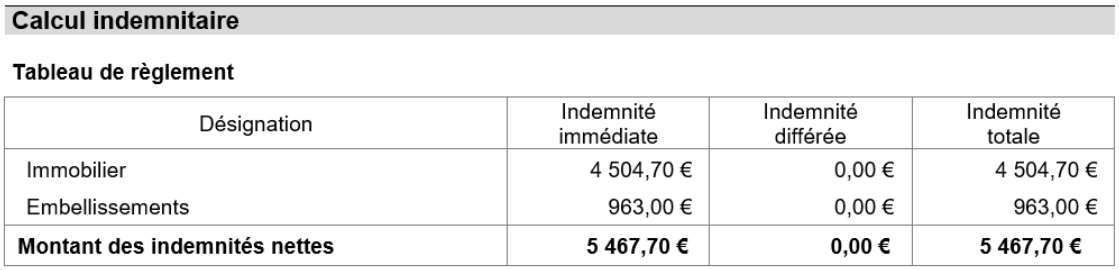
Sinistre

Date réelle du sinistre :	01/02/2019		
Suppression de la cause :	Oui	Réparation faite :	Oui
Convention applicable :	Aucune	Réserve sur la garantie :	Non
Nature garantie :	Dégâts des eaux	Mesure de prévention :	Non

Figure 13 – Convention applicable

Approche de localisation des éléments en zone tableau :

Suite à un sinistre, le rapport d'expertise peut présenter un ou plusieurs types de dommages indemnisés. C'est dans ce sens que, dans le cadre de l'extraction des données tableaux, il a fallu aussi réaliser les mêmes retraitements que précédemment puis mettre en place un dictionnaire de référence alimenté par les différents types de dommage indemnisés suite à un sinistre expertisé en dommage aux biens. Bien entendu il faudra dans un premier temps traiter le rapport de sorte à ne conserver que la partie désirée (dans notre cas, il s'agira de toujours localiser la zone correspondante à l'estimation des dommages ou calcul indemnitaire). Avec des formats pdf, il aurait été possible de passer par l'option tabula qui est une surcouche de pdfminer permettant d'analyser les pdf et de ressortir les dataframes. Toutefois, avec le format actuel, notre stratégie sera de recourir au module **re** de python pour parcourir le contenu des rapports afin d'identifier les mots clés prédéfinis et stocker en mémoire la zone cible délimitée dans laquelle sera réalisée les différents travaux. Par ailleurs, dans le cadre de l'extraction des montants en zone tableau, il a fallu réaliser un tri afin de procéder à une extraction par lot de rapports présentant un même nombre de dommages recensés suite à l'expertise. Le principe sera encore de jouer sur les mots clés afin d'affecter à chaque type de dommage recensé par l'expert le montant à indemniser qui lui est associé.



Calcul indemnitaire

Tableau de règlement

Désignation	Indemnité immédiate	Indemnité différée	Indemnité totale
Immobilier	4 504,70 €	0,00 €	4 504,70 €
Embellissements	963,00 €	0,00 €	963,00 €
Montant des indemnités nettes	**5 467,70 €**	**0,00 €**	**5 467,70 €**

Figure 14 – Calcul indemnitaire

4.1.2 Les données règlements :

Lors de la rédaction du rapport par l'expert, les montants d' évaluation des différents dommages à indemniser se présentent souvent sous forme de tableau de règlement avec 4 colonnes. Le tableau le plus courant dans les rapports se décline comme suit :

Calcul indemnitaire

Tableau de règlement

Désignation	Indemnité immédiate	Indemnité différée	Indemnité totale
Immobilier	4 504,70 €	0,00 €	4 504,70 €
Embellissements	963,00 €	0,00 €	963,00 €
Montant des indemnités nettes	**5 467,70 €**	**0,00 €**	**5 467,70 €**

Figure 15 – Exemple de tableau de règlement

Ainsi, comme nous pouvons voir dans le tableau de règlement ci-dessus, différents types de dommage peuvent se présenter suite à un sinistre dégât des eaux. Ci-dessous, nous allons présenter les plus fréquents :

- **Les dommages à l'immobilier :** Un bien immobilier est un bien par définition immobile. On parlera de bien immobilier **par nature** quand il s'agira par exemple de terrains ou de bâtiments fixés au sol et de bien immobiliers **par destination** quand il s'agit de meubles scellés ou fixés au logement et dont le détachement est susceptible de causer des dommages au support.
- **Les dommages à l'embellissement :** Les embellissements ont une définition issue de la convention CIDRE résultant d'un accord commun entre les compagnies d'assurance. Cette convention définit les embellissements comme : « Les peintures et vernis, miroirs fixés aux murs, revêtements de boiseries, faux plafonds, éléments fixés de cuisine ou de salles de bains aménagées, ainsi que tous revêtements collés de sol, de murs et de plafonds, à l'exclusion des carrelages et parquets »
- **Les dommages au mobilier** : Les biens mobiliers comprennent l'ensemble des meubles et objets (y compris animaux domestiques). Les contrats d'assurance habitation couvrent le mobilier personnel de l'assuré, des membres de sa famille. Mais garantissent également ceux appartenant aux employés, ouvriers ou toute autre personne résidant ou se trouvant momentanément dans les locaux assurés.

Remarque :
Dans un rapport d'expertise, il est possible d'avoir un seul type de dommage dans le tableau d'évaluation tout comme il est possible d'avoir une liste de plusieurs types de dommage recensés par l'expert. Voici ci-dessous la liste des types de dommages retenus suite à une expertise en dégât des eaux.

Types de dommages	Détails	Type
Immobilier	Dommages relatifs à l'immobilier	Cible
Embellissement	Dommages relatifs à l'embellissement	Cible
Contenu ou Mobilier	Dommages relatifs au mobilier	Cible
Autres	Frais annexes, pertes immatérielles	Cible

Pour rappel, certains dommages portant des libellés très spécifiques ont été regroupés dans le poste **"Autres"**. Nous présentons dans le tableau suivant la liste non exhaustive de ces postes :

Autres postes
Pertes immatérielles
Frais afférents
Frais annexes
Recherche de fuite
Nettoyage
Assechement

Figure 16 – Liste des indemnisations regroupées dans Autres

Cette liste reflète ainsi les indemnisations versées par Generali au titre de ce dernier poste et qui sont parfois chiffrées au niveau du tableau récapitulatif des évaluations de l'expert. Ces postes sont très peu sinistrés dans notre base reflétant une part négligeable des coûts indemnisés.

La récupération des variables **Type de dommage** et **Nature de la garantie enjeu** disponibles dans la base de données des rapports d'expertise a permis de réaliser les premières statistiques portant notamment sur les différents postes de dommage recensés généralement après un sinistre dégât des eaux. Cette étude a permis d'avoir une visibilité sur la sinistralité par type de dommage suite à un dégât des eaux. Il apparaît que les deux postes les plus sinistrés en terme de nombre concernent l'embellissement et l'immobilier.

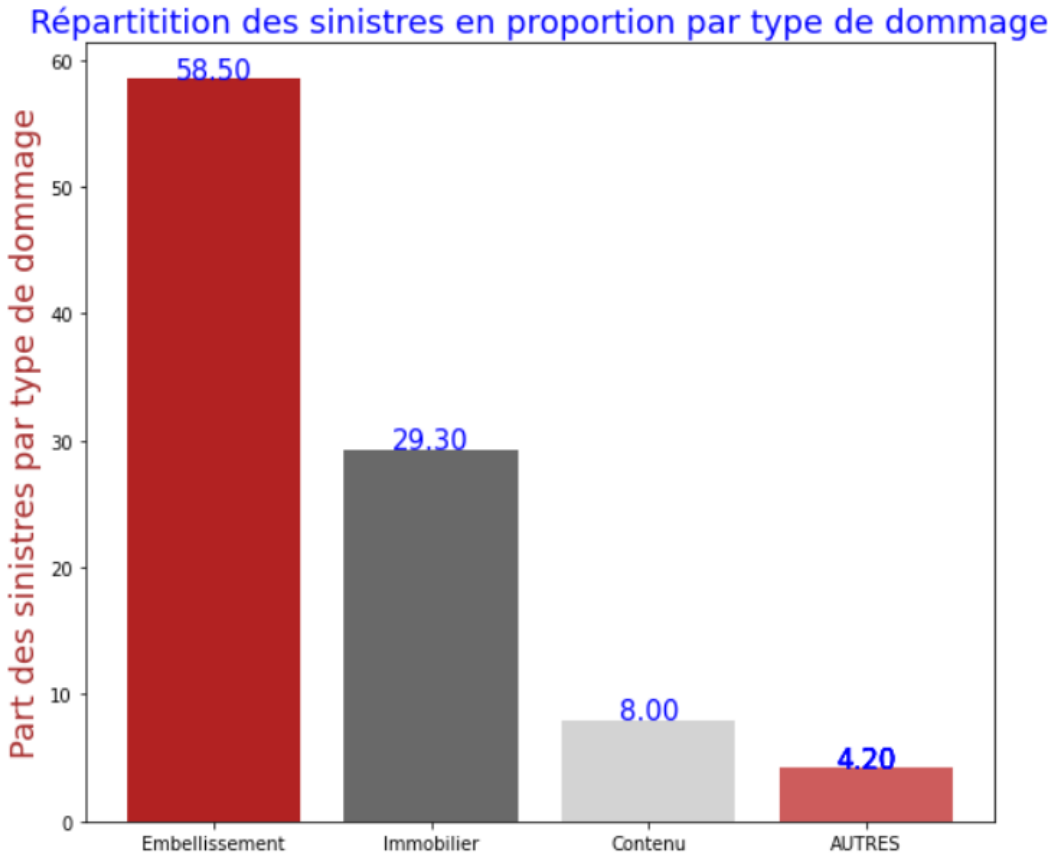


Figure 17 – Statistiques sur les types de dommage en dégât des eaux

En effet, il apparaît qu'environ 58% des sinistres survenus et expertisés en dégât des eaux habitation sont liés à des dommages à l'embellissement, 29% à de l'immobilier, 8% à du mobilier et 4% sont liés à des frais annexes. Nous pouvons remarquer que plus de la moitié des sinistres expertisés concerne les dommages à l'embellissement. Le mobilier ou contenu ainsi que le dernier poste regroupant les autres libellés spécifiques sont des postes très peu sinistrés dans le portefeuille.

4.1.3 Statistiques sur la récupération des données rapports d'expertise :

Quelques statistiques ont été réalisées sur la base des rapports notamment pour avoir une idée sur la part de l'échantillon potentiellement exploitable par la suite. En effet, l'extraction a été généralisée sur un certains volume de rapports. Cependant, il arrive que le rapport ne présente pas de tableau de règlement. Il s'agit la plupart du temps de compte rendu de rapport, des notes d'experts, photos diverses, notes d'honoraires etc... D'ailleurs, ces types de rapport n'ont pas pour vocation d'apporter une visibilité concernant le détails des règlements et ne nous intéressent dans le cadre de ce projet. Toutefois, certains de ces documents peuvent servir de complément au rapport d'expertise initial afin d'apporter plus de détails par rapport au sinistre. Il a donc fallu extraire et utiliser des variables permettant d'identifier les rapports d'expertise (titre du rapport) ou chercher la présence de mots spécifiques aux rapports d'experts. Par ailleurs, certains sinistres peuvent également faire l'objet d'une expertise sans faire suite à un versement de prestation que nous allons mieux détailler dans la suite. Nous avons réalisé quelques statistiques à partir de notre échantillon concernant les rapports relatifs à nos sinistres et qui seront exploités dans le cadre de nos études.

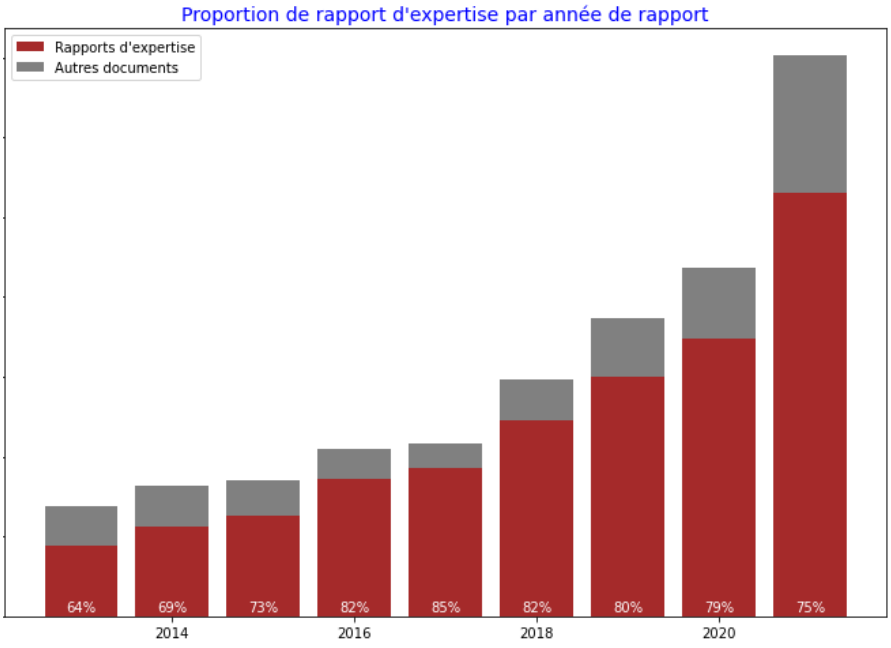


Figure 18 – Proportion de rapports d'expertise par année de rapport en dégât des eaux

Ces statistiques révèlent qu'en moyenne 20% des rapports ne serviront pas dans le cadre de notre étude. Autrement dit, environ 20% ne sont pas réellement des rapports d'expertises et donc ne seront pas intégrer dans la suite de l'étude.

4.2 Les données sinistres:

La base sinistre est le résultat d'un croisement de plusieurs tables alimentées au fil des années afin de retracer l'historique de la sinistralité de la compagnie.
Plusieurs variables dans cette base ne seront pas utiles à la modélisation mais elles serviront dans le cadre de certaines analyses.

Champs	**Détails**
Numéro de sinistre	Référence du dossier
Numéro de contrat	Référence du contrat
Etat du dossier	En-cours, clos, sans-suite
Garantie	
Sous garantie	
Le Bilan	
Les charges nettes et brutes	charge sinistre
Le montant des honoraires expert	coût d'expertise sinistre
Top expert	Liste des cabinets
Top expert honoraire	Top sur les sinistres expertisés
Date de survenance	
Nature du sinistre	

Cette base de données contient au départ tous les sinistres dommages aux biens de l'historique de Generali. Dans le cadre de notre étude, des filtres ont été réalisés notamment en faisant un focus sur la branche **multirisque habitation** et sur la garantie **dégâts des eaux**. Ne seront considérés dans cette étude que les sinistres clos pour des raisons qui seront présentées dans la suite de ce mémoire.

4.3 Les données risques:

Cette base de données contient l'ensemble des variables caractéristiques de l'habitation et qui ont été collectées au moment de la souscription de l'assuré.
Ces variables permettent d'avoir une vision sur le risque à assurer. Parmi les différentes variables présentes dans cette base il y a :

Champs	Détails	Type
Numéro de contrat	Permettant de référencer les contrats	Jointure
La surface	en m^2	Explicative
Le statut de l'assuré	Propriétaire/ Locataire	Explicative
Le type d'habitation	Maison/Appartement	Explicative
Le nombre de pièces	Le nombres de pièces disponibles dans l'habitation	Explicative
Le nombre d'enfants	Cette variable se décline en 4 modalités : 1,2,3, >3 enfants	Explicative
Le réseau distribution	Agents/Courtiers	Explicative
Surface dépendance	En tranche	Explicative
Le zonier		Explicative
La franchise		Explicative
Le type de résidence	Principal/Secondaire	Explicative

La variable zonier coût selon l'adresse de l'assuré est la variable utilisée dans le modèle de tarification en MRH. Cette variable est construite selon la méthode de l'entreprise à segmenter les zones en des groupements homogènes de risques.

Par ailleurs, la variable franchise étant indexée sur le FFB a été obtenue en appliquant le produit :

$$\boxed{Franchise = COEF * IndiceFFB}$$

Le coefficient ou niveau de franchise est toujours précisé et est situé entre [0,1]. Cependant l'indice ne l'est pas forcément. De ce fait, afin de bien le définir, l'idée était de se baser sur la date de survenance pour la rattacher à l'indice FFB correspondant. Dans le cadre du premier modèle (le GLM), cette variable fera l'objet d'une catégorisation en groupements homogènes via la méthode des quantiles. Cela aura pour objectif de simplifier le modèle et de renforcer son efficacité durant la phase d'apprentissage.

4.4 La table des dépenses de Generali (DWBDPR) :

Cette table permet d'avoir des informations concernant l'indemnisation des sinistres déclarés. Elle est aussi appelée table des **Dépenses/Recettes** et permet de se faire une idée sur les différents mouvements des dépenses/recettes (règlement principal, complément principal, annulation de règlement, acompte etc), mais également d'avoir une visibilité sur les dates auxquelles ces différents mouvements ont été effectués.

4.5 Synthèse des variables extraites des rapports:

Cette base de données est relative à une partie des variables obtenues à l'issu des travaux d'extraction des données contenues dans les rapports d'expertise. Ces variables viendront compléter nos données initiales dans le cadre de l'étude.

Champs	**Détails**
Nom du rapport	Clé de jointure
Type de dommage	Embellissement, Immobilier, Mobilier, Autres
Dommage embellissement	Évaluation du dommage associé à l'embellissement
Dommage immobilier	Évaluation du dommage associé à l'immobilier
Dommage mobilier	Évaluation du dommage associé au mobilier
Indemnisation	Oui/Non
Numéro de sinistre	Référence du sinistre
Numéro du contrat	Référence du contrat
Dommage total	Total des dommages du sinistre
Date du rapport	Date de rédaction du rapport
Franchise	Montant
Qualité assuré	
Nombre de pièces	Renseigne sur le nombre de pièces présents dans l'habitation
Titre du rapport	Intitulé du rapport (en début de page)

La variable **Nom du rapport** est présentée ici comme une clé de jointure car elle est unique. En effet, il est important de noter qu'un même numéro de sinistre peut être lié à plusieurs rapports d'expertise. Toutefois, chacun de ces rapports possède un nom qui lui est propre, d'où l'unicité.

La construction de cette base a nécessité un travail important. Sa mise en place a été suivie d'un processus de retraitement se déclinant comme suit :

— Correction des fautes d'orthographe
— Retraitement des montants par type de dommage
— Contrôle de cohérence
— Identification de la dernière vision des rapports (un sinistre pouvant être rattaché à plusieurs rapports d'expertise).
— Fiabilisation des montants extraits des rapports

L'étape d'identification de la dernière vision des rapports a été une étape cruciale. L'extraction et l'usage des variables telles que la **date du rapport** et le **nom du rapport** a permis d'identifier la dernière mise à jour des montants évalués suite à l'expertise. La difficulté principale de ce traitement est liée au fait que les formats des dates peuvent différer d'un rapport à l'autre. Il a fallu pour pallier ce problème homogénéiser les dates afin de disposer d'un format unique. Aussi, certains rapports d'expertise ne présentaient pas de date de rapport. Une alternative était de réaliser une extraction de la variable relative à la **date de visite** de l'expert qui sera utilisé dans le cadre de la récupération de la dernière vision pour ces types de rapports. Ce traitement a ainsi permis dans la suite d'entamer le processus de fiabilisation des montants. L'objectif de cette dernière étape consistait notamment à vérifier si les montants présents au niveau de la base rapport reflètent fidèlement la réalité des indemnisations de Generali. Toutefois, une étape intermédiaire consistait à réaliser en amont des jointures successives entre nos différentes bases que nous détaillerons dans la suite de ce mémoire. La base principale qui sera utilisée pour vérifier l'alignement des montants entre les deux visions est la base des **Dépenses/Recettes** retraçant l'historique des prestations allouées par Generali.

5 Analyse préliminaire sur les données:

Nous allons nous intéresser ici à la réalisation de quelques statistiques descriptives qui permettront dans un premier temps d'appréhender la décomposition de la charge sinistre. Mais également de mieux orienter l'étude et de pouvoir définir le périmètre de nos travaux.

5.1 Statistiques sur la décomposition de la charge globale des sinistres dégât des eaux habitation :

La base sinistre a permis d'obtenir quelques statistiques concernant la répartition de la charge. Le périmètre d'étude concerne l'ensemble des sinistres dégât des eaux du produit MRH survenus entre 2013 et 2021.

<table>
<tr><th colspan="4">Répartition de la charge en dégât des eaux MRH</th></tr>
<tr><th colspan="4">Répartition par type d'expertise</th></tr>
<tr><th>Non expertisés</th><th colspan="3">Expertisés</th></tr>
<tr><td>38%</td><td colspan="3">62%</td></tr>
<tr><td rowspan="4"></td><th colspan="3">Répartition par gravité</th></tr>
<tr><th>Grave</th><th colspan="2">Non Grave</th></tr>
<tr><td>17%</td><td colspan="2">83%</td></tr>
<tr><td rowspan="2"></td><th colspan="2">Répartition par cabinet</th></tr>
<tr><th>Saretec</th><th>Autres cabinets</th></tr>
<tr><td></td><td>34%</td><td>66%</td></tr>
</table>

Figure 19 – Répartition de la charge des sinistres clos

Les résultats montrent que **38**% de la charge est portée par les sinistres non expertisés et **62**% par les sinistres expertisés. Nos travaux se concentrent sur les sinistres dégât des eaux qui ont été expertisés par le cabinet saretec car Generali récupère pour l'instant que les rapports de ce cabinet. Aussi, nous nous sommes focalisés spécifiquement sur l'étude des sinistres attritionnels. Hormis quelques valeurs considérées comme extrêmes le coût de ces sinistres ne dépassent pas dans la plupart du temps un certains seuil. Le seuil de grave retenu sera celui de Generali pour la garantie dégât des eaux du produit multirisque habitation et qui est fixé à 24.000€. Par ailleurs, étant donné que notre périmètre d'étude se limite aux sinistres expertisés, il est important de noter que **83**% de la charge des expertisés est liée à des sinistres non graves et que **34**% de ces sinistres sont expertisés par Saretec.

5.2 Tendance du coût moyen des sinistres expertisés par Saretec par rapport aux autres cabinets :

Quelques analyses préliminaires ont été réalisées avec la première version de la base sinistre . Pour rappel, cette base contient l'ensemble des sinistres survenus entre 2013 et 2021 en MRH avec comme seule garantie présente le **dégât des eaux**. Le but de ces premières analyses est d'étudier l'évolution du coût moyen des sinistres expertisés par Saretec.

En effet, étant donné que seuls les sinistres expertisés par un cabinet d'expert sont pris en compte dans cette étude, nous avons jugé intéressant de voir la tendance du coût moyen de Saretec par rapport à l'évolution du coût moyen de l'ensemble des sinistres expertisés. Le coût moyen étant défini à partir de la charge nette.

Ainsi, ces premières analyses préliminaires sont essentielles pour pouvoir comprendre :

- Si le comportement du coût moyen des sinistres expertisés par Saretec
- Si la tendance du coût moyen des sinistres expertisés par Saretec reflète la tendance du coût moyen de l'ensemble des sinistres expertisés par les autres cabinets

A la suite de cette étude comparative, les résultats des analyses ont montré que le coût moyen des sinistres Saretec suit la même tendance que le coût moyen de l'ensemble des sinistres expertisés. Mais aussi, le coût moyen de Saretec suit la même tendance que le coût moyen de l'ensemble des sinistres expertisés par les autres cabinets (saretec exclu).

6 Mise en place de la base de modélisation :

Cette section est consacrée à la construction de la base de données finale. En effet, différentes jointures ont été effectuées en partant des données sinistres comme base principale. Nous détaillerons ci-dessous le cheminement menant à la mise en place de la base de données qui sera utilisée dans le cadre des modélisations.

6.1 Retraitement de la base de données des sinistres :

Avant de procéder au premier croisement de la base sinistre avec la base de données des rapports, quelques retraitements de variables ont été effectués. Il s'agit notamment du traitement des variables telles que le numéro de sinistre, ou encore la police. Le but de ce traitement étant de faire en sorte que le format de ces données de jointure soient similaire au données des bases risques et sinistres. Par ailleurs, quelques filtres ont été dans un premiers temps appliqués au niveau de la base sinistre.

Filtre sur les expertisés en dégât des eaux:

En effet, étant donné que l'étude ne sera menée que sur le périmètre des sinistres expertisés, un filtre a été réalisé que sur les sinistres pour lesquels Generali a mandaté un expert pour l'évaluation des dommages.

Filtre sur les expertisés du cabinet saretec :

Aussi, étant donné qu'aujourd'hui, Generali ne récupère que les rapports du cabinet Saretec, l'étude sera donc restreinte sur ce périmètre. Ainsi grâce à la variable **Top expert honoraire** un top a été réalisé sur les sinistres expertisés par Saretec. De ce fait, nous obtenons par la suite **37 528** sinistres. Cet échantillon est représentatif de l'ensemble de nos sinistres expertisés en dégât des eaux habitation par ce cabinet.

6.2 Présentation des différentes jointures et retraitements :

Le rapprochement des différentes bases était une partie délicate qui a été réalisé en différentes étapes que nous allons présenter. Les principales difficultés rencontrées étaient les suivantes :

- Une écriture différente des clés de jointure au niveau des différents tables
- L'absence de clé similaire entre deux bases
- Les retraitements effectués au niveau des variables de jointure
- la gestion des doublons

Les principales clés utilisées seront le numéro de sinistre, la police et le nom du rapport d'expertise. Les erreurs relatives à l'écriture de ces clés ont été identifiées et corrigées.

6.2.1 Jointure avec la base des rapports :

Afin de rajouter les données des rapports dans notre base de données sinistre une jointure à gauche a été réalisée avec comme clé de jointure le **numéro de sinistre**.
Après jointure, nous arrivons à retrouver **43 331** rapports (à relativiser avec le fait qu'un sinistre peut être lié à plusieurs rapports) d'expertise à savoir **36 715** sinistres uniques. Ce qui fait qu'il s'avère nécessaire dans la suite de nos travaux d'identifier la dernière vision des rapports. L'objectif sera donc par la suite de conserver pour chaque sinistre le rapport d'expertise le plus récent présentant la dernière mise à jour des montants indemnisés.

- **Quelques numéro de sinistres mal renseignés sur la base des rapports :** La variable **numéro de sinistre** est dans certains cas non renseignée ou bien mal renseignée. Ces erreurs engendrent une perte de volumétrie d'environ 0.5% de la base initiale.

- **La non prise en compte des sinistres du réseau salarié :** Les sinistres du réseau salarié ont été écartés de l'étude du fait que plusieurs contrats étaient en cours d'avenant. Il n'était donc pas possible d'avoir accès à ces tables dans le cadre de nos travaux. Ce qui a généré une perte d'environ **3**% de la base.

A la fin de cette étape, la base obtenue est constituée de **35 606** sinistres.

6.2.2 Suppression des rapports ne présentant pas de tableau d'évaluation :

Dans un premier temps, l'étude a montré que tous les rapports d'expertise ne présentaient pas des tableaux d'évaluation des dommages. De ce fait, il a fallu identifier et écarter l'ensemble des rapports textuels, c'est-à-dire les rapports d'expertise non exploitables car ne présentant pas de montants susceptibles d'être extraits. Il s'agit souvent de compte rendu de rapport d'expertise faisant suite à l'évaluation de l'expert. Cette suppression génère une légère perte au niveau de nos sinistres (cela engendre une suppression quelques doublons). En effet, dans la plupart des cas, ces types de rapport sont des sortes de compte

rendu de rapport en complément du rapport d'expertise initial portant ainsi le même numéro de sinistre. Cette phase a permis ainsi de conserver **35 512** sinistres uniques.

6.2.3 Identification et prise en compte que de la dernière vision des rapports d'expertise

Il s'avère qu'un sinistre peut avoir plusieurs rapports (rapports d'expertise complémentaires, rectificatifs etc...) en lien avec le fait qu'un rapport peut évoluer notamment en cas de correction, d'aggravation ou de désaccord par exemple entre les deux parties. Cela mène de ce fait à la rédaction d'un nouveau rapport. De ce fait, le but de ce traitement sera de réussir à conserver la vision la plus récente de l'évaluation du sinistre pour pouvoir récupérer la dernière mise à jour des montants indemnisés. Par ailleurs, il est parfois possible que le rapport complémentaire soit intitulé rapport d'expertise rectificatif et que le second rapport d'expertise complémentaire soit intitulé rapport d'expertise supplémentaire.

Figure 20 – Évolution possible d'un rapport

6.2.4 Jointure avec la base risque :

Cette jointure s'est réalisée sur la base de la **police** et a permis de ramener l'ensemble des données risques qui seront utilisées dans le cadre de l'étude. Un retraitement a été fait dans le but de rattacher pour chaque sinistre la situation du risque au moment de la survenance. Notamment en appliquant la règle suivante :

6.3 Contrôles qualité des données et traitements préliminaires :

Dans cette partie, un travail important a été effectué sur la qualité des données. En effet, une fois avoir finalisé les différentes jointures, il était nécessaire de procéder à quelques traitements et vérifications afin de réussir à fiabiliser la base de données. Ci-dessous les différentes étapes du retraitement :

6.3.1 Contrôles de cohérence et traitements des données :

Nous avons effectué des contrôles de cohérence dans le but de fiabiliser les données.
Ainsi, dans un premier temps, nous avons cherché à respecter la règle suivante sur les dates :

$$date\ de\ survenance \preceq date\ du\ rapport \preceq date\ de\ règlement \preceq date\ clôture$$

Les quelques dates de rapport qui étaient antérieures à la survenance du sinistre ou qui se situaient après le règlement ou la clôture du sinistre ont été supprimées de la base.
Aussi, l'extraction de la variable **Indemnisation** au niveau des rapports a permis d'écarter les rapports d'expertise pour lesquels un versement de prestation suite au sinistre n'a pas eu lieu. Il s'agissait notamment des sinistres dont l'évaluation du dossier est nulle au niveau du système de gestion de Generali mais qui présentaient tout de même des montants au niveau des rapports. Ci- dessous un exemple de cas en guise d'illustration :

Estimation des dommages

Bénéficiaire : Pas d'indemnité

Taux de TVA : Voir en annexe le tableau Evaluation des dommages

Pour information, réclamation initiale : 2 464,62 €

Tableau estimatif

Désignation	Indemnité Immédiate	Indemnité Différée	Indemnité Totale
Embellissements (pour info, pas de règlement à l'assuré)	2 182,16 €	242,46 €	2 424,62 €

Observations :
Pas d'indemnité à régler, les dommages sont à prendre en charge par un autre assureur (dommages chiffrés).

Figure 21 – Exemple de rapport n'ayant pas été indemnisé par Generali

Par ailleurs, il existe des sinistres qui suite à une expertise ne font pas l'objet d'une indemnisation. Dans ces cas de figure, nous constatons une absence de montants au niveau du tableau d'évaluation des règlements. Ci-dessous un exemple des ces typologies de rapport en guise d'illustration :

Estimation des dommages et calcul indemnitaire

Indemnisation : Pas d'indemnité

Nature estimation :

Pour information, réclamation initiale : 1,00 €

Tableau de règlement

Désignation	Indemnité Immédiate	Indemnité Différée	Indemnité Totale
Total avant application de la franchise			

Observations : Absence de dommages constatés lors de l'expertise

Figure 22 – Exemple 2 de rapport n'ayant pas l'objet d'une indemnisation

Ces traitements ont permis ainsi de supprimer les sinistres sans réelle indemnisation allouée . A la suite de cette étape, nous nous retrouvons avec **29 203** sinistres. Ci-dessous, un exemple de la fonction de récupération de cette variable en illustration :

Variable indemnité

```
def indemnisation(txt):
    try:
        texte_verif_indemnite = re.split("beneficiaire")[1]
        if ("pas d indemnite" in texte_verif_indemnite) | ("pas de reglement"  in texte_verif_indemnite) :
            indemnite = "Non"
        else :
            indemnite = "Oui"
    except :
        indemnite = "Non renseigne"

    return(indemnite)
```

Figure 23 – Extraction de la variable relative à l'indemnisation de l'assuré

6.3.2 Périmètre des clos :

Le problème avec les sinistres en cours est qu'ils sont entachés d'une certaine incertitude. Toutefois, il est très souvent possible de les prendre en compte dans le cadre d'une étude et de déterminer la charge ultime de ces sinistres via une méthode de Chain-Ladder par exemple. Dans le cadre de cette étude, nous allons prendre en compte que les sinistres clos. Cela nous conduit à travailler sur un portefeuille de sinistres déjà réglés . C'est ainsi que la base obtenue à la suite de la prise en compte des sinistres ayant atteint leur charge ultime est constituée de **26 782** sinistres uniques. La considération que des clos génère une perte d'environ **7.5**% de nos sinistres. Ce choix se justifie par le fait que nos vérifications nous ont permis de comprendre que certains montants issus des rapports peuvent parfois différer des prestations versées à l'assuré. Ci-dessous quelques cas susceptibles d'expliquer ces non alignements et qui rendent difficile l'analyse lorsque le sinistre n'est pas encore clos :

Le paiement des différés :

Le paiement des différés n'est pas toujours effectué pour les sinistres encore ouverts, entraînant un remboursement partiel à l'assuré faisant qu'une partie des montants ne soit pas parfaitement réconciliée.

Les conventions applicables :

Les conventions viennent parfois modifier les prestations versées à l'assurée (remboursement de la franchise notamment pour la convention IRSI par exemple) ou que des frais se rajoutent suite au passage de l'expert générant parfois un déphasage entre l'évaluation de l'expert et le montant réellement indemnisé. Mais aussi, la vie du sinistre peut être amenée à évoluer dans le temps et l'on regarde seulement une vision à l'instant « rapport d'expertise » ce qui peut engendrer certains écarts.

6.3.3 Fiabilisation des montants extraits des rapports :

Après avoir obtenu la base retraitée, la dernière étape est d'arriver à la fiabiliser. En effet, l'idée est d'essayer de voir si les montants qui ont été évalués dans les rapports reflètent bien la réalité des indemnisations de Generali (coût total renseigné dans le système). Cette étape permet également de s'assurer du bon fonctionnement de l'algorithme en comparant les montants des dommages des sinistres issus des rapports d'expertises et les montants que l'on a payé à nos assurés. Autrement dit, si les montants de la base créée à partir des rapports matchent parfaitement avec les montants renseignés au niveau des outils de gestion de Generali. Pour cela, il a fallu dans un premier temps faire un croisement avec la base **DWBDPR** (base des dépenses recettes présentée précédemment) et étudier les cas où les montants ne coïncidaient pas. Ainsi, après plusieurs travaux de vérification ainsi que des investigations auprès du service de gestion, il s'est révélé que plusieurs raisons faisaient que les montants différaient notamment :

La franchise :

Souvent, cette différence était occasionnée par la franchise qui tantôt n'était pas déduite de l'évaluation finale de l'expert. Ainsi, au moment de procéder à l'indemnisation du bénéficiaire, cette franchise est déduite de l'évaluation initiale du total des dommages. De ce fait, il a fallu prendre en compte ces cas dans la suite du traitement de fiabilisation

L'existence d'un nouveau rapport (complémentaire ou rectificatif) :

Le rapport d'expertise peut évoluer faisant suite à la rédaction d'un nouveau rapport. Dans le meilleur des cas ce nouveau rapport d'expertise est censé retracer les informations du précédent dans le but d'avoir des chances d'aligner les montants pendant la fiabilisation. Toutefois, en fonction de l'expert, les détails des rapports varient. En effet, en cas de rédaction d'un nouveau rapport, certains experts ne retracent pas l'historique des évaluations précédentes et donc le dernier rapport **n'annule et remplace pas** le(s) précédent(s). C'est dans ce sens que lorsque le rapport initial fait suite à un rapport complémentaire ou rectificatif deux cas peuvent se présenter :

1 - La dernière vision du rapport retrace toute l'historique des dommages précédemment indemnisés :

Dans ce cas figure, le rapport peut être considéré comme un rapport qui annule le(s) précédent(s). De ce fait, pour ces types de rapports d'expertise, l'idée sera d'identifier et de conserver dans la base de données la dernière vision du rapport reflétant la dernière mise à jour des montants relatifs au sinistre.

2 - Rajout d'informations :

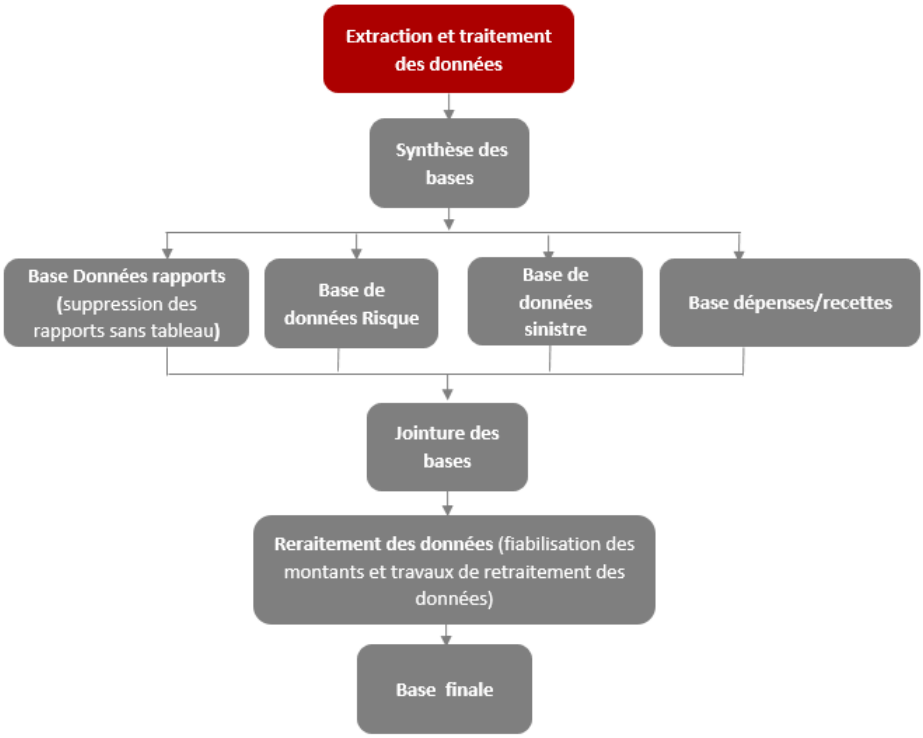
Dans certains cas, l'expert peut décider de rajouter un poste de dommage par exemple et peut se permettre de ne pas retracer l'historique des indemnisations passées. Dans ce cas de figure les montants ne correspondent pas avec ceux renseignés au niveau des outils de gestion Generali. De ce fait, l'identification de la dernière vision ne nous permet pas de fiabiliser les montants. Ainsi, la stratégie adoptée dans le cadre du processus de fiabilisation a été de réconcilier les montants des différents rapports afin de reconstituer le coût total des prestations indemnisées.

La gestion :

Dans certains cas de figure, il est possible d'identifier des écarts entre la vision du rapport et celle des systèmes en lien avec la gestion du sinistre. En effet, il est possible que des frais se rajoutent à posteriori du passage de l'expert par le biais des gestionnaires sinistres.

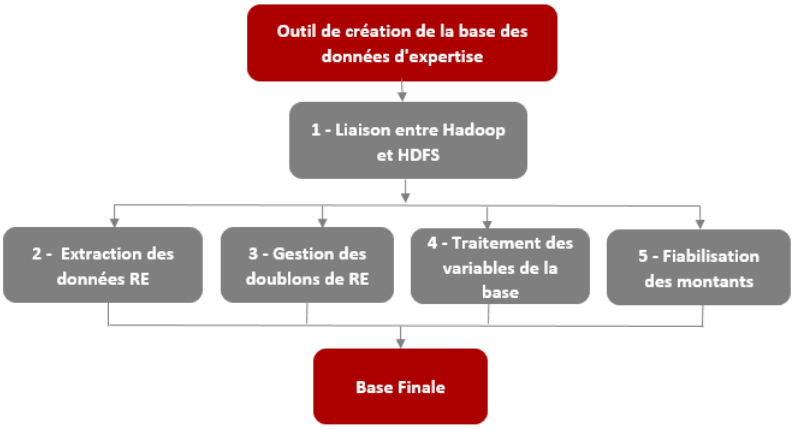
Autres explications :

D'autres écarts sont aussi possiblement expliqués par des erreurs d'algorithme ou sont à relativiser avec des situations particulières. Toutes ces raisons ont fait qu'au final, la base obtenue et qui sera exploitée pour l'étude contient environ **24 707** sinistres. Le schéma ci- dessous permet de visualiser le rapprochement des différents bases de données:



Par ailleurs, ces différents travaux ont ainsi permis de mettre en place des scripts permettant à Generali d'automatiser le traitement depuis l'extraction jusqu'à l'obtention

de la base fiabilisée en passant par la gestion des doublons et l'homogénéisation des informations obtenues. Le schéma ci-dessous permet de visualiser la synthèse des différentes étapes aboutissant à l'obtention d'une base détaillant les sinistres :



Le lancement de l'outil permet aujourd'hui à Generali d'obtenir une base de données permettant d'atteindre une certaine granularité. Cette dernière se décline comme suit :

Tableau de règlement

Désignation	Indemnité immédiate	Indemnité différée	Indemnité totale
Embellissements	963,00 €	0,00 €	963,00 €
Immobilier	4504,70 €	0,00 €	4504,70 €
Contenu	0,00 €	0,00 €	0,00 €
Montant des indemnités nettes	**5 467,70 €**	**0,00 €**	**5 467,70 €**

Liste_Nom_RE	Num_sinistre	Police	Garantie	EMBELISSEMENT	IMMOBILIER	CONTENU	AUTRES	TOTAL
IARD-R214369000.docx	**40432717**	**000AH482665**	**Degat des eaux**	**963,00**	**4 504,70**	**0,00**	**0,00**	**5 467,70**
IARD-R214564456.docx	38339888	000AD784859	Degat des eaux	784,00	0,00	0,00	0,00	784
IARD-R215393209.docx	65004835	000AL871982	Degat des eaux	2 040,85	0,00	0,00	0,00	2040,85
IARD-R198487313.docx	66102141	48989338	Degat des eaux	2 384,80	0,00	0,00	0,00	2384,8
IARD-R247364481.docx	66162040	000AL601391	Degat des eaux	563,47	0,00	0,00	2 250,00	2 813,47
IARD-R198647994.docx	48579830	000AH251060	Degat des eaux	0,00	3 674,89	0,00	0,00	3 674,89
IARD-R215476598.docx	40434278	000AL746743	Degat des eaux	935,68	1 200,21	0,00	0,00	2135,89
IARD-R215208715.docx	67086205	40214673	Degat des eaux	0,00	3 560,62	0,00	0,00	3560,62
IARD-R214597297.docx	37351612	000AA762600	Degat des eaux	2 754,05	0,00	1 162,00	0,00	3916,05
IARD-R198390857.docx	48578578	56353594	Degat des eaux	1 351,30	1 486,25	0,00	0,00	2837,55
IARD-R222302478.docx	38340843	000AA823820	Degat des eaux	904,76	0,00	1 414,75	0,00	2319,51
IARD-R198547881.docx	66103906	000AH514746	Degat des eaux	2 282,31	0,00	0,00	0,00	2 282,31

Figure 24 – Création de la base de données des rapports d'expertise

La base obtenue permet ainsi de décomposer poste par poste les prestations allouées au titre des dommages présents dans les rapports d'expertise. Par ailleurs, l'outil permet également de récupérer d'autres variables en lien avec les données risques. En effet, pendant l'expertise, l'expert essaye de décrire la situation du risque au moment de la survenance du sinistre. Il peut s'agir d'un descriptif du logement par exemple. D'ailleurs ces informations ont été extraites des rapports et nous servirons au traitement des données manquantes quand elles seront absentes dans la base risque. Mais également beaucoup d'autres éléments comme la localisation des dommages, la convention applicable, la conformité du risque etc (**Cf** . [ici]). Les variables relatives au nom du rapport au numéro de dossier et à la police ont été extraites des rapports pour les travaux de jointure.

6.4 Retraitement des données manquantes :

La base d'étude présente quelques données manquantes localisées au niveau de certaines variables. Voici ci-dessous les statistiques concernant les proportions :

Variable	Proportion manquante
Année de construction	63 %
Étage	46 %
Présence de cheminée	56 %
Nombre de pièces	5.2%
Qualité de l'assuré	2.92 %
Franchise	1.8 %

La plupart de ces proportions de manquantes sont importantes (en rouge). Cela peut être relativisé avec le fait que ce sont des informations qui ont commencé à être récoltées que depuis très peu. Par exemple, la question relative à l'année de construction a commencé à être posée qu'en 2016 aux assurés et celle relative à la présence de cheminée est adressée uniquement qu'aux propriétaires de maison. Ainsi, étant donné le volume non négligeable de données manquantes au niveau de ces 3 premières variables (Année de construction, étage, présence de cheminée), nous préférons les écartées de l'étude au lieu de les traiter et d'introduire un biais à nos modèles. Toutefois, nous allons réaliser le remplissage des variables relatives au nombre de pièces, à la franchise et à la de l'assuré (en vert) présentant les proportions les plus acceptables.

6.4.1 Traitement données manquantes via l'extraction et l'usage de données d'expertise :

Initialement les variables **nombre de pièces**, **qualité de l'assuré** et **franchise** de la base risque présentaient respectivement environ 5.2%, 2.92% et 1.8% de données manquantes. Elles ont été complétées par l'extraction des données relatives aux champs équivalents dans les rapports. Au final, la complétion des variables qualité de l'assuré et montant de la franchise a été totale. La variable nombre de pièces quant à elle se retrouve avec une faible proportion de données manquantes (2.2%) par la suite.

Variable	Proportion manquante après premier retraitement
Nombre de pièces	2.2%
Qualité de l'assuré	0 %
Franchise	0 %

Par ailleurs, il a été décidé de ne pas réaliser de suppression mais d'appliquer une méthode mathématique pour en assurer le remplissage. L'approche retenue pour gérer ces quelques données manquantes de la variable nombre de pièces est la technique des K plus proches voisins que nous allons présenter.

6.4.2 Traitement données manquantes : Algorithme des k-plus proches voisins :

L'algorithme des K plus proches voisins ou k-NN (k-Nearest Neighbors) est une méthode basée sur les voisins des données déjà connues. Le principe de la méthode est de trouver lors de la phase d'entraînement les k plus proches voisin puis d'évaluer la moyenne.
Les étapes de la mise en place de l'algorithme des k plus proches voisins se présentent comme suit :

- Dans un premier temps, il faut partitionner nos données en un jeu d'apprentissage et de test
- Choisir un nombre arbitraire de voisin k qu'il faudra optimiser (dans notre cas, nous allons commencer par une petite valeur de k)
- L'algorithme calcul l'ensemble des distances existantes entre les observations afin de trouver les k plus proches voisins selon la distance utilisée. Celle qui sera prise en compte dans notre étude est la distance euclidienne :

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}$$

- Affectation de la moyenne évaluée à chaque nouvelle observation. Il s'agit de la moyenne des k plus proches voisins retenus.

Dans la suite, l'idée est d'évaluer modèle en comparant les données modélisées et les données observées sur l'échantillon de test. Une mesure d'erreur classique qui sera utilisée et que l'on détaillera plus tard est le RMSE permettant de quantifier l'écart entre les observations et les estimations :

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

Pour conclure, une fois avoir finalisé l'implémentation du modèle, la phase suivante est de l'appliquer aux données manquantes de la variable à compléter.

6.5 Retraitement des données aberrantes :

Nous rappelons que la modélisation concerne principalement les sinistres attritionnels. Au global, quelques sinistres étaient au delà du seuil de grave prédéfini et été considérés comme des valeurs extrêmes après vérification. Aucun sinistre est détaché de l'étude à ce stade de l'analyse. Les contrats présentant une indemnisation qui peut paraître élevée sont conservés dans la base compte tenu de leur faible matérialité (0.23% de la base totale) Dans la suite, nous allons modéliser l'embellissement et l'immobilier reflétant 94% de la charge de de dommage indemnisée. Les deux derniers postes **"Contenu et Autres"** sont des postes très peu sinistrés dans notre échantillon.

Ainsi compte tenu du faible volume de sinistres et du faible poids de ces derniers en terme d'enjeu financier, ils ne seront pas inclus dans la suite de l'étude.

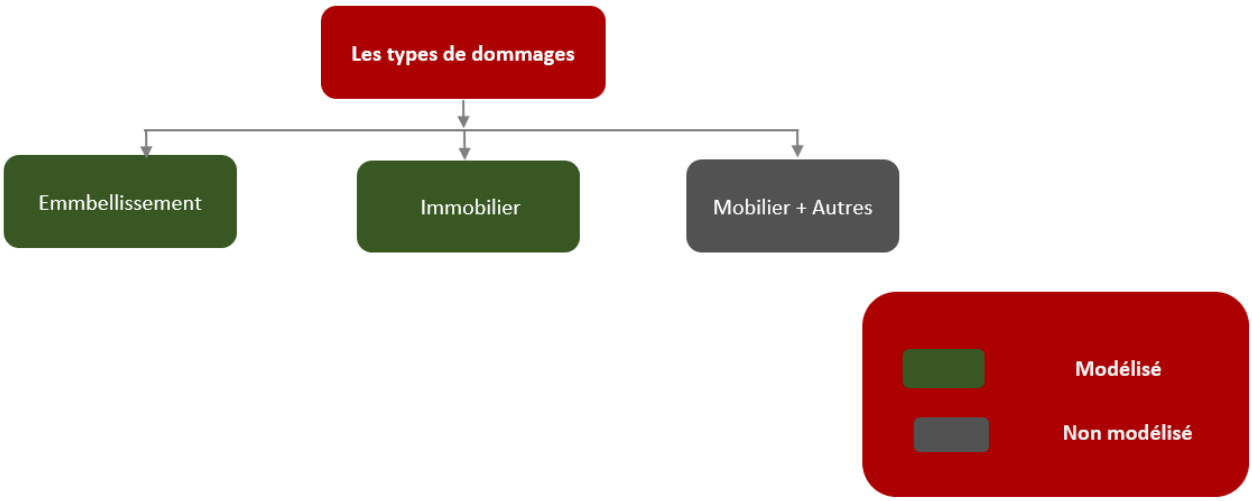


Figure 25 – Périmètre de modélisation

Par la suite, des travaux d'analyse univariée sur la sinistralité permettra de détecter les potentielles valeurs extrêmes voir aberrantes. D'abord, via une analyse l'évolution du coût moyen de l'embellissement et de l'immobilier par modalité de variable. Mais aussi, nous avons analysé la distribution de chacune des charges modélisées par modalité sur toutes les variables explicatives à disposition a l'aide d'un boîte à moustache. Ci-dessous un graphe décrivant le fonctionnement de cette dernière :

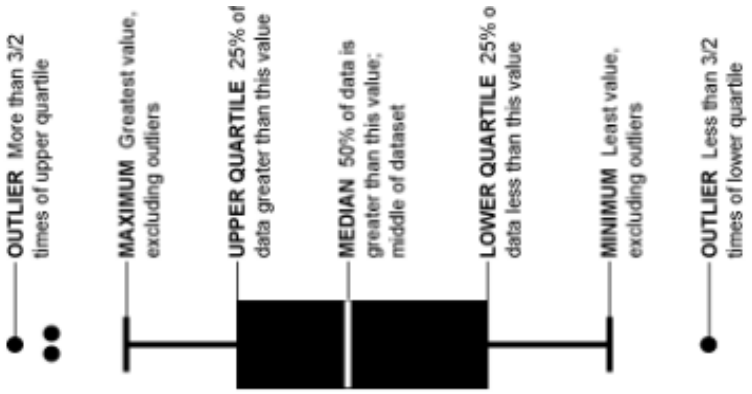


Figure 26 – Boîte à moustache

Il permet ainsi de visualiser les quartiles, la médiane ainsi que les bornes min et max. Mais également de repérer et d'étudier les potentielles valeurs aberrantes de l'échantillon (outliers) susceptibles de polluer l'apprentissage des modèles.

6.6 Statistiques descriptives sur la répartition de la charge des expertisés :

Répartition de la charge relative aux dommages indemnisés :

Nous avons réalisé quelques statistiques descriptives concernant la ventilation de la charge des indemnisations versées au titre des différents dommages recensés par l'expert suite à un sinistre. Nous rappelons que l'étude concerne que les sinistres clos et donc qui ont déjà fait l'objet d'un règlement. Le graphe ci-dessous permet d'apprécier la répartition de ces sinistres expertisés clos de notre échantillon d'étude :

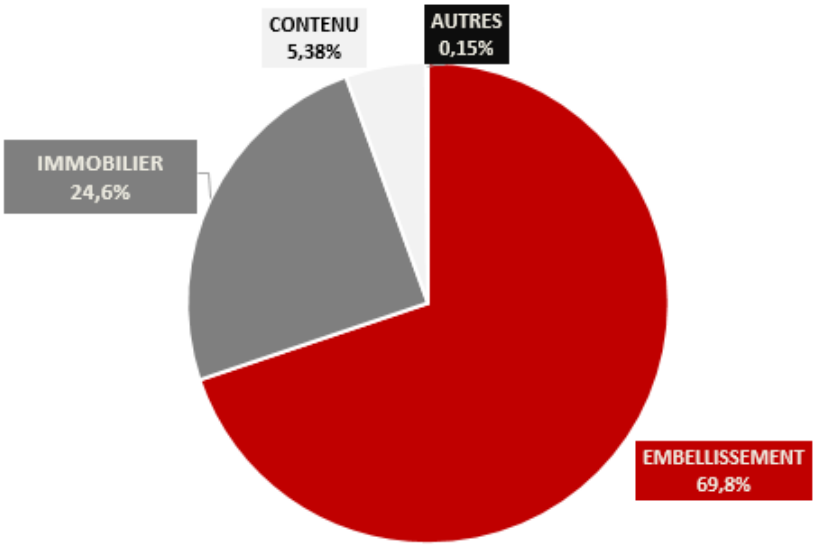


Figure 27 – Statistiques sur les types de dommages en dégât des eaux

Ces statistiques montrent que l'embellissement se présente comme étant le type de dommage le plus sinistré concentrant environ 70% de la charge totale des dommages, suivi de l'immobilier comptant pour environ 25% et enfin moins de 6 % est expliquée par le mobilier et autres frais spécifiques. Il s'en suit que l'embellissement est le type de dommage le plus important à la fois en terme de nombre de sinistres et de charge. Le "contenu" ainsi que les autres dommages regroupés dans le poste "Autres" sont très peu sinistrés dans le portefeuille présentant une volumétrie relativement faible au niveau de la base d'étude.

6.7 Statistiques descriptives sur l'embellissement et l'immobilier:

Qualité de l'occupant:

A titre informatif, il peut être intéressant de remarquer que notre portefeuille sinistré est plus composé de propriétaires que de locataires. Pour cette variable, les locataires semblent être moins risqués en terme de coût moyen que les propriétaires quel que soit le poste (embellissement/immobilier). Il est probable que cette différence de coût moyen soit dû au fait que les propriétaires ont plus tendance à investir sur le logement que les locataires susceptibles de déménager à tout moment.

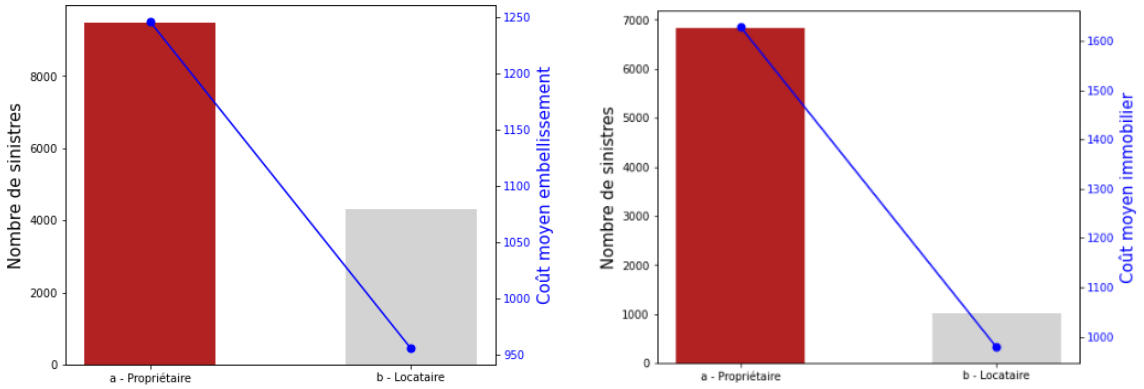


Figure 28 – Répartition du coût moyen de l'embellissement et de l'immobilier selon la qualité de l'assuré

Le réseau:

La variable réseau présente 2 modalités (Agent/courtier) considérées comme des distributeurs. Cette variable est prise en compte dans la segmentation actuelle du risque.
En effet, elle peut être révélatrice du degré de risque auprès des distributeurs étant donné qu'un mauvais intermédiaire va avoir tendance à ramener des contrats pas très rentables pour la compagnie. Cette variable apparaît comme étant peu discriminante dans notre étude et ne semble donc pas avoir un impact significatif dans notre portefeuille.

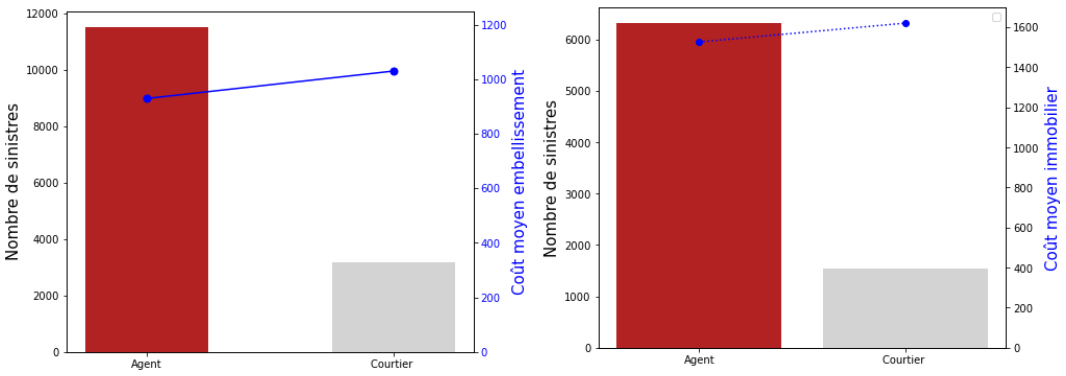


Figure 29 – Répartition du coût moyen de l'embellissement et de l'immobilier en fonction du réseau

Le type d'habitation:

Le variable type d'habitation apparaît peu discriminante pour l'explication du coût moyen de l'embellissement et très discriminante pour celui de l'immobilier. Cette différence de coût entre type de logement semble venir du fait que les maisons étant souvent plus grandes auront plus tendance à présenter des biens d'aménagements immobiliers (terrasse, clôture, piscine etc.) que les appartements ainsi qu'une surface habitable plus développée. De ce fait en cas de dégâts des eaux affectant un parquet par exemple, le dommage peut être beaucoup plus important dans les maisons que dans les appartements (généralement plus petits).

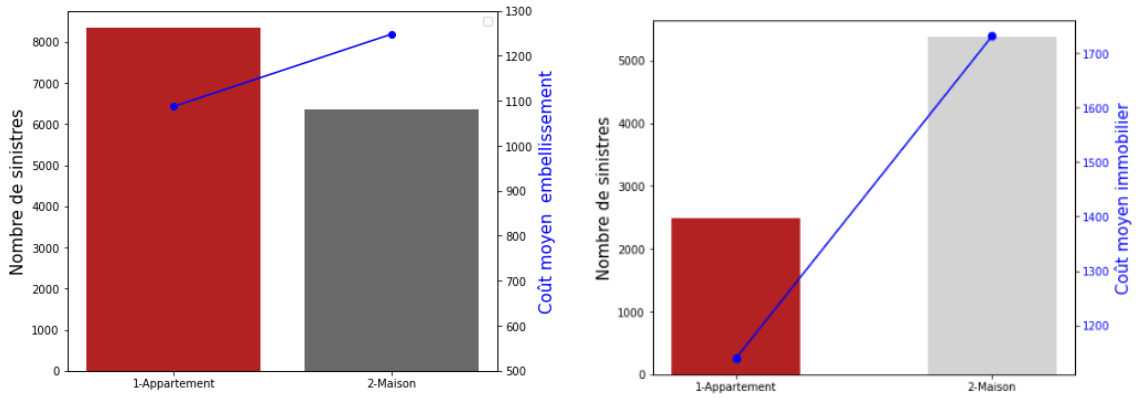


Figure 30 – Répartition du coût moyen de l'embellissement et de l'immobilier en fonction du type d'habitation

La présence de dépendance:

Une dépendance peut être considérée comme un bâtiment qui n'est pas rattaché directement au bâtiment principal et qui n'est pas destinée à l'habitation. Initialement, cette variable renseignait sur la surface de la dépendance subdivisée en tranches. Toutefois, dans notre échantillon la majorité des habitations ne possédaient pas de surface de dépendance ou étaient affectées à la tranche 1. Nous avons décidé de transformer cette variable. En effet, dans la suite, elle sera considérée comme une indicatrice renseignant sur la présence ou l'absence de dépendance. Ainsi, l'analyse montre que le fait de disposer d'une dépendance a un impact sur le coût moyen pour les deux postes de dommage. Malgré que le portefeuille soit minoritairement composé de domiciles avec dépendance, le coût moyen est beaucoup plus élevé pour ces types d'habitation.

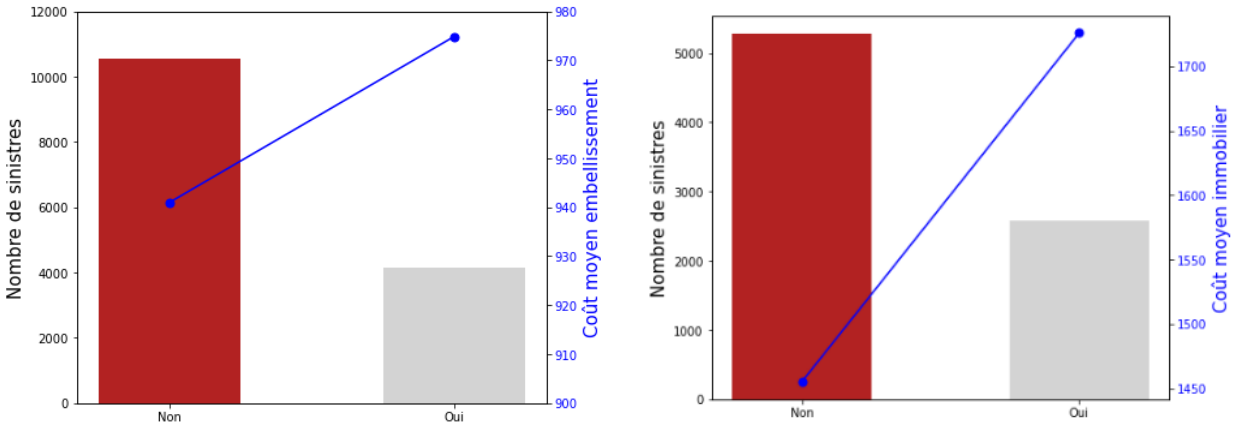


Figure 31 – Répartition du coût moyen de l'embellissement et de l'immobilier selon la présence ou non de dépendance

Le type de résidence :

Pour cette variable aussi, nous observons que le portefeuille sinistré est majoritairement composé de résidences principales. Toutefois, la modalité **résidence secondaire** apparaît comme étant plus risquée en terme de coût moyen que les autres modalités. Cela est sûrement à relativiser avec le fait que ces types de logements sont fortement exposés à des risques liés à l'inhabitation, au défaut d'entretien pouvant créer des sinistres dégâts des eaux via par exemple des phénomènes de rupture des canalisations liées au froid.

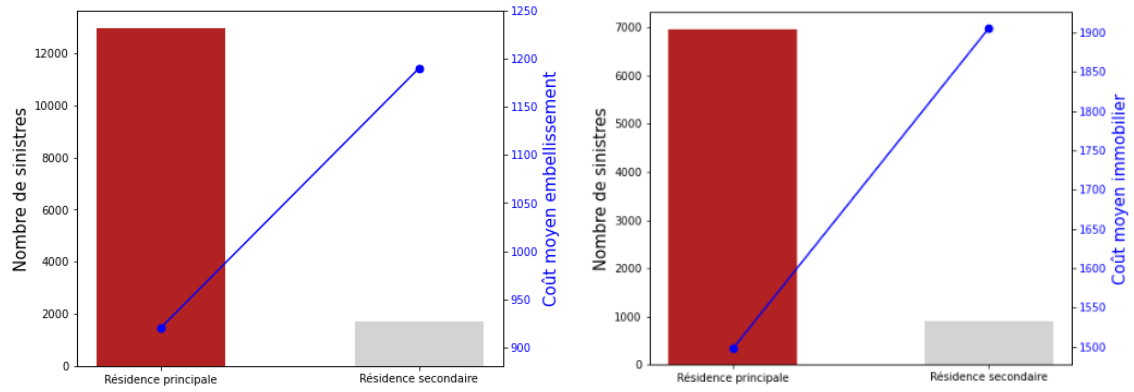


Figure 32 – Répartition du coût moyen de l'embellissement et de l'immobilier selon le type de résidence

Le nombre de pièces:

Cette variable présente plusieurs modalités. On peut remarquer ici une forte concentration du portefeuille au niveau des classes 1 à 9 et une tendance à la hausse du coût moyen en fonction du nombre de pièces pour ces premières classes. Toutefois, au-delà de 10 pièces cette évolution ne se confirme pas. Cela peut être relativisée avec la faible représentativité des classes supérieures engendrant ainsi une forte volatilité du coût moyen.La même tendance est observée concernant la répartition du coût de l'immobilier en fonction du nombre de pièces dans le logement (avec une forte volatilité cette fois-ci observée à partir de 8 pièces ainsi qu'une tendance croissante beaucoup moins visible). Cette variable sera de ce fait retravaillée dans la suite avant d'être injectée dans un modèle. Ci-dessous, nous présentons cette tendance du coût moyen pour l'embellissement.

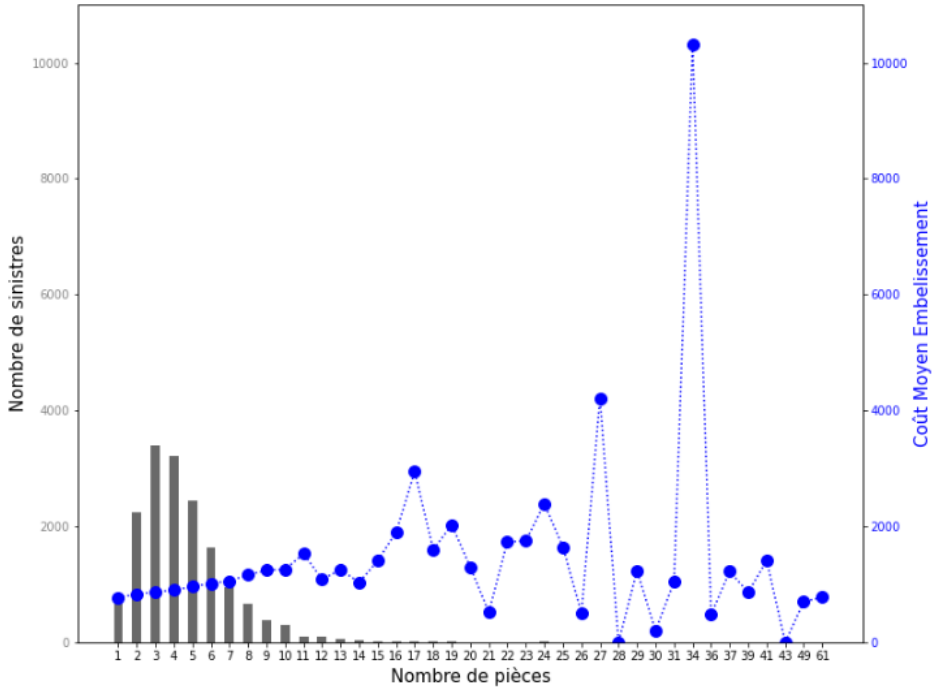


Figure 33 – Répartition du nombre de sinistres et du coût moyen de l'embellissement en fonction du nombre de pièces

A travers cette répartition, nous avons pu identifier 4 sinistres présentant un coût moyen d'environ 10.000 euros. L'identification de ces sinistres et la consultation des rapports d'expertise a permis de vérifier l'exactitude de ces montants qui pourraient à priori sembler erronées. Nous décidons ainsi de ne pas les supprimer étant donné que ces valeurs ne sont pas aberrantes.

6.8 Classification des variables:

L'une des problématiques de la mise en place d'un modèle est le nombre de modalités que peut présenter une variable. La discrétisation est une méthode en réponse à ce problème consistant à prendre des données numériques afin de les transformer en un ensemble de valeurs discrètes par regroupement selon des critères de ressemblance.
La discrétisation n'est pas une méthode de pré-traitement indispensable, mais est utile pour rendre un algorithme d'apprentissage plus efficace. En effet, elle permet de regrouper les valeurs statistiques d'une variable en classes facilitant ainsi l'apprentissage du modèle. Il existe d'autres méthodes permettant de mettre en place ces groupements de valeurs en vue de rajouter de la valeur au modèle.

La variable nombre de pièces :

Comme dit précédemment, cette variable présente une tendance à la hausse du coût moyen qui se dessine pour les modalités les plus sinistrées. Toutefois, le même cas de figure ne s'observe pas pour les modalités à plus de 10 pièces. En effet, la représentativité dans notre portefeuille des modalités au-delà de 10 pièces est faible créant une forte volatilité pour ces classes. Ainsi, il a été décidé de les regrouper en une seule classe. Cela permet de renforcer la segmentation du risque. La tendance du coût moyen après regroupement est la suivante:

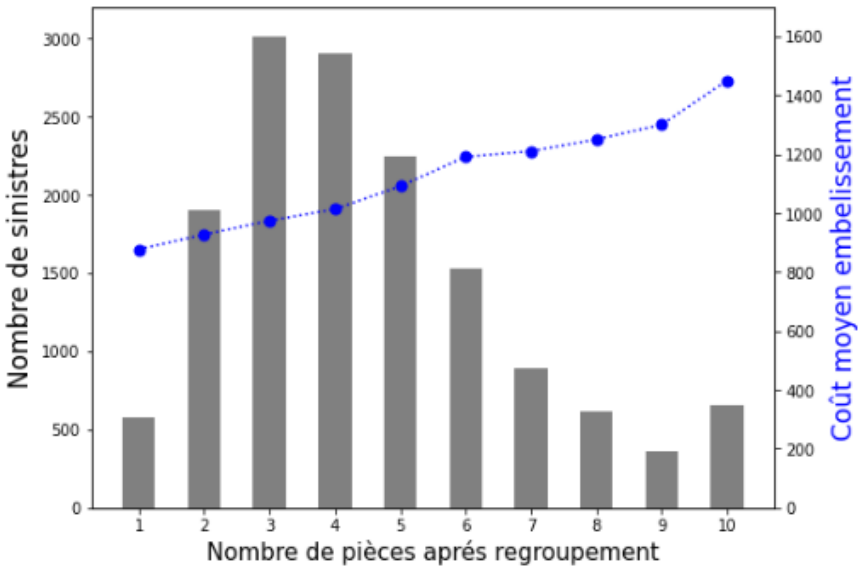


Figure 34 – Répartition du coût moyen de l'embellissement en fonction du nombre de pièces après regroupement

Nous pouvons observer une tendance à la hausse du coût moyen en fonction du nombre de pièces qui se dessine ce qui est assez logique. En effet, il paraît assez cohérent de penser que le coût moyen augmente en fonction du nombre de pièces disponibles dans le logement.

6.8.1 L'algorithme des KMeans (considérations théoriques et application) :

a - Approche mathématique :

Le principe d'un clustering est de laisser la machine apprendre à classer les données selon leur ressemblance. Ainsi, un des algorithmes les plus utilisés est le KMeans clustering.
Il s'agit d'un algorithme itératif qui pour un ensemble de données $(x_1, .., x_n)$ va minimiser la distorsion par rapport au centre des différents clusters $(\boldsymbol{\mu_1}, ..., \boldsymbol{\mu_k})$ et les classes $(\mathbf{z_1}, .., \boldsymbol{z_n})$.
L'algorithme des centres mobiles s'exécute comme suit :

— Étape d'affectation des points du dataset au centre le plus proche :

$$z_{ik}^{(t)} = \begin{cases} 1 & \text{si } k = \underset{z \in 1,...,k}{\arg\min} ||x_i - \mu_z|| \\ 0 & \text{sinon.} \end{cases}$$

— Après l'étape d'affectation, il s'agira de calculer la moyenne de chaque cluster pour y déplacer par la suite le centre : $\forall k = 1, .., K$

$$\boxed{\mu_k^{(t+1)} = \sum_{i=1}^{n} \frac{z_{ik}^{(t)} x_i}{z_{ik}^{(t)}}}$$

Toutefois, il est possible parfois selon la position initiale de nos centroïdes que ces derniers convergent vers de mauvaises positions. D'où la nécessité d'exécuter plusieurs fois l'algorithme en modifiant à chaque fois la position initiale des centroïdes afin d'assurer une bonne classification.
De ce fait, pour chaque résultat donné, la solution retenue sera celle qui va minimiser la distance entre chaque point d'un groupe et son centre. L'idée était à la fin de créer des groupes avec la distance la plus faible au sens d'une métrique. La métrique retenue pour le clustering sera à nouveau la distance euclidienne qui sera définie comme suit :

$$||x_i - \mu_k|| = d(x_i; \mu_k) = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2}$$

Dans le cadre pratique, cet algorithme prend en entrée une matrice et un nombre de clusters, puis donne en sortie un objet de la classe Kmeans présentant la liste des clusters ligne à ligne. C'est ainsi que cette méthode nous permettra ainsi de mettre en place une base de données qui pour chaque sinistre associera la zone et la moyenne du coût des sinistres de cette dernière.

b -Choix du nombre de classes:

Dans le cadre du clustering, il est nécessaire de déterminer le nombre de classes optimal. Étant dans le cadre d'un apprentissage non supervisé (il n'y a pas de vérité absolue), le fait de choisir soi même le nombre de clusters peut poser un problème. Ainsi, un mauvais choix du nombre de clusters **k** peut engendrer une mauvaise segmentation.
Il existe différentes méthodes permettant de déterminer le k optimal. Toutefois, la technique qui sera utilisée ici sera la méthode de la coude ou **EBLOW Method** en anglais qui consiste à représenter l'évolution du coût du modèle (inertie) en fonction du nombre de clusters. Le but étant de minimiser la variance intra-classe et de déterminer une zone de coude afin d'être en mesure de trouver le nombre de clusters optimal. C'est-à-dire celui qui nous permet de réduire au mieux le coût du modèle tout en veillant à conserver un nombre raisonnable de clusters. La variance des clusters se calculent de la manière suivante :

$$\mathbb{V} = \sum_{j} \sum_{x_i -> C_j} D(C_j, x_i)^2$$

— C_j : le centroïde
— $D(C_j, x_i)$: distance séparant le centre du cluster et l'observation x_i
— x_i : l'observation i du cluster ayant pour centroïde C_j

c - Exemple d'application : Le zonier de la garantie dégât des eaux

Les analyses préliminjaires ont montré que cette variable semble être intéressante dans le sens où elle discrimine fortement le coût moyen. Toutefois, elle se décline en plusieurs modalités ce qui peut avoir un impact sur le modèle durant l'apprentissage ainsi qu'une forte volatilité pour certaines classes qui sont très peu représentées dans l'échantillon. Il convient donc de procéder à son retraitement en regroupant les modalités qui se ressemblent le plus et apportent la même information sur la sinistralité. Ainsi, cette variable a été retraitée via cet algorithme des KMeans pour plusieurs valeurs de K.

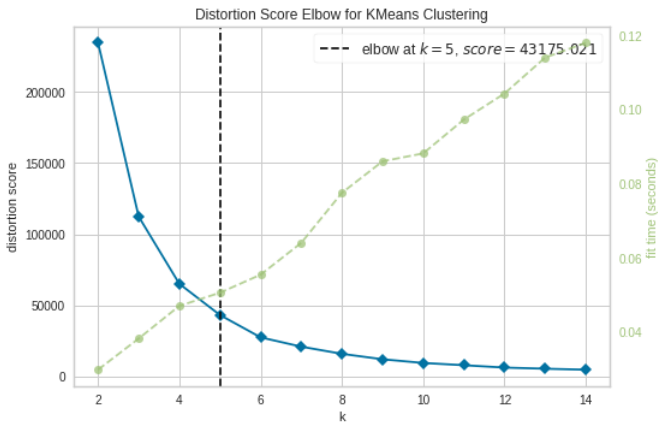


Figure 35 – Détermination du k optimal par la méthode de la coude

La visualisation du graphe ci-dessus permet de voir l'évolution de l'inertie intra-groupe en fonction du nombre de classes et de trouver la zone de coude. C'est à ce niveau que le clustering ne s'améliore plus quand la valeur de k augmente. La valeur optimale de K est 5, ce qui signifie que nous allons créer 5 groupes ou classes. On observe que ces classes formées segmentent le portefeuille en cinq risques distincts.

Cluster	Coût moyen	% sinistre
1	805	48.7%
2	983	22.8%
3	1102	19.2%
4	1271	8.6%
5	1304	0.7%

Nous constatons un déséquilibre au niveau de la répartition des sinistres mais aussi observons une certaine variabilité des coûts moyens. Cette méthode de classification regroupe dans la classe 5 que 0.7% des sinistres. Il y'a donc de grande chance de ne pas avoir de sinistres relatifs à cette catégorie dans la base de test. Ainsi, il semble que cette méthode ne soit pas la plus adaptée à nos données. Dans la suite, nous décidons de réaliser le même traitement qui a été effectué sur notre variable nombre de pièces en procédant par un regroupement des classes les moins représentées. Nous obtenons les résultats suivants :

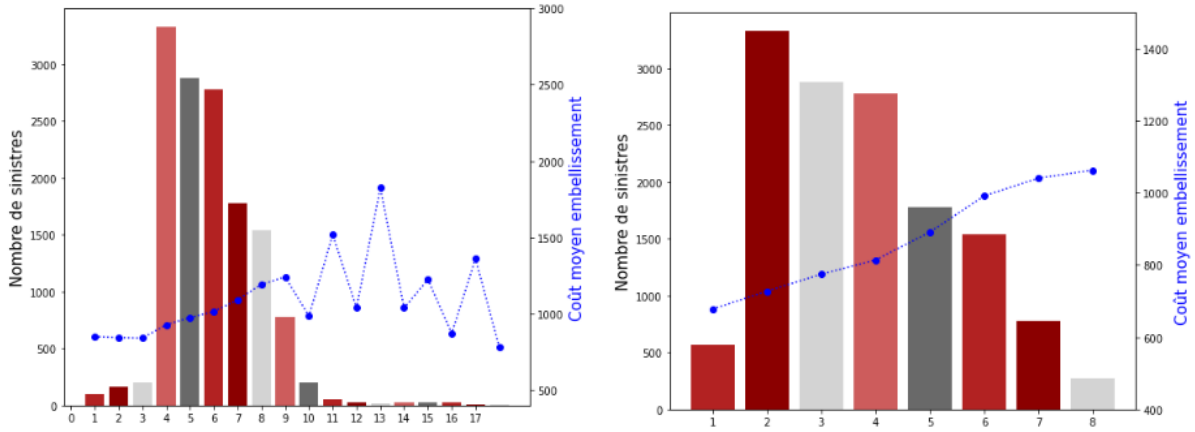


Figure 36 – Zonier avant et après retraitement

Les modalités 1,2 et 3 étant proches et présentant quasiment les mêmes coûts moyens ont été regroupées au sein d'une seule modalité (modalité 1-3). Aussi, nous constatons qu'à partir de la variable brute plus le niveau de la zone est élevé, plus le risque l'est. Toutefois, il est plus difficile de capter une tendance pour les zones de niveau 10 et plus. En effet, au même titre que la variable relative au nombre de pièces, nous observons une certaine volatilité du coût moyen au delà de la 10 ème modalité. Ce phénomène est dû à la faible représentativité de ces modalités dans notre échantillon. Laisser ces modalités telles qu'elles ne nous facilite pas la modélisation. Nous faisons ainsi le choix de les regroupées en une seule et même modalité (10-18). Les résultats après retraitement peuvent être visualisés dans le graphe de droite ci-dessus.

6.9 Études de dépendance:

Cette partie sera focalisée sur l'étude des corrélations entre nos différentes variables. En effet, l'étude de corrélation est une étape importante dans le sens où une multicolinéarité prononcée entre les variables prévisionnelles peut s'avérer problématique.
Ainsi, pour évaluer la liaison entre nos différentes variables explicatives du modèle, il existe différents tests. Le test utilisé dépendra de la nature du croisement de variables.

Nature du croisement	**Test**
Qualitative vs Qualitative	V de Cramer
Quantitative vs Qualitative (2 modalités)	Test de Man-Withney
Quantitative vs Qualitative (>2 modalités)	Test de Krustal-Wallis

Dans le cadre de cette étude, étant donné que nous sommes en présence que de variables qualitatives, nous utiliserons par la suite que le test du V de Cramer. Ce test permettra de détecter l'intensité des liaisons existantes entre nos différentes variables.

6.9.1 Le V de Cramer (considérations théoriques)

Le V de Cramer est l'une des mesures d'association les plus connues en statistiques. Il peut être considéré comme une amélioration du test du χ^2. C'est dans ce sens que nous allons dans un premier temps présenter ce dernier.
Le principe du test du χ^2 est de mettre en place une comparaison entre la répartition théorique T_j et la répartition observée O_j. Ainsi, quand le test du χ^2 permet juste de détecter un lien entre 2 variables, le V de Cramer lui va plus loin en évaluant l'intensité du croisement des deux variables qualitatives. Il est défini comme étant la racine carré du rapport entre le χ^2 et l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1.

$$V = \sqrt{\frac{\chi^2}{n * (Min(k,l) - 1)}}) \tag{1}$$

— n : Taille de l'échantillon
— k,l: le nombre de modalités relatif à chacune des deux variables en jeu .
— χ^2 : Le test du chi 2
Ainsi, les valeurs du test sont comprises entre 0 et 1. De ce fait :
— Plus V proche de 1, plus l'intensité de la liaison sera forte.

L'analyse montre des corrélations entre : la qualité de l'occupant et le type habitation. Mais aussi entre le type d'habitation et le zonier. Dès lors, il serait judicieux de réaliser un croisement de variables corrélées et mettre à jour le calcul du V de Cramer avec ce nouveau croisement. De ce fait, nous décidons de réaliser dans un premier temps un croisement entre qualités de l'occupant et le type d'habitation qu'on appellera "statut de l'occupant".Ci-dessous les résultats obtenus :

V Cramer	Qualite occupant	Résidence	Type habitation	Zonier	Nbre_enfant	Dépendance	véranda	Franchise	Nombre pièces	Survenance
Qualite occupant	1	0.19	0.51	0.23	0.08	0.19	0.12	0.14	0,28	0,03
Résidence		1	0.10	0.075	0.17	0.03	0.038	0.05	0,17	0,06
Type habitation			1	0.37	0.11	0.18	0.2	0.13	0,6	0,046
Zonier				1	0.07	0.11	0.10	0.16	0.15	0,05
Nbre_enfant					1	0.06	0.01	0.01	0.18	0,05
Dépendance						1	0.04	0.07	0.21	0,16
veranda							1	0.03	0.22	0,03
Franchise								1	0.09	0,19
Nombre pièces									1	0,03
Survenance										1

Figure 37 – V de Cramer sur les différentes variables

Le croisement a clairement diminué l'intensité des liens qui existaient entre certaines variables.

V Cramer	Qualite x Habitat	Résidence	Zonier	Nbre_enfant	Dépendance	véranda	Franchise	Nombre pièces	Survenance
Qualite x Habitat	1	0.22	0.23	0.15	0.19	0.20	0.16	0,24	0,03
Résidence		1	0.075	0.17	0.03	0.038	0.05	0,17	0,06
Zonier			1	0.07	0.11	0.10	0.16	0,15	0,05
Nbre_enfant				1	0.06	0.01	0.01	0.18	0,05
Dépendance					1	0.04	0.07	0.21	0,16
veranda						1	0.03	0.22	0,03
Franchise							1	0.09	0,19
Nombre pièces								1	0,03
Survenance									1

Figure 38 – V de Cramer actualisé après croisement

Ci-dessous la répartition du coût moyen et des sinistres de la nouvelle variable issue du croisement entre la qualité de l'occupant et le type d'habitation.

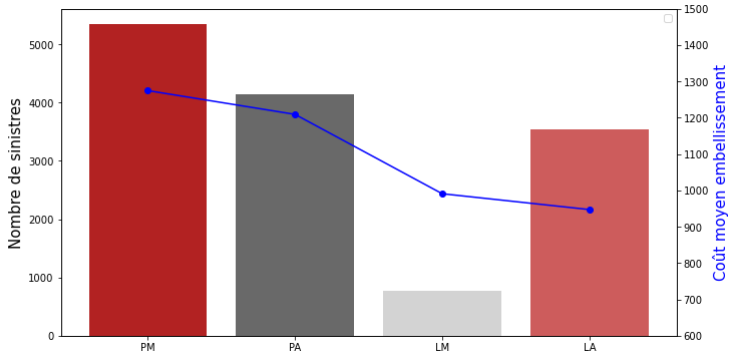


Figure 39 – Répartition du nombre de sinistres et du coût moyen de l'embellissement en fonction de la variable QualitéxHabitat

- LA : Locataire - Appartement
- LM : Locataire - Maison
- PA : Propriétaire- Appartement
- PM : Propriétaire - Maison

Par ailleurs, le principal point fort du test du V de Cramer est la fiabilité. En effet, contrairement au test du χ^2 qui a tendance à être très sensible à la structure par modalité et à la variation de la taille de l'échantillon, le V de Cramer va corriger ce défaut dans ses évaluations. Ci-dessous, la représentation du corrélogramme des différentes variables :

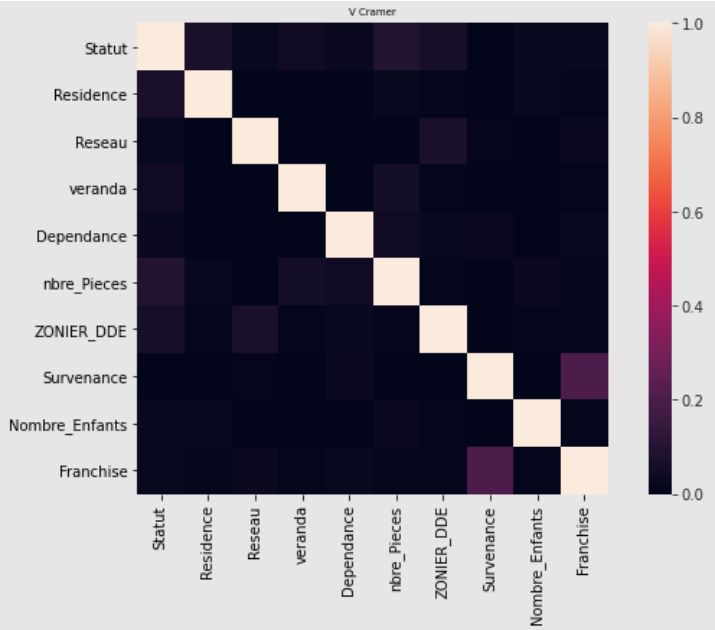


Figure 40 – Corrélogramme du V de Cramer sur les différentes variables

Certaines variables telle que par exemple la surface de l'habitation a été exclue de l'étude car très corrélée avec le nombre de pièces. Ce qui est assez cohérent étant donné que cette dernière est définie à partir de la surface du logement. Cette approche est en ligne avec ce qui se fait dans la segmentation actuelle du risque.

Part III

Mise en place des modèles statistiques et analyse des résultats

7 Outils théoriques et application

7.1 Indicateurs de performance du modèle

7.1.1 Racine de la moyenne des erreurs au carré (RMSE):

Elle permet d'évaluer l'écart existant entre les valeurs prédites par un modèle et les valeurs réelles afin de se faire une idée sur la performance du modèle.
Ainsi, il va permettre de mesurer l'erreur de prédiction mais à tendance à donner un poids élevé aux erreurs larges. Elle est définie comme suit :

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

— y_i : la valeur de l'observation i
— $\hat{y}_i$: la valeur estimée du modèle pour l'observation i

De manière plus détaillée, en supposant que $y = f(x) + \epsilon$ représente la fonction à modéliser et que $\hat{f}$ notre estimateur créé à partir de la base d'apprentissage. Alors l'erreur quadratique pourra être définie comme étant :

$$\begin{aligned} RMSE &= \sqrt{\mathbb{E}[(y - \hat{f}(x))^2]} \\ &= \sqrt{\mathbb{E}[(y^2 - \hat{f}^2(x) + 2y\hat{f}(x))]} \\ &= \sqrt{\mathbb{V}\text{ar}(y) + \mathbb{V}\text{ar}(\hat{f}(x)) + \mathbb{E}[f(x) - \hat{f}(x)]^2} \\ &= \sqrt{\mathbb{V}\text{ar}(\hat{f}(x)) + \mathbb{V}\text{ar}(y) + Biais^2(\hat{f}(x))} \end{aligned}$$

C'est ainsi que l'erreur est subdivisée en 3 termes :

$$MSE = Biais^2 + Variance + irréductible$$

$$RMSE = \sqrt{Biais^2 + Variance + irréductible}$$

7.1.2 L'erreur moyenne absolue (MAE):

L'erreur moyenne absolue (Mean Absolute Error en anglais) se présente comme étant ma moyenne des écarts entre les valeurs observée et prédites. Elle se définie comme suit :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

Remarque : Ces fonctions de perte classiques seront utilisées notamment pour évaluer les performances du modèle mais aussi principalement pour les challenger entre eux.

7.1.3 L'indice de Gini et la courbe Lorenz :

La courbe de Lorenz permet de représenter la distribution d'une variable aléatoire dans une population et est une illustration graphique qui était initialement utilisée dans le cadre de la mesure des inégalités salariales. Cette courbe permet de mesurer l'indice de Gini. En effet, l'indice de Gini se définit comme étant un coefficient permettant d'évaluer la dispersion d'une population donnée et se calcul à partir de la courbe de Lorenz. L'idée est de voir si une bonne partie de l'information est captée par une partie de la population.

$$Gini = \frac{AireA}{AireA + AireB}$$

Dans le cadre de cette étude de coût, l'abscisse représentera la part cumulée des contrats sinistrés et l'ordonnée désignera la part cumulée des prestations.

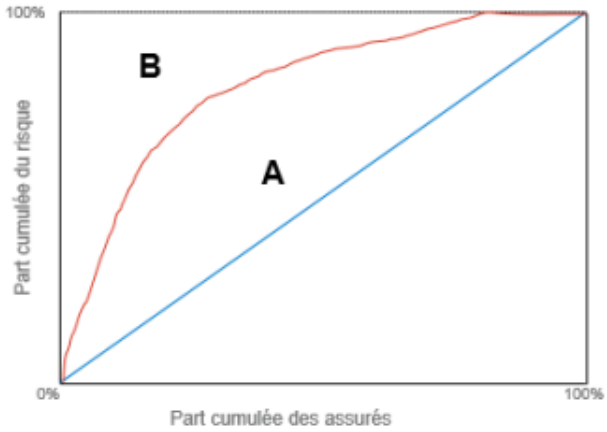


Figure 41 – Illustration de la courbe de Lorenz

7.1.4 Autres mesures graphiques :

D'autres mesures de performance à caractère graphique sont également utilisées pour visualiser les biais que peuvent potentiellement présenter les modèles. Il s'agit notamment de la comparaisons des distributions prédites et observées par modalité de variable. Mais également, une étude sera portée sur la distribution et l'analyse des résidus.

8 Modèle linéaire généralisé (GLM):

Les modèles linéaires généralisés permettent d'étudier la liaison entre un ensemble de variables explicatives et une variable dite variable réponse ou cible **Y**. Ces modèles sont très utilisés dans le secteur de l'assurance notamment dans le cadre de la tarification. Dans les modèles linéaires généralisés, la variance de cette variable cible est très dépendante des variables explicatives. Le modèle linéaire généralisé se présente comme étant une extension de la régression linéaire. Pour rappel, le but de la régression linéaire est de modéliser l'espérance $E[Y|X]$ définie comme étant une fonction g des variables explicatives de sorte que :

$$\boxed{Y = g(X) + \epsilon}$$

ϵ étant l'erreur de la modélisation de Y par son espérance conditionnelle.
Par ailleurs, les hypothèses de la régression linéaire se déclinent comme suit :

$$\begin{cases} E[Y|X] = X\beta \\ Y|X \sim \mathcal{N}(X\beta, \sigma) \\ X \perp Y \\ X \ est \ une \ matrice \ de \ rang \ plein \end{cases}$$

8.1 Composante déterministe :

La composante déterministe du GLM aussi appelé prédicteur linéaire peut être définie sous la forme d'une combinaison linéaire comme on peut le voir ci-dessous :

$$g(E[Y]) = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_p\beta_p$$

Ainsi, dans le cadre de la modélisation via un GLM, de nouvelles variables peuvent se déduire des variables explicatives initialement introduites dans le modèle :

— $X_4 = X_2^2$ pour la prise en compte du caractère non linéaire de X_2 par exemple

Durant le processus de modélisation, il est important de sélectionner les variables les plus pertinentes pour expliquer la variable réponse.

8.2 Composante aléatoire :

La loi de **Y** doit appartenir à une famille exponentielle. Une loi de probabilité appartient à une famille exponentielle si elle admet une densité de la forme :

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

Dans ce cas, l'espérance et la variance s'écrivent comme suit :

$$\begin{cases} \mathbb{E}[Y] = \mu = b'(\theta) \\ Var[Y] = \sigma^2 = a(\phi) * b''(\theta) \end{cases}$$

— $\Phi \in \mathbb{R}$: Paramètre de dispersion souvent considéré comme un paramètre de nuisance (c'est à dire un paramètre qui n'est pas d'un intérêt immédiat mais qui doit être prise en compte dans l'analyse des paramètres d'intérêt.

— $\theta \in \mathbb{R}$: Paramètre canonique

— a : Une fonction de $\mathbb{R}^*$

— b : Une fonction appartenant à $\mathbb{R}$ et dérivable deux fois

Il existe différentes lois qui appartiennent à la famille exponentielle. La loi utilisée dépendra du phénomène modélisé. Dans le cadre de la modélisation du coût des sinistre il est souvent privilégié d'utiliser une loi de type Gamma.

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$	$c(y_i,\phi)$
Normale$(\mu_i\,;\sigma^2)$	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\}$
Poisson(μ_i)	$\log(\mu_i)$	1	ϕ	$\exp(\theta_i)$	$-\log y!$
Binomiale$\frac{1}{m_i}(m_i\,;\mu_i)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	ϕ	$\log(1+\exp\theta_i)$	$\log\binom{m_i}{m_i y_i}$
Gamma$(\mu_i\,;\alpha)$	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-\log(-\theta)$	$\alpha\log(\alpha y) - \log y - \log\Gamma(\alpha)$
Inverse Gaussienne$(\mu_i\,;\sigma^2)$	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$

Figure 42 – Exemple de familles exponentielles

8.3 Fonction de lien :

Elle permet de spécifier la relation entre la composante aléatoire et la composante déterministe. La fonction de lien est supposée monotone et différenciable. En effet, contrairement au modèle linéaire où l'espérance est directement modélisé, le GLM va permettre de modéliser la fonction de lien de cette espérance. Ainsi, les fonctions de liens les plus courantes sont :

- La fonction logarithme qui est multiplicative $log(x) = x$ qui va donner :

$$g(E[Y]) = \exp \beta_0 + \exp \beta_1 * X_1 + ... + \exp \beta_p * X_p$$

- La fonction identité $g(x) = x$ qui est linéaire :

$$g(E[Y]) = \beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p$$

- La fonction logit $g(x) = log(\frac{x}{1-x})$ souvent utilisée pour une distribution de Bernoulli:

$$g(E[Y]) = \frac{\exp(\beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p)}{1 + \exp(\beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p)}$$

- La fonction inverse $g(x) = log(\frac{1}{x})$:

$$g(E[Y]) = \frac{1}{1 + \exp(\beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p)}$$

Loi	Fonction de lien	$g(\mu)$	Support
Gamma	Inverse	$\frac{1}{\mu}$	$\mathbb{R}^*$
Poisson	Logarithme	$\ln \mu$	$\mathbb{N}$
Normale	Identité	μ	$\mathbb{R}$
Binomiale	Logit	$\ln \frac{\mu}{1-\mu}$	$\{0, 1\}$

8.4 Estimation des paramètres :

Une fois la loi déterminée et la fonction de lien choisie, la phase suivante consiste à faire l'estimation des paramètres. Cela est possible par maximisation de la vraisemblance. $\beta_1, ..., \beta_p$. Étant donné une loi de probabilité Y et les réalisations $(y_1, ..., y_n)$ indépendantes et identiquement distribuées. En supposant que les Y_i suivent une loi exponentielle, la vraisemblance du modèle se présente comme suit :

$$L(y|\theta, \phi) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi)$$

Toutefois, dans la pratique, il est beaucoup plus aisé de prendre la log-vraisemblance afin de transformer le produit en somme :

$$l(y|\theta, \phi) = \ln(L(\beta)) = \ln(\prod_{i=1}^{n} f(y_i|\theta_i, \phi))$$

$$\Rightarrow l(y|\theta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi)$$

$$\Rightarrow l(y|\theta, \phi) = \sum_{i=1}^{n} \underbrace{c(y_i, \phi) + \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}}_{:=l_i}$$

Ainsi, déterminer les paramètres optimaux revient à trouver les β_j $j \in 1, ..., p$ de sorte que :

$$\hat{\beta} = argmax\ l(y|\theta, \phi)$$

$$\frac{\partial l(y|\theta, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} ln(f(y_i|\theta_i, \phi)) = 0$$

Cela revient ainsi à résoudre l'équation :

$$\frac{\partial l_i}{\partial \beta_j} = 0 \iff \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \nu_i}\frac{\partial \nu_i}{\partial \beta_j} = 0$$

Ce qui nous amène à avoir :

$$\begin{cases} \frac{\partial l_i}{\partial \theta_i} = \frac{Y_{i-\mu_i}}{a(\phi_i)} \\ \frac{\partial \theta_i}{\partial \mu_i} = \frac{Var(Y_i|X_i)}{a(\phi_i)} \\ \frac{\partial \mu_i}{\partial \nu_i} = \frac{g^{-1}(\nu_i)}{\partial \nu_i)} \\ \frac{\nu_i}{\partial \beta_j} = \frac{X_i\beta}{\partial \beta_j} \end{cases}$$

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{Var[Y_i]}\frac{\partial \mu_i}{\partial \nu_i} x_{i,j} = 0, j = 1, ..., p$$

Dans le cadre pratique, ces équations ne possèdent pas de solutions explicites. Ce problème est souvent résolu grâce à des méthodes itératives comme l'algorithme de Newton-Raphson.

8.5 Sélection de variables :

Dans le cadre de la modélisation GLM, il est possible de faire appel à des méthodes statistiques à l'exemple des méthodes de sélection automatiques présentées ci-dessous :

Sélection pas à pas :

a - forward selection (méthode ascendante) :

C'est une méthode qui va tester dans un premier temps le modèle avec comme seule variable explicative la constante mais qui va rajouter pas à pas d'autres variables dans le modèle suivant les résultats obtenus des tests de significativité. Cela permet de choisir à chaque itération la variable la plus significative et s'arrêtera quand la performance du modèle ne s'améliore plus. L'inconvénient avec cette méthode est qu'une fois une variable introduite elle ne peut plus être retirée même si elle n'est plus considérée comme étant significative.

b - Backward selection (méthode descendante) :

Cette méthode est assez similaire à la première mais va procéder en sens inverse. En effet, elle va partir d'un modèle de départ avec toutes les variables susceptibles d'expliquer la variable réponse et à chaque itération va supprimer la moins pertinente en se basant sur des tests de significativité. De ce fait, si après ce test des variables ne deviennent plus pertinentes, cette méthode va supprimer la moins significative d'entre elles. Ainsi, nous pouvons comprendre qu' à chaque étape du processus, des variables peuvent perdre leur significativité d'une itération à une autre.

c - Stepwise selection (méthode mixte) :

Cette méthode est une approche assez similaire à la méthode ascendante mais avec une possibilité de retirer des variables déjà introduites en amont. En effet, à chaque étape elle va introduire une nouvelle variable et tester si toutes les variables précédemment retenues restent significatives. Ainsi, en présence de variables non significatives pour le modèle, la moins significative d'entre elles est retirée. L'itération continue jusqu'à ce que toutes les variables soient significatives. Pour faire un choix de modèle, il est très courant de faire appel à des critères fondées sur la pénalisation de la vraisemblance telles que l'AIC (Akaike Informative Criterion) et le BIC.

d - Test de significativité (Type III) :

Ce test permet de mesurer la significativité des variables en terme de contribution à l'explication de la variable réponse. Il est basé sur la déviance du modèle.
Soit les hypothèses suivantes :

$$\begin{cases} H_0 : la\ variable\ X_j\ n'est\ pas\ significative\ au\ modèle \\ H_1 = X_j\ est\ significative \end{cases}$$

La statistique de test associé étant :

$$S = D_M^j - D_M^{j-1}$$

S suivant une loi du χ^2 qui aura pour nombre de degrés de liberté la différence des degrés de liberté du modèle et du sous modèle.

- M^{j-1} : Modèle étudié avec prise en compte de la variable X_j
- M^{j-1} : Le modèle est étudié sans la prise en compte de la variable X_j

Ainsi, la statistique de S suit une loi du type χ^2 à n-p degrés de liberté.

- n : nombre de degrés de liberté du modèle de base
- p : nombre de degrés de liberté du sous-modèle

Le principe du test étant de trouver le quantile d'ordre alpha de la loi du χ^2 puis d'évaluer la p-value $P(S > q_\alpha)$ associée :

$$\begin{cases} P(S > (q_\alpha)) < \alpha \quad On\ rejette\ H_0 \\ Sinon \qquad X_j\ n'est\ pas\ très\ significative\ pour\ expliquer\ la\ variable\ réponse \end{cases}$$

Avec :

- α : représente le risque de première espèce
- q_α : le quantile d'ordre α de la loi du χ^2

8.6 Sélection de modèle :

La déviance :

Comme dans le cas des modèles linéaires, une fois que les paramètres sont estimés pour un GLM, il est important de vérifier que le modèle reflète bien la réalité et qu'il décrit bien les données. Toutefois, étant donné que dans un GLM on capte très peu d'information sur les résidus, l'idée derrière sera donc d'établir une comparaison ou un arbitrage entre le modèle étudié et le modèle saturé (le modèle contenant toutes les variables explicatives).

Ainsi, un des critères permettant d'évaluer cet écart est la déviance. En effet, cette dernière va permettre de calculer la différence entre la vraisemblance du modèle estimé et celle du modèle saturé (modèle présentant autant de paramètres que d'observations).

$$D = 2\phi log(\frac{L_{sat}}{L}) = 2\phi \left\{log(L_{sat}) - log(L)\right\}$$

— L_{sat} : vraisemblance du modèle saturé
— L : vraisemblance du modèle étudié

Les critères de sélection automatique :

L'AIC et le BIC se présentent comme étant des compromis entre la qualité du modèle et sa complexité. Ce sont des critères qui utilisent le principe du maximum de vraisemblance et vont pénaliser les modèles contenant plusieurs variables afin de pallier à des problèmes de sur-apprentissage en mettant en avant le modèle le plus parcimonieux. Avec cette méthode,le meilleur modèle qui sera choisi sera celui qui aura tendance à minimiser l'AIC. ou le BIC. Pour un modèle avec k paramètres, l'AIC sera défini comme étant :

$$\begin{cases} AIC = -2ln(L) + 2k \\ BIC = -2ln(L) + k * ln(n) \end{cases}$$

- k : représente le nombre de paramètres
- L : représente la vraisemblance du modèle
- 2k : représente la pénalisation
- n : le nombre d'observations

Toutefois, la parcimonie est plus accentuée avec le BIC étant donné qu'il pénalisera plus le nombre de paramètres dans le modèle. De ce fait en cas d'une égalité d'AIC entre 2 modèles, le choix sera orienté sur celui qui minimisera plus le BIC.

Les résidus de déviance :

Ces résidus permettent d'évaluer la contribution de chaque observation à la déviance du modèle. Ils sont définis via un terme noté d_i reflétant la contribution de l'i-ème observation y_i à la déviance du modèle.

$$R = signe(y_i - \hat{y}_i)\sqrt{d_i} \;\; avec \;\; d_i = 2\log L(y_i, y_i, \phi) - 2\log L(y_i, \hat{y}_i, \phi)$$

Avec ϕ représentant le paramètre de dispersion. Ils permettent de réaliser une validation du modèle. L'examen de ces résidus est essentiellement graphique en veillant à s'assurer que les observations ne s'écartent pas trop des valeurs observées.

Paramétrage du modèle :

Afin d'ajuster au mieux le modèle, une validation croisée sera effectuée afin de s'assurer qu'il n'est pas impacté par le sur-apprentissage pendant l'entraînement. Cela permettra d'évaluer la performance du modèle à se généraliser sur de nouvelles données.
Ainsi, dans une approche basique, la base de données d'origine est séparée en 2 : une base d'entraînement (servant d'apprentissage) et une base de validation pour le tester. Toutefois, en procédant ainsi le modèle est évalué qu'une seule fois ce qui ne semble pas être une approche idéale. De ce fait, la validation croisée permet d'évaluer le modèle plusieurs fois et de lui présenter la totalité de l'échantillon de la base. Le principe est le suivant :

- Dans un premier temps, la base d'origine est divisée en base d'apprentissage et de test contenant respectivement 80% et 20% des données.
- Dans un second temps, la base d'apprentissage est à son tour partitionnée aléatoirement en K blocs.
- Par la suite, l'idée sera d'entraîner et d'évaluer k fois le modèle. Chaque itération permettra de lancer l'apprentissage du modèle sur K-1 blocs et de le tester sur le dernier bloc.

	All data (100 %)				
	Training data (80 %)				Test data (20%)
	Fold 1 (20%)	Fold 2 (20%)	Fold 3 (20 %)	Fold 4 (20%)	
Itération 1	Test	Train	Train	Train	
Itération 2	Train	Test	Train	Train	
Itération3	Train	Train	Test	Train	
Itération 4	Train	Train	Train	Test	
Validation finale					Test data

Figure 43 – Cross validation

La figure ci-dessus permet d'avoir une illustration visuelle de la méthode. Par ailleurs, le score de cross validation final est obtenu en faisant la moyenne des scores individuels.

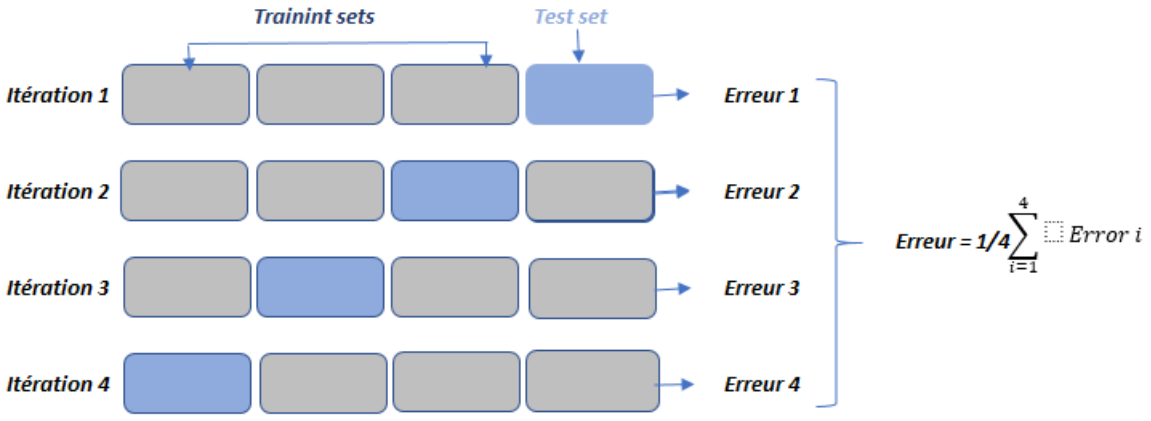


Figure 44 – Illustration du calcul de l'erreur moyenne en cross validation

8.7 Application sur la mise en place d'un modèle pour l'embellissement :

Nous avons décidé d'appliquer un modèle linéaire généralisé à la distribution Gamma et à la fonction de lien logarithme. En effet, cette distribution est très cohérente avec nos observations étant donné que la loi gamma ne prend en compte que les coûts strictement positifs. Ainsi, la base de données utilisée pour la modélisation de ce poste est obtenue en filtrant sur l'ensemble des montants non nuls de la colonne correspondante aux coûts de l'embellissement. Cela permet d'obtenir une base constituée d'environ 14500 sinistres.

Sélection de variables :

Dans le cadre de la sélection des variables, nous allons privilégier la méthode du stepwise avec comme critère de minimisation l'AIC et le BIC. Nous avons également étudié la variation de la déviance a l'ajout successif des variables pour avoir une visibilité sur la contribution de chacune d'entre elles à la baisse de la déviance. Par ailleurs, des tests ont été également réalisés sur les variables les moins significatives pour juger de leur apport ou non d'informations. Il s'en est suivi que ces variables n'ont pas trop d'incidence sur l'explication de la variable réponse. Ainsi, les différents prédicteurs retenus au final sont :

- Le nombre de pièces
- Le type de résidence
- Le zonier
- La survenance
- Le statut de l'assuré

Ces variables sont les plus contributrices pour expliquer le coût moyen de l'embellissement. Par ailleurs, les variables relatives à la franchise, la localisation de véranda ou encore au nombre d'enfants dans l'habitation semblent être les moins discriminantes pour le modèle.

Regroupement de modalités :

Durant la phase d'implémentation du GLM, nous avons eu à réaliser quelques regroupements de modalités. Par exemple, notre première sortie GLM montrait que majoritairement l'ensemble des modalités semblait être significatives mise à part la modalité correspondante à l'année de survenance 2014. Cela est à relativiser avec le fait que soit cette modalité présente une exposition très faible, soit son coefficient estimé est très proche d'un des coefficients d'une modalité existante. La seconde hypothèse a été validé consistant à regrouper cette modalité avec celle de l'année de survenance 2015 présentant le même coefficient β.

8.7.1 Analyse des résultats et Validation du modèle :

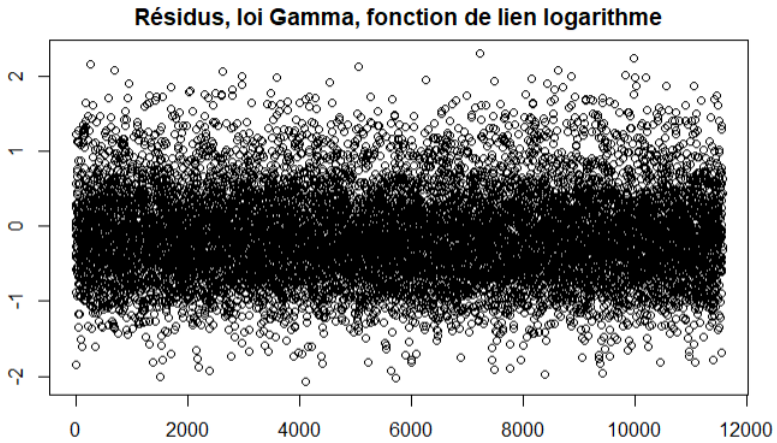
Nous allons évaluer la performance du modèle sur quelques indicateurs numériques notamment le RMSE, l'erreur totale ou encore le MAE. Mais aussi un indicateur graphique tel que la courbe de Lorenz. Une analyse sera faite également sur les coefficients estimés du modèle GLM ainsi que les résultats fournis par le GLM.

Analyse des résidus et des coefficients estimés par le modèle :

Le modèle GLM retourne des valeurs relatives aux coefficients estimés. Dés lors il se trouve important d'analyser la cohérence de ces coefficients. Ainsi, l'analyse qui a été faite sur ces derniers repose sur le principe suivant :

— Un coefficient positif montre que le coût moyen de la modalité étudiée est plus élevé que le coût moyen de la modalité fixée comme référence.
— A contrario, un coefficient négatif pour une des modalités considérées montre que le coût moyen de cette modalité en question est inférieur au coût moyen de la modalité de base ou modalité de référence.

Les résultats obtenus sont en accord avec l'attendu.
Par la suite, au delà du fait que nos résidus suivent loi normale centré en 0, nous avons représenté la distribution de ces résidus en fonction des observations pour une éventuelle analyse des cas extrêmes . Ci-dessous le graphe illustratif :



Nous pouvons remarquer une bonne cohérence du modèle en examinant la distribution du nuage de points des estimations en fonction des observations. Notamment au niveau des valeurs les plus représentées. Les résidus semblent être centrés et ne présentent pas de forme particulière synonyme que notre modèle est effectivement validé. Nous constatons également que l'écrasante majorité de ces résidus est localisée dans l'intervalle [–2;+2].Une faible proportion d'observations contribuent au mauvais ajustement du modèle et sont celles qui sont moins représentées dans l'échantillon impliquant probablement une plus grande difficulté de modélisation.

Analyse de la déviance :

La déviance standardisée permet d'avoir une idée sur la validité du modèle. En effet, le modèle pourra être considéré comme valide dans le cas où la déviance standardisée du modèle est plus petite que le quantile d'ordre 95% d'une χ^2 à n-p degrés de liberté.

Quelques statistiques du modèle	Valeurs
Déviance	4395
AIC	179065
ddl	11550
Quantile χ^2 d'ordre 95 %	11801

La déviance standardisée du modèle étant largement inférieure au quantile à 95% de la loi du χ^2 à 11 550 degrés de liberté, nous ne rejetons pas l'hypothèse de validité du modèle.

Validation croisée :

Nous décidons par la suite de réaliser un 4-fold cross validation.
A chaque itération, le Gini et la RMSE ont été évalués. Les résultats de la validation croisée se présentent comme suit :

Indicateur	Itération 1	Itération 2	Itération 3	Itération 4
RMSE	716.87	734.33	713	718.04
GINI	25.40%	24.92%	25.79%	25.22%

Nous obtenons des variations assez stables du Gini et du RMSE à chaque itération. Le RMSE moyen s'évalue à 720.56 et le le Gini moyen à 25.33 %.

Les indicateurs numériques :

L'évaluation du RMSE, du MAE et de l'indice de Gini sur la base d'apprentissage et la base de test ne montre pas d'écart significatif. En effet, ces indicateurs sont assez stables sur les échantillons d'entraînement et de test. Cela permet de déduire que le modèle n'est pas en sur-apprentissage.

Indicateur	Base d'apprentissage	Base de test
RMSE	708.33	728.25
MAE	510	518
GINI	25.58%	25.02%

Représentation de la courbe de Lorenz :

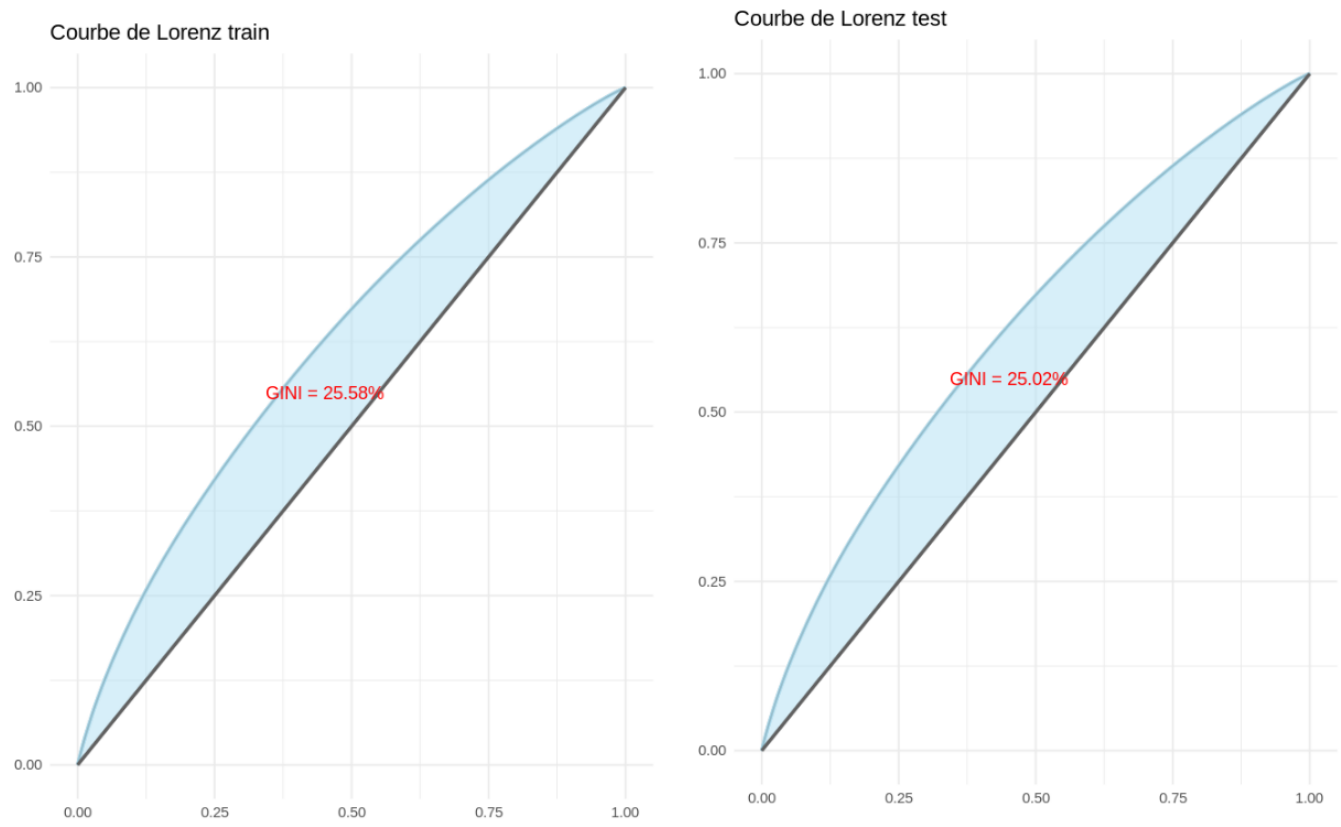


Figure 45 – Courbes de Lorenz du GLM sur les bases train et test de l'embellissement

Nous obtenons des indices Gini proches entre l'apprentissage et le test ce qui permet de confirmer que notre modèle ne sur-apprend pas. Toutefois, contrairement à un modèle de fréquence, le Gini évalué sur notre modèle de coût peut paraître faible. Cela est en lien avec le fait que le Gini à tendance à pénaliser la variabilité des montants sur un modèle de coût.

8.8 Application à la mise en place d'un modèle pour l'immobilier :

Cette partie est dédiée à la mise en place d'un GLM coût moyen pour l'immobilier. La base de modélisation est obtenue en appliquant un filtre sur l'ensemble des valeurs strictement positives de la colonne correspondante aux montants des dommages relatifs a l'immobilier et permet d'obtenir constituée d'environ 7800 sinistres uniques.

8.8.1 Sélection des variables explicatives :

Résultats avec la méthode stepwise

Pour la modélisation du coût moyen de l'immobilier, nous allons comme précédemment associer la méthode du stepwise avec comme critère de minimisation l'AIC et le BIC à un test de significativité des variables afin de sélectionner les variables les plus pertinentes. La méthode de sélection basée sur l'AIC à tendance à rendre significative une variable à plusieurs modalités nous allons en parallèle analyser la variation de la déviance à l'ajout de variables. En effet,la méthode **stepwise** avec comme critère de minimisation l'AIC sélectionne l'ensemble des variables. Toutefois, en analysant en parallèle la sensibilité de la déviance a l'ajout successif des variables, nous nous avons remarqué que les variables les plus contributrices à la baisse de la déviance sont :

- Le nombre de pièces
- Le statut de l'assuré
- La dépendance

Le zonier et les autres variables ne semblent pas être significatives en terme d'amélioration du modèle. Par ailleurs, le réseau de distribution semble être celle qui impacte moins la baisse de la déviance du modèle. Par la suite, un test de significativité a été appliqué aux variables les moins significatives. Il s'en est suivi que l'hypothèse d'apport d'information de ces variables n'a pas été retenu raison pour laquelle elles ne serviront pas à la modélisation de ce poste. Par ailleurs, concernant la modélisation de l'immobilier, c'est la modalité **Locataire-Maison** qui apparaît comme très peu exposée et pas très significative. En effet, cette seconde modalité affiche un coefficient β assez proche de celui de la modalité de référence (**Locataire-Appartement**) permettant de conclure qu'elles ne sont pas significativement différentes. Il semblerait donc que le type d'habitation n'a pas un impact sur le coût moyen de l'immobilier étant locataire. Toutefois, elle est très significative étant propriétaire. De ce fait, nous décidons donc de les regrouper en une seule et même classe.

Analyse des statistiques du modèle :

Vérifions que la déviance standardisée du modèle est plus petite que le quantile d'ordre 95% d'une χ^2 à n-p degrés de liberté. Au vu des sorties du GLM, la déviance standardisée du modèle étant largement inférieure au quantile à 95% de la loi du χ^2 à 6369 degrés de liberté, nous ne rejetons pas l'hypothèse de validité du modèle.

Quelques statistiques du modèle	Valeurs
Déviance	3053
ddl	6369
Quantile χ^2 d'ordre 95 %	6475

8.8.2 Analyse des résultats et Validation du modèle :

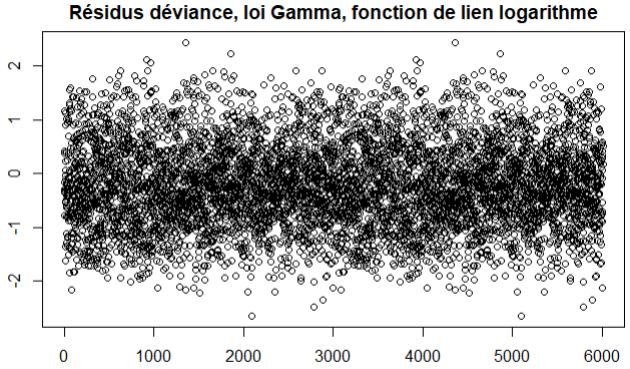
Nous allons évaluer la performance du modèle sur quelques indicateurs numériques notamment le RMSE, l'erreur totale ou encore le MAE. Mais aussi un indicateur graphique tel que la courbe de Lorenz. Une analyse sera faite également sur les coefficients estimés du modèle GLM ainsi que les résultats fournis par ce dernier.

Analyse des résidus et coefficients estimés par le modèle :

Le modèle GLM retourne des valeurs relatives aux coefficients estimés. Dés lors il se trouve important d'analyser la cohérence de ces coefficients. Ainsi, l'analyse qui a été faite sur ces coefficients repose sur le principe suivant :

- Un coefficient positif montre que le coût moyen de la modalité étudiée est plus élevé que le coût moyen de la modalité fixée comme référence.
- A contrario, un coefficient négatif pour une des modalités considérées montre que le coût moyen de cette modalité en question est inférieur au coût moyen de la modalité de base ou modalité de référence.

Par ailleurs, au delà de l'analyse de la répartition de nos résidus qui suivent une loi normale centrée en 0, nous avons analysé le graphique des résidus de déviance du modèle. Ci-dessous, une illustration du nuage de points des résidus en fonction des observations :



Graphiquement, il est possible de voir que globalement les résidus sont symétriques. Ces derniers sont majoritairement situés entre -2 et 2 et ne présentent pas de structuration particulière. Certaines observations peuvent sembler mal estimées. Cela reste négligeable car quelques points sont concernés et ne sont pas influents, ils ne faussent donc pas le modèle.

Les indicateurs numériques :

Ci-dessous les métriques évaluées sur les deux bases :

Indicateur	Base d'apprentissage	Base de test
RMSE	890	928
MAE	758	787
GINI	21.68%	19.29%

L'évaluation des métriques d'erreur à savoir le MAE, la RMSE ainsi que l'indice de Gini sur la base d'apprentissage et la base de test ne montre pas d'écart très significatif sur les 2 échantillons. Toutefois, ces métriques d'erreurs de prévision sont un peu plus élevées que celles obtenues pour l'embellissement. En effet, les métriques obtenues semblent être un peu hétérogènes d'un poste à l'autre dans l'échantillon.

Représentation de la courbe de Lorenz :

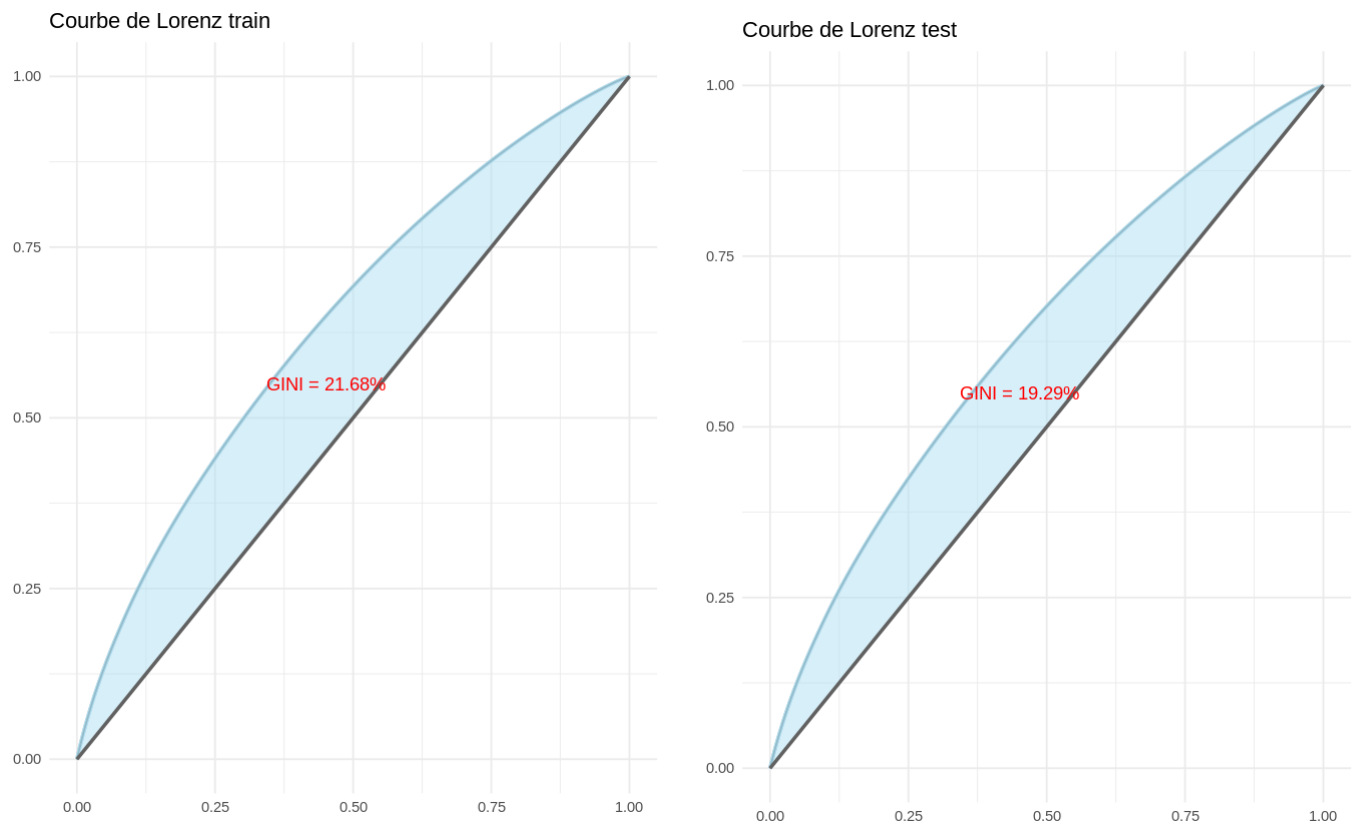


Figure 46 – Courbes de Lorenz sur les bases train et test de l'immobilier

Une fois encore nous obtenons des indices de Gini qui peuvent paraître faibles. Mais cela est expliqué par le fait que ce dernier à tendance à pénaliser la variabilité des données observées sur un modèle de coût.

9 Cadre théorique du Random forest:

9.1 Les arbres de décision (CART) :

Nous allons dans un premier temps faire un petit rappel des arbres de décision avant d'aborder le principe du Random Forest.
Les arbres de décision sont des algorithmes très utilisés pour répondre à des problématiques de classification ou encore de régression. La construction de ces arbres est basée sur une hiérarchie de questions/réponses permettant d'aboutir à une décision.

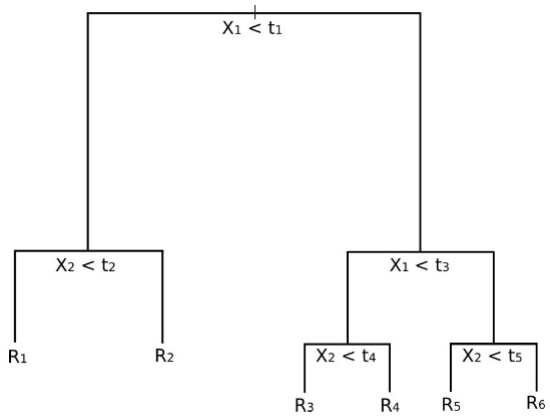


Figure 47 – Arbre CART

L'algorithme CART cherche à mettre en place un arbre idéal en procédant comme suit :

Arbre maximal :

Le principe est de construire des noeuds et de favoriser le meilleur partitionnement des variables. Le critère de découpage étant basé sur l'optimisation du niveau d'hétérogénéité. Dans le cadre d'une régression, l'hétérogénéité d'un noeud est définie par sa variance :

$$D_N = \frac{1}{|N|} \sum_{i \in N} (\bar{Y}_i - Y_N)^2$$

— $|N|$ désigne l'effectif au sein du noeud N
— $\bar{Y}_i$ désigne la moyenne de la variable des y_i dans le noeud N

Dans le cadre d'une classification, l'indice de Gini est utilisé pour évaluer l'hétérogénéité d'un noeud. Mathématiquement, il se présente comme suit :

$$D_N = \sum_{k=1}^{n} p_k(N)(1 - p_k(N))$$

— $p_k(N)$ représente la part des individus de classe k dans le noeud N

Phase d'élagage :

Le principe de l'élagage consiste à mettre en place une sous suite de l'arbre maximal présenté précédemment que l'on nommera l'arbre T_{max} et d'en choisir l'arbre optimal.
Dans la suite, on définit un critère de pénalisation de l'erreur d'ajustement :

$$Crit_\alpha(T) = \hat{S}(T) + \alpha|F_T|$$

— $|F_T|$ désigne le nombre de feuilles de l'arbre T
— $\hat{S}(T)$ désigne quant à lui l'erreur d'ajustement de l'arbre T

9.2 Le Bagging :

Le Random forest est basé sur la théorie des arbres de décisions. En effet, ces derniers présentent un certains nombres de défauts à savoir la sensibilité à l'ajout de nouvelles observations par exemple ou une dépendance assez forte à l'échantillon initiale. Ainsi, le principe du Bagging est de mettre en place plusieurs estimateurs décorrrélès et prendre la moyenne des prévisions issues de ces différents modèles.
Cela permet ainsi de réduire la variance et donc d'améliorer les performances du modèle. La méthode du Bagging consiste à construire différents jeux de données en faisant un tirage aléatoire avec remise sur la base de données initiale . En considérant $D_n = (x_1, y_1), ..., (x_n, y_n)$ ainsi que M échantillons boostrap tirés de manière aléatoire :

$$\hat{f}_{bag} = \frac{1}{M}\sum_{m=1}^{M} \hat{f}_m(x)$$

— $\hat{f}_m$: Modèle entraîné sur l'échantillon m.
Le Bagging sera un moyen très efficace qui permettra en agrégeant plusieurs modèles de minimiser la variance de l'estimation.
Ainsi, si les arbres sont 2 à 2 corrélés nous aurons :

$$V_{forêt} = Var\left(\hat{f}_{bag}\right) = \frac{1}{M^2} Var\left(\sum_{m=1}^{M} \hat{f}_m(x)\right)$$

$$= \frac{1}{M^2}\sum_{m=1}^{M}\left(\sigma^2 + \sum_{l=1, l\neq k}^{M} \Psi\sigma^2\right)$$

$$= \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\left(1 + (M-1)\Psi\right)$$

$$= \frac{\sigma^2}{M}\left(1 + (M-1)\Psi\right) = \Psi\sigma^2 + \frac{1-\Psi}{M}\sigma^2$$

— Ψ : Coefficient de corrélation entre 2 arbres
— σ^2 : Variance des arbres

9.3 L'algorithme du Random Forest :

Les forêts aléatoires peuvent être considérés comme une amélioration de l'algorithme des arbres de décision. En effet, l'inconvénient majeur avec ces derniers est qu'ils ont tendance à sur-apprendre au niveau de la base d'entraînement.
Ainsi, le Random Forest est un regroupement de plusieurs arbres de décision chacun travaillant de son côté de façon indépendante sur un sous ensemble des données d'entraînement.
Cela est un moyen permettant de limiter le volume de sur-apprentissage et d'avoir une sorte de moyenne des prédictions des arbres décisionnels de la forêt.

L'algorithme de Random Forest est basé sur la méthode du Bagging. Ainsi, pour une base contenant n observations il se présente comme suit :

- **1 -** Tirage de M échantillons BOOSTRAP (un boostrap pouvant être considéré comme un ré-échantillonnage avec remise de la taille de l'échantillon initial)
- **2 -** Construction de l'arbre sur chaque échantillon boostrap précédemment généré tout en veillant pour chaque segmentation à :

  a - Sélectionner aléatoirement m variables parmi l'ensemble des variables à disposition.

  b - Créer le meilleur noeud à partir de ces m variables.

Ainsi, M arbres sont créés à la fin et l'algorithme effectue des prévisions en faisant une prédiction pour chacun des arbres. La prédiction finale dans le cas d'un Random Forest de régression est obtenue par calcul de la moyenne de l'ensemble des estimations de ces M arbres.

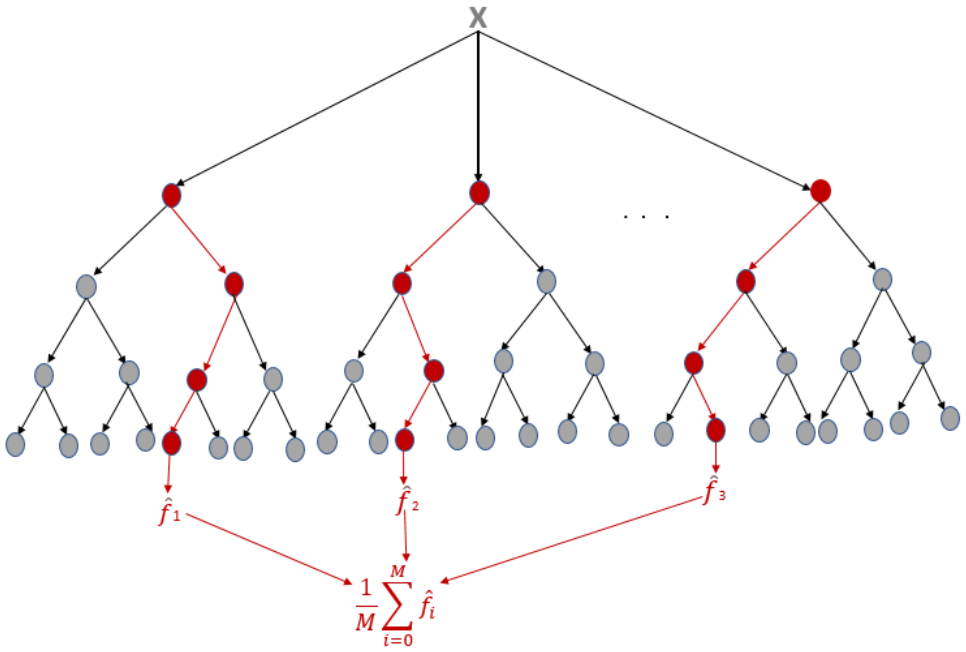


Figure 48 – Agrégation du Random Forest

9.4 Cadre théorique:

L'algorithme de Random Forest propose deux sorties très importantes à savoir l'erreur out of bag et la mesure de l'importance des différentes variables.

Erreur Out of bag :

La procédure Out of bag permet d'estimer les erreurs de prédiction de l'algorithme. En effet, dans chacune des échantillons boostrap, une fraction des données est laissée de côté lors du tirage. Ainsi, les individus de cette fraction restante (Out-of-bag) serviront à l'estimation des erreurs de prédiction.
Soit $(X_1, Y_1), ..., (X_n, Y_n)$ l'échantillon servant d'apprentissage à l'algorithme. Étant donné (X_i, Y_i) une observation de cet échantillon initial, nous définissons par I_β l'ensemble des arbres décisionnels de l'algorithme ne présentant pas cette observation dans leur échantillon boostrap servant d'entraînement. Ainsi, l'estimation de la prévision du random forest est obtenue par agrégation de ces arbres de I_β :

$$\hat{Y_i} = \frac{1}{|I_\beta|} \sum_{k \in I_\beta} T(X_i, k)$$

L'erreur Out-of-bag est ainsi définie comme suit :

. $OOB = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\hat{Y}_i \neq Y_i}$ dans le cadre de la classification
. $OOB = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)$ dans le cadre de la régression

Importance des variables :

L'erreur Out Of Bag permet également d'évaluer l'importance des variables du Random Forest.
On note E_{OOB_k} l'erreur de prédiction du prédicteur k évaluée sur cet échantillon. Elle est définie de la manière suivante:

$$E_{OOB_k} = \frac{1}{|OOB_k|} \sum_{i \in OOB_k} (h(x_i, k) - Y_i)^2$$

La mesure de l'importance d'une variable passe dans un premier temps par une perturbation des valeurs de la variable. Soit OOB_k^j l'échantillon OOB associé au k-ème arbre dans lequel nous avons une perturbation aléatoire des valeurs de la variable V_j.
L'importance de la variable est ainsi obtenue comme étant l'écart entre la moyenne des erreurs de l'arbre sur l'échantillon aléatoirement perturbé et celle obtenue sur l'échantillon initial. C'est ainsi qu'on a :

$$Imp_j = \frac{1}{M} \sum_{k=1}^{M} (E^j_{OOB_k} - E_{OOB_k})$$

Encodage des variables discrètes:

Les modèles de machine learning peuvent parfois être sensibles à la stratégie d'encodage utilisée. En effet, pour l'instant les algorithmes de machine learning sous python ne prennent en compte que des variables numériques. Raison pour laquelle les données catégorielles doivent être recodées sous forme de valeurs numériques pour être intégrés aux modèles de machine learning. Il est toutefois important de noter que le choix de la stratégie d'encodage des variables catégorielles est fortement dépendant de la base d'étude et de la qualité des variables explicatives disponibles. L'encodage peut être fait de plusieurs manières :

- Label Encoding :

Cette stratégie d'encodage consiste à affecter un nombre entier à chaque modalité de la variable discrète en question. Cette approche a l'inconvénient d'introduire un ordre dans les modalités, ce qui n'est pas toujours souhaitable.

- One Hot Encoding :

L'application de cette méthode permet de transformer les n modalités d'une variable discrète en n-1 variables binaires. Cette méthode aura tendance à favoriser les variables présentant beaucoup de modalités et ainsi augmenter la variance de l'estimation. En effet, les prédicteurs de la forêt aléatoire vont parfois donner un poids très important à cette variable par rapport aux autres variables lors du tirage aléatoire (https://turf-ia.com/encodage-a-chaud-one-hot-encoding/).

Dans le cadre pratique, ces méthodes ont été testées puis comparées. Il s'en est suivi que même avec un nombre d'arbres important, nous obtenons des résultats très similaires sur ces différentes stratégies.

Hyperparamétrage du modèle :

L'implémentation d'un modèle de Random Forest nécessite l'ajustement de quelques paramètres en vu de réduire au mieux les erreurs liées à la prédiction et de limiter le risque de sur-apprentissage.
Les paramètres importants à ajuster sont :

- **n-estimators (nombre d'arbres) :** il est souvent préférable que ce paramètre soit élevé. En effet, plus il y'a d'arbres, plus le modèle sera robuste pour limiter le risque de sur-apprentissage
- **max-features** : il est préférable que ce paramètre prenne une petite valeur. En effet, cela va permettre de réduire le sur-apprentissage lors de l'entraînement.
- Et possiblement des paramètres de pré-étalage comme par exemple **max-depth** qui permet d'ajuster ou de contraindre la profondeur des arbres.

9.5 Application à la mise en place d'un modèle pour l'embellissement :

Dans cette section, nous allons mettre en place un modèle de coût moyen via l'algorithme de Random Forest sous python qui pourra être challengé avec le modèle GLM précédemment étudié. Ainsi, les mêmes données d'entraînement et de test que le GLM seront utilisées. Pour rappel, 80 % des données avaient servi à l'apprentissage contre 20 % pour le test.

Calibrage des paramètres :

Dans le cadre de la mise en place du modèle de Random Forest, nous allons essayer de déterminer les paramètres optimaux en comparant deux approches :

— Détermination des paramètres par minimisation de l'erreur quadratique moyenne via k-fold cross validation

— Sélection des hyper-paramètres optimaux via la technique de Grid Search

Le GridSearch :

L'ajustement des hyperparamètres est une phase très importante dans le cadre de la mise en place de l'algorithme. Une des méthodes très répandue d'optimisation est le GridSearch. Cette technique permet d'estimer le meilleur paramétrage du modèle à travers la comparaison d'une série de paramètres. Les modèles sont par la suite évalués et challengés via une approche de validation croisée. Toutefois, la difficulté avec cette technique réside sur le fait que les paramètres à rensigner sont choisis en amont par l'exécuteur et le temps de calcul peut être considérable lorsqu'il y'a beaucoup de paramètres à tester.

Figure 49 – Illustration du Grid search

C'est ainsi que le calibrage sera axé sur les paramètres suivants :

— le nombre d'arbres

— le nombre de variables testées

— la profondeur maximale des arbres décisionnels

En effet, il est aussi possible de contraindre la profondeur des arbres notamment lorsqu'on a un grand échantillon. Dans le cadre de la détermination des paramètres notre base d'apprentissage a été partitionnée en 4. L'idée est de mettre en oeuvre l'approche de validation croisée en entraînant à chaque itération le modèle sur le 3 premières partitions et à l'évaluer sur la base de validation. Nous avons commencé par un ajustement simultané des deux paramètres les plus importants. A chaque itération, une analyse de l'erreur quadratique moyenne est effectuée afin d'avoir un premier à priori visuel :

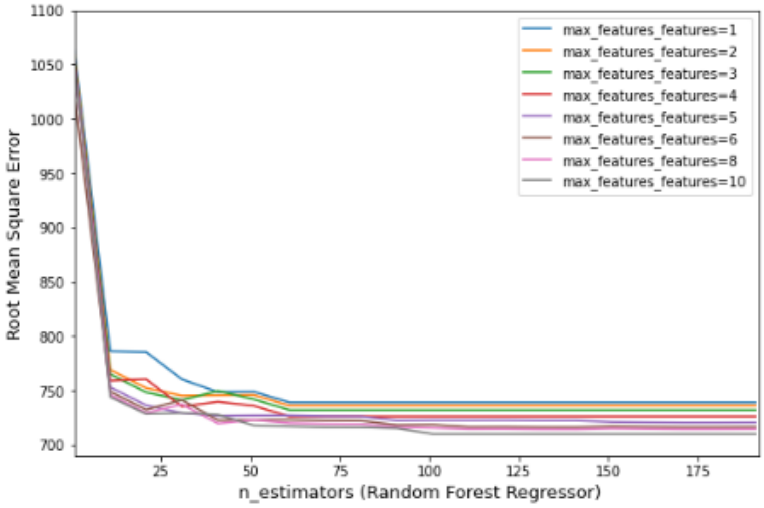


Figure 50 – RMSE en fonction du nombre d'arbres et du nombre de variables

Nous pouvons remarquer que l'erreur converge très rapidement avec l'augmentation du nombre d'arbres et se stabilise à partir d'une soixantaine d' arbres.

En fixant un nombre d'arbres à au moins 60, la performance du modèle est minimale lorsque toutes les variables sont utilisées. Par ailleurs, l'ajustement individuel du paramètre profondeur au delà de 6 voir 7 étages montrant qu'un arbre profond n'améliore pas forcément la qualité du modèle. C'est ainsi qu'après quelques simulations de ce processus sur les 4 partitionnements distincts de la base d'apprentissage, nous décidons de fixer le nombre d'arbres à 100 pour renforcer la stabilité du modèle. Puis maximiserons l'optimalité de la recherche des paramètres relatifs au nombre de variables et à la profondeur grâce à la technique du **GridSearch** autour de ces différentes valeurs obtenues. En principe, cette méthode permet de trouver le modèle avec les meilleurs hyperparamétres en comparant les performances des différentes combinaisons via la technique de k-fold cross-validation. Ainsi, après l'exécution de la méthode autour des valeurs définies, nous retenons au final un random forest issue de la moyenne des contributions de 100 arbres, 5 variables testées et de profondeur 5.

Paramètres	Valeurs
Nombre d'arbres	100
Nombre de variables testées	5
Profondeur	5

Importance des variables :

Le modèle de Random Forest est difficilement interprétable. Toutefois, avec les méthodes de Bagging il est possible de mesurer l'importance des variables.

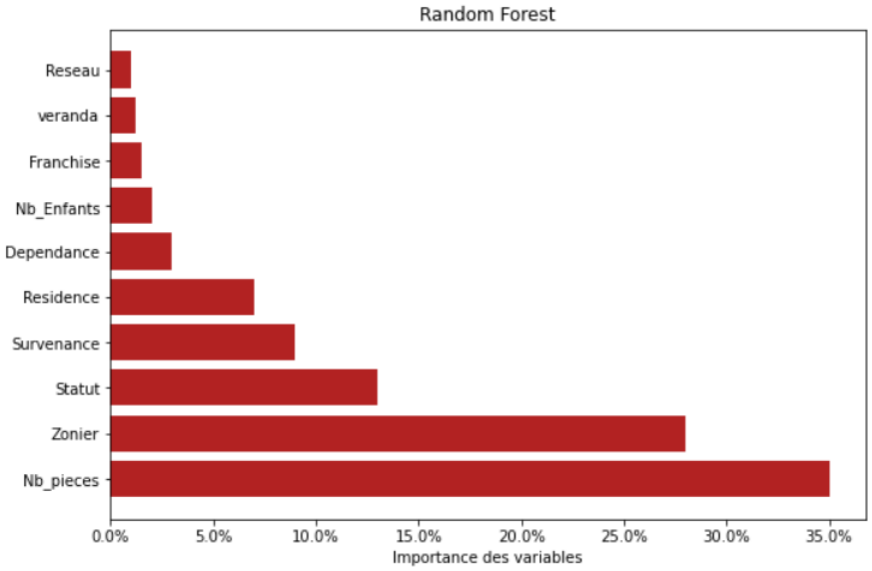


Figure 51 – Importance des variables

La variable qui contient le plus d'information dans l'explication du coût moyen de l'embellissement est le nombre de pièces disponible dans l'habitation suivi du zonier. Vient ensuite la variable relative au statut de l'assuré. La variable la moins contributrice au modèle est le réseau de distribution (Agents/Courtiers).

Performance du modèle :

Par la suite, en évaluant les erreurs de prédiction sur les bases d'apprentissage et de test, les résultats suivants sont obtenus :

Indicateur	Base d'apprentissage	Base de test
RMSE	714.58	725.58
MAE	513.32	526
Gini	25.32%	24.83%

Les métriques obtenus sur les base d'apprentissage et de test sont proche permettant de dire que le modèle ne souffre pas d'un phénomène de sur-apprentissage. Les résultats obtenus pour le random forest sont proches des résultats du GLM. Toutefois, ce dernier performe mieux même si le gain reste minime.

Courbe de Lorenz :

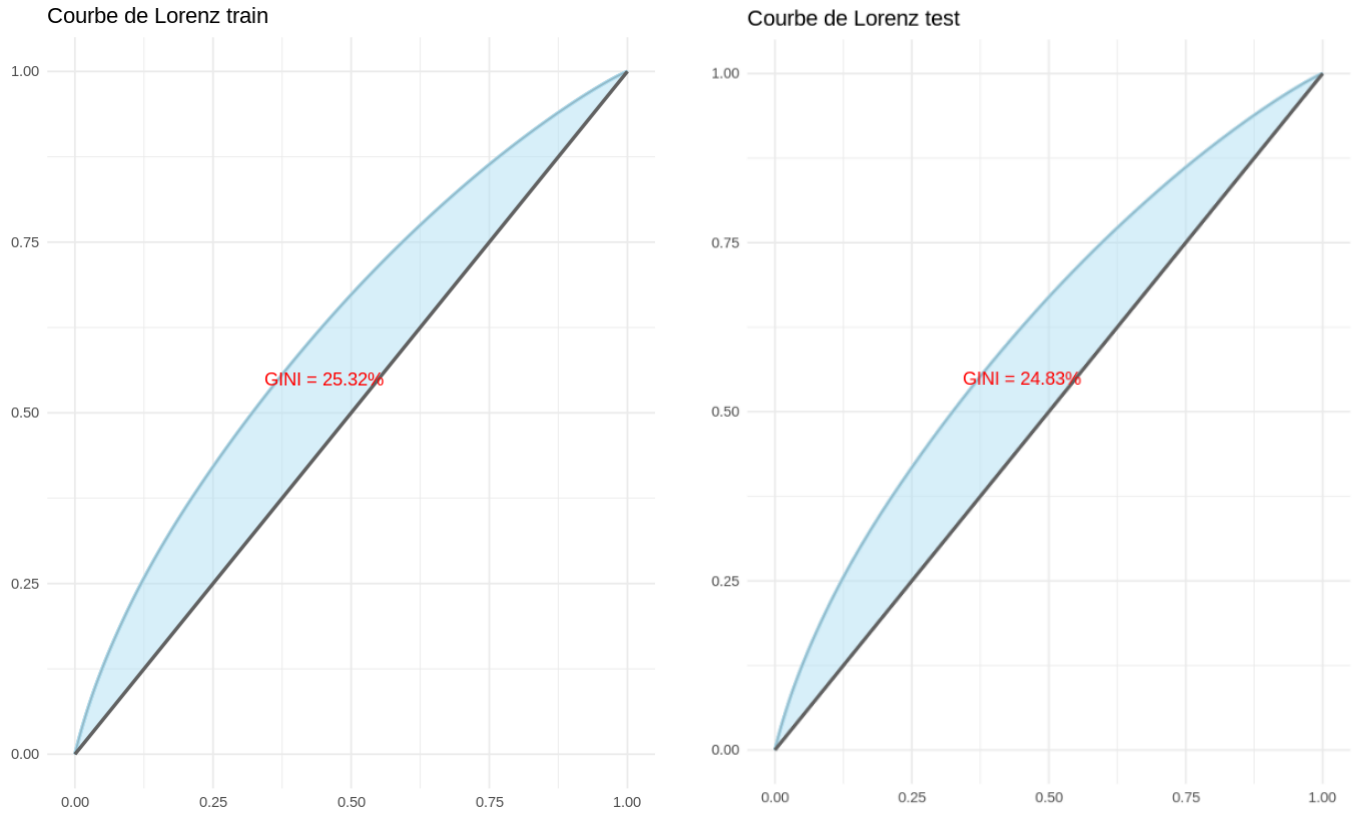


Figure 52 – Courbes de Lorenz sur les bases train et test de l'embellissement

9.6 Application à la mise en place d'un modèle pour l'immobilier :

Dans cette partie, nous allons à nouveau implémenter un modèle de coût moyen via l'algorithme de Random Forest dont les résultats seront comparés à ceux du GLM de l'immobilier. De la même manière que précédemment, les mêmes données d'entraînement et de validation que le GLM seront utilisées. Ainsi, 80 % des données serviront à l'entraînement et à la calibration du modèle contre 20 % pour le test.

Optimisation de l'algorithme :

Comme pour l'embellissement, nous allons dans un premier temps analyser le comportement du RMSE en fonction du nombre d'arbres et du nombre de variables à tester.
L'ajustement simultané des deux paramètres sur le jeu de test montre que l'erreur décroît de manière significative au tout début puis commence à se stabiliser par la suite. Nous remarquons qu'augmenter le nombre d'arbres contribue à renforcer le modèle et que le RMSE commence à se stabiliser au bout d'environ 100 arbres. Ainsi, en considérant un nombre d'arbres égal à au moins 100, l'erreur du modèle est minimale au bout et 3 variables testées et sa performance se dégrade petit à petit avec l'ajout de nouvelles variables. Cela pouvant être relativisé avec le fait qu'une valeur basse de ce paramètre fait que les arbres du Random Forest sont assez différents et que chaque arbre pourrait avoir besoin d'être profond (Andreas C.Muller et Sarah Guido) pour renforcer l'ajustement des données (d'où la nécessité de rajouter la profondeur des arbres comme paramètre supplémentaire à ajuster). Par ailleurs, le fait que le RMSE soit maximal lorsque toutes les variables disponibles sont testées peut être dû au fait qu'entraîner le modèle sur un nombre considérable de variables pourrait réduire les chances que les variables les plus prépondérantes soient utilisées dans chaque noeud. Le fait d'introduire des variables très peu contributrices peut augmenter l'erreur de prédiction. Cela s'explique par le fait qu'il y'a des chances que ces variables dites perturbatrices soient choisies à la place des variables les plus discriminantes.

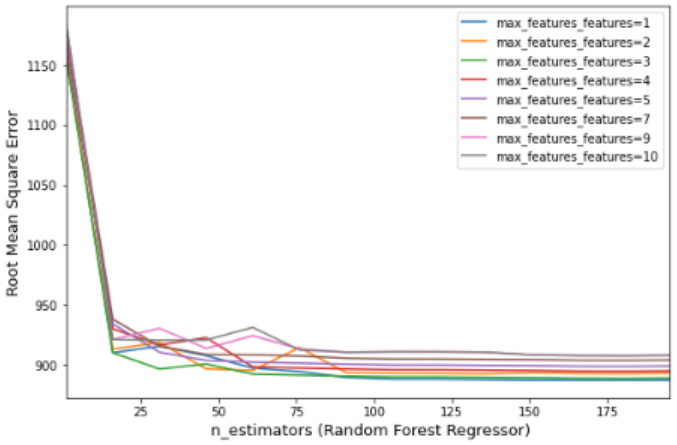


Figure 53 – RMSE en fonction du nombre d'arbres et du nombres variables testées

Dans la suite, après quelques itérations de ce processus, nous décidons comme précédemment de fixer le nombre d'arbres à 120 puis testons la combinaison des paramètres : profondeur et nombre de variables à tester grâce à la mise en oeuvre de l'approche GridSearch autour de ces premières valeurs. Pour le nombre de variables nous testons les valeurs $\{2, 3, 4, 5, 6, 8, 10\}$ et pour la profondeur les valeurs $\{1, 2, 3, ..., 10\}$. Après quelques simulations de cette technique, nous constatons que la combinaison de paramètres fournissant le meilleur score de validation croisée cette fois-ci est un Random Forest issu de la moyenne des contributions de 120 arbres, 4 variables testées et de profondeur 6. Ci-dessous un récapitulatif des résultats :

Paramètres	**Valeurs**
Nombre d'arbres	120
Profondeur	6
Nombre de variables testées	4

Importance des variables :

Le modèle de Random Forest est difficilement interprétable. Toutefois, avec les méthodes de Bagging il est possible de mesurer l'importance des variables.

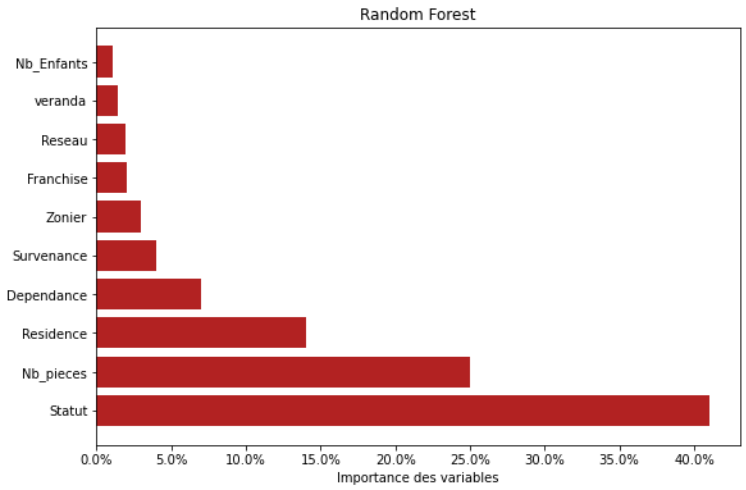


Figure 54 – Importance des variables

La variable qui capte le plus d'information concernant le coût moyen de l'immobilier est le statut de l'assuré qui concentre une part non négligeable de l'information. Vient par la suite la variable relative au nombre de pièces dans l'habitation et au type de résidence. Par ailleurs, les variables les moins importantes ou les moins contributrices au modèle sont le réseau de distribution (Agents/Courtiers), la présence de véranda, ou encore le nombre d'enfants dans le logement qui semblent avoir peu d'incidence dans l'explication du coût moyen de l'immobilier.

Performance du modèle :

Par la suite, en évaluant les erreurs de prédiction sur les bases d'apprentissage et de validation, les résultats suivants sont obtenus :

Indicateur	Base d'apprentissage	Base de test
RMSE	881	914
MAE	757	768
Gini	21.67%	19.29%

Nous obtenons des métriques du même ordre de grandeur que celles obtenues avec le GLM. Le Random Forest ne semble pas apporter une amélioration très significative et présente des performances assez similaires au GLM.

Courbe de Lorenz

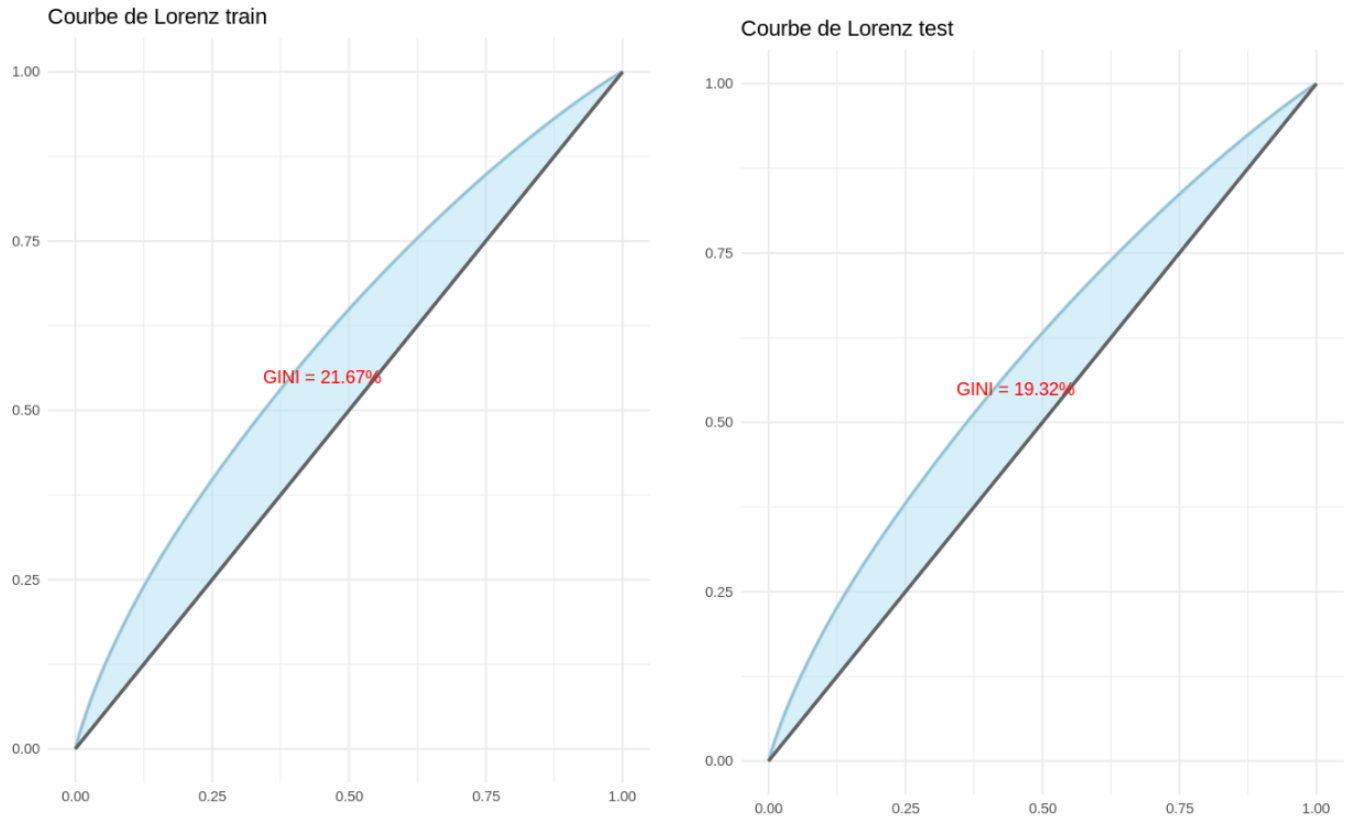


Figure 55 – Courbes de Lorenz sur les bases train et test de l'immobilier

10 Analyse des résultats et limites de l'étude:

10.1 Bilan modélisation de l'embellissement (comparaison des performances) :

10.1.1 Comparaison des performances et choix de modèle :

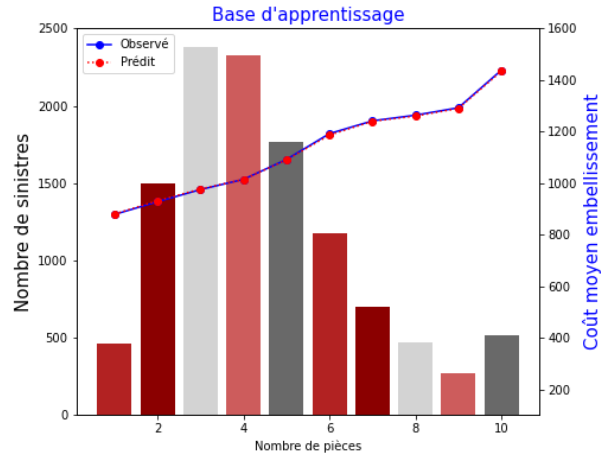
Dans le cadre de la modélisation du coût moyen de l'embellissement, les résultats obtenus sont assez satisfaisants. Toutefois, un modèle de Random Forest a été lancé pour le comparer au GLM en terme de performance. Ci-dessous, un tableau récapitulatif des performances :

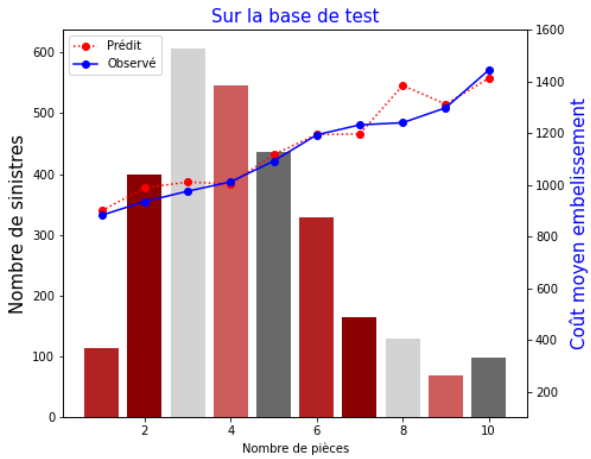
Méthode	RMSE test	MAE test	Gini test
GLM	728.25	518	25.02%
RF	725.58	526	24.83%

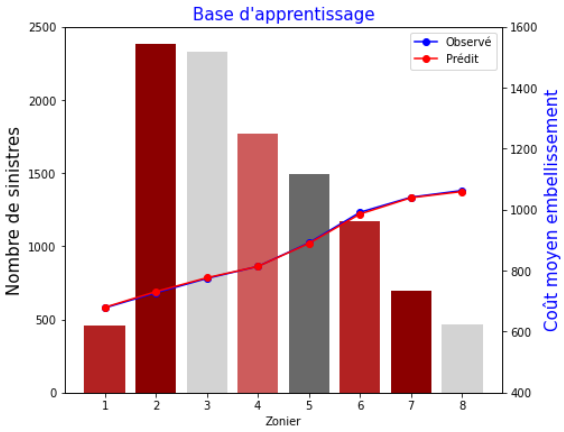
Les deux modèles présentent des performances comparables au regard des différentes métriques évaluées mais qui se révèlent être assez satisfaisants. Le MAE qui est un indicateur mesurant les écarts par observation est meilleur pour le GLM . De même, au sens de l'indice de Gini, la segmentation du coût est de meilleur qualité pour le GLM. Toutefois, au sens du RMSE, les résultats laissent à croire que ce modèle aura tendance à être légèrement moins bon pour les valeurs extrêmes. Par ailleurs, afin de se faire une idée sur les biais de la prédiction et la capacité du modèle à conserver tout son sens de généralisation sur de nouvelles données, une comparaison entre coût moyen prédit et observé par modalité de variable a été effectuée.

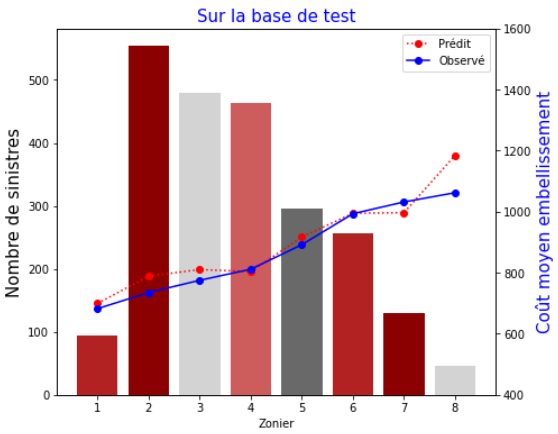
10.1.2 Analyse de la prédiction du meilleur modèle :

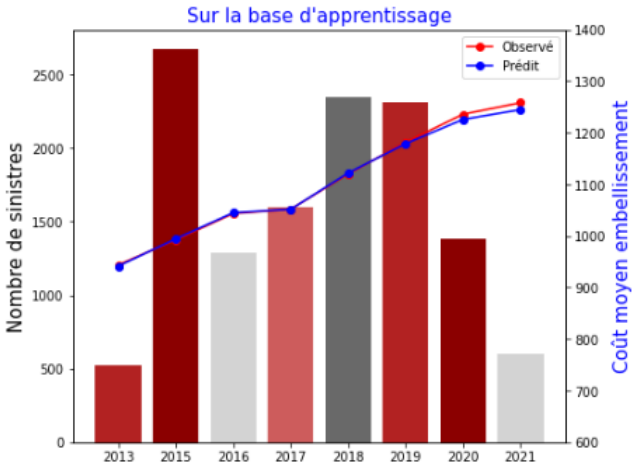
Afin d'évaluer la capacité du modèle à se généraliser sur de nouvelles données, nous allons essayer d'étudier pour chaque variable les coûts moyens observés et prédits sur la base d'apprentissage et celle de validation . Ci-dessous, nous présentons les résultats obtenus sur les variables :

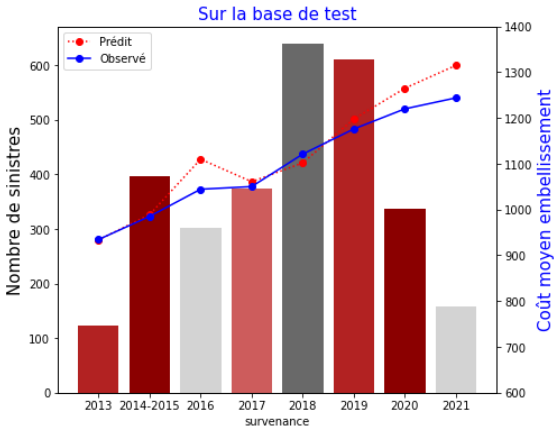












Ci-dessus les prédictions du modèle GLM qui est l'approche retenue et qui est destinée à expliquer le coût moyen de l'embellissement via la fonction de lien logarithme. Il apparaît que le modèle s'ajuste bien sur les données d'entraînement et n'a pas perdu tout sens de généralisation sur la base de validation. Nous pouvons remarquer que sur la base de test ce dernier s'ajuste moins bien que sur la base d'entraînement car on note quelques sur-estimations et sous-estimations du modèle sur certaines modalités de variables.
Le GLM présente des résultats globaux satisfaisants. Les résultats de l'apprentissage et de test semblent être assez concluants via une prédiction cohérente avec les observations.

10.2 Bilan modélisation de l'immobilier :

10.2.1 Comparaison des performances et choix de modèle :

Nous allons comparer les deux modèles à savoir le GLM et le Random Forest afin de se faire une idée sur les performances de chacun. Ci-dessous les métriques obtenues pour chaque modèle :

Méthode	RMSE test	MAE test	Gini test
GLM	928	787	19.29%
RF	914	768	19.38%

Les deux modèles cherchent à prédire le coût moyen du dommage immobilier. Les métriques d'erreurs évaluées pour les deux modèles et les résultats montrent que le Random Forest performe mieux que le GLM pour la modélisation de l'immobilier. Toutefois, on reste sur des performances comparables. Nous ne notons pas d'écarts significatifs au niveau des métriques d'erreurs évaluées sur l'apprentissage et le test. Toutefois, l'analyse des graphiques de comparaison entre coûts moyens prédits et observés révèle que les modèles présentent un biais de sur-évaluation des modalités ne présentant pas suffisamment d'observations. Un phénomène qui est beaucoup plus visible au niveau du GLM. En effet, la forêt aléatoire réduit cet effet de sur-estimation locale en atténuant ce phénomène au niveau de ces modalités. Raison pour laquelle ce dernier sera choisi pour la modélisation du poste immobilier. Dans la section suivante, nous présentons la comparaison entre coûts moyens prédits et observés du random forest (modèle retenu) pour avoir plus de visibilité sur les biais en lien avec la qualité de prédiction.

10.2.2 Analyse de la prédiction du meilleur modèle :

Dans cette section nous allons comparer comme décrit précédemment le prédit et l'observé afin de bien analyser les biais que pourrait potentiellement présenter le modèle.

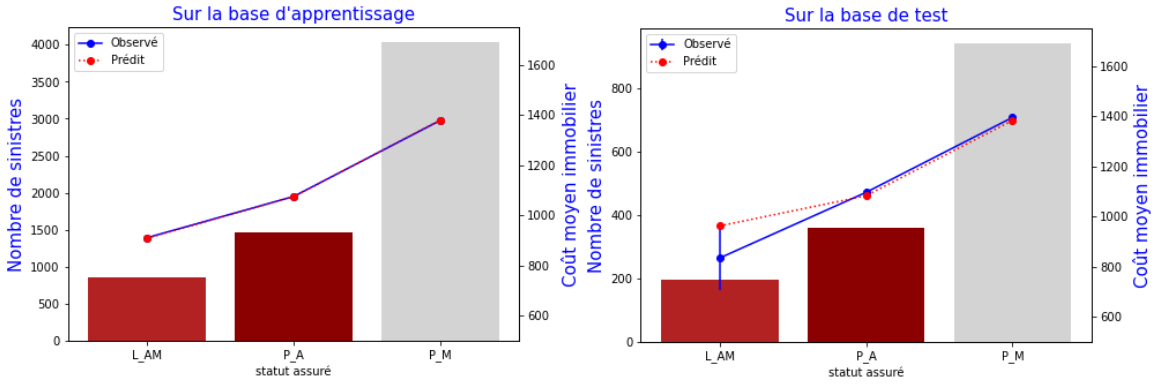


Figure 56 – Comparaison entre valeurs observées et prédites du Random Forest sur le test et le train de la variable statut de l'assuré

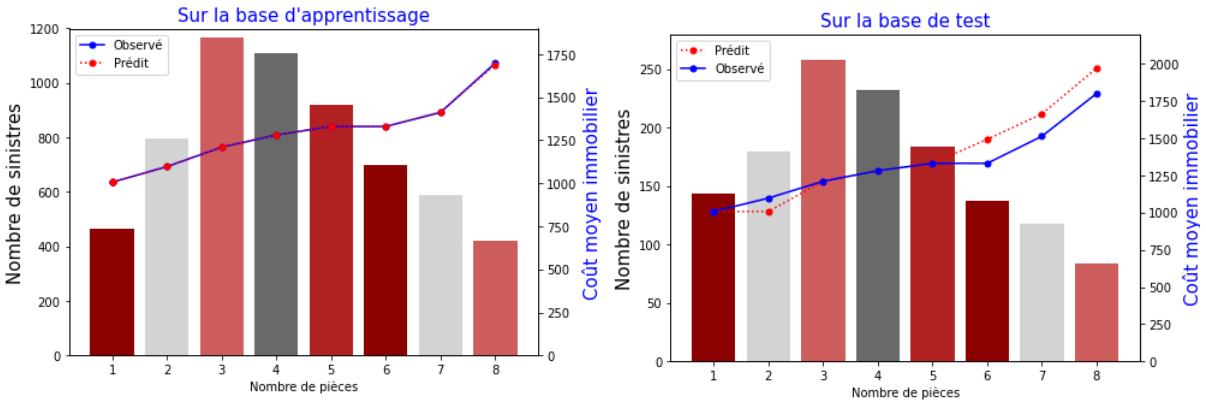


Figure 57 – Comparaison entre valeurs observées et prédites de sur le test et le train de la variable nombre de pièces

L'analyse de ces graphes montre que l'apprentissage du modèle se passe plutôt bien sur la base d'entraînement. Toutefois, ce dernier s'ajuste moins bien à de nouvelles données sur la base de test en présentant des écarts entre valeurs prédites et observées au niveau de certaines modalités. En effet, un phénomène de sur-apprentissage local peut s'observer au niveau de certaines modalités. De la même façon que le GLM, le Random Forest peut parfois avoir tendance à sur-évaluer le coût moyen sur des modalités avec peu d'observations. Cela peut s'expliquer par le fait que le modèle ne dispose pas d'assez d'informations pour capturer tous les effets risque de ces modalités. Toutefois les prédictions se retrouvent sur les modalités présentant suffisamment d'observations.

10.3 Résultats de la modélisation directe :

Pour des fins de comparaison, nous avons réalisé une approche de modélisation directe du montant des indemnisations allouées au titre de l'embellissement et de l'immobilier. En effet, nous avons modélisé cette fois-ci directement le coût total relatif à ces deux dommages par application des deux méthodes précédentes (GLM et Random Forest). La procédure de construction est la même que celle des deux postes de dommage modélisés individuellement raison pour laquelle nous ne nous attarderons pas sur les détails des différentes étapes. Toutefois, nous avons dans un premier temps implémenté un GLM de type Gamma en sélectionnant les variables via la méthode stepwise suivant les critères AIC et BIC. Mais aussi, en analysant la significativité des variables ainsi que leur contribution à la baisse de la déviance. Le but étant encore de trouver le modèle le plus parcimonieux via une recherche d'équilibre entre qualité et simplicité. Par la suite, nous avons également implémenté un Random Forest comme précédemment en ajustant principalement le nombre d'arbres, le nombre de variables et la profondeur. Ci-dessous les résultats obtenus au niveau des indicateurs sur le jeu de test :

Méthode	RMSE test	MAE test	Gini test
RF	1166	978	18.35%
GLM	1183	989	17.02%

Les valeurs des différents indicateurs de performance sont proches. Toutefois, nous remarquons que pour une approche de modélisation directe le Random Forest performe mieux que le GLM sur tous les indicateurs et sera retenu pour la suite. En effet, ce dernier présente des niveaux d'erreur plus faibles et un Gini un plus élevé.

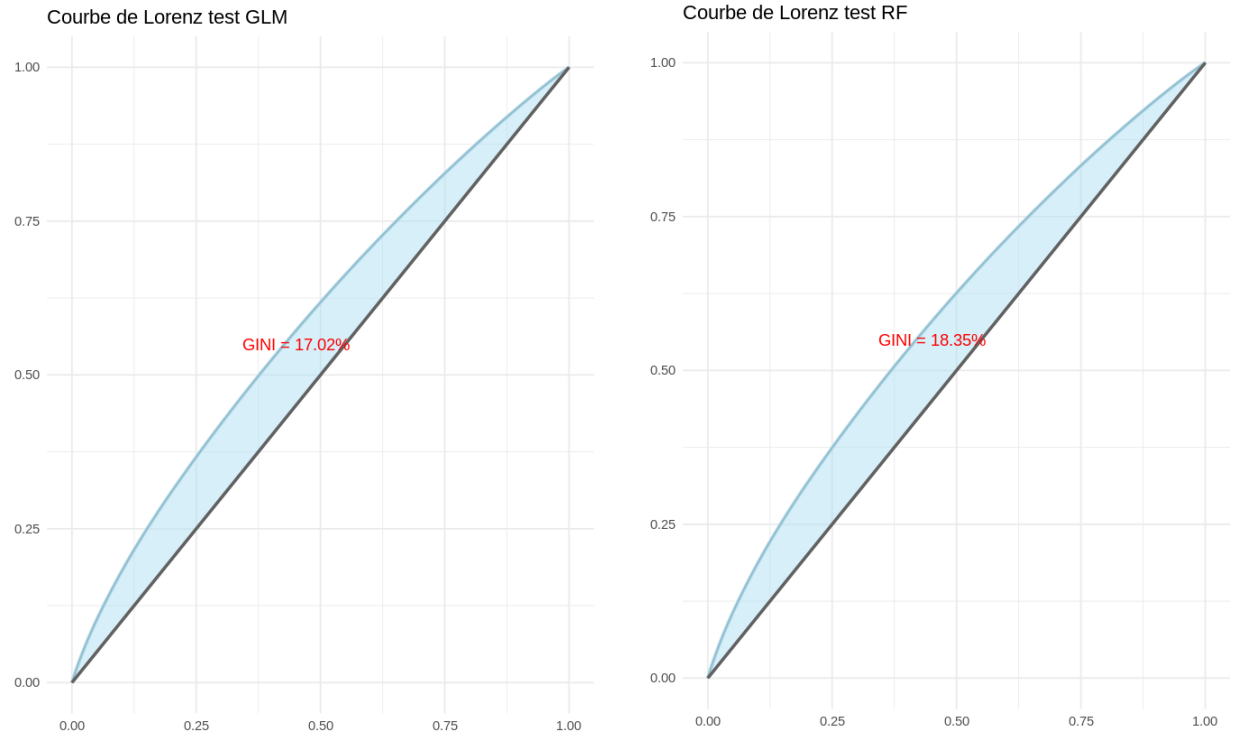


Figure 58 – Courbes de Lorenz du GLM sur les bases train et test de l'embellissement

Ainsi, modèle de Random Forest ainsi retenu est issu de la moyenne des contributions de 160 arbres, 5 variables testées et de profondeur 6. L'analyse de l'importance des différentes variables donne les résultats suivants :

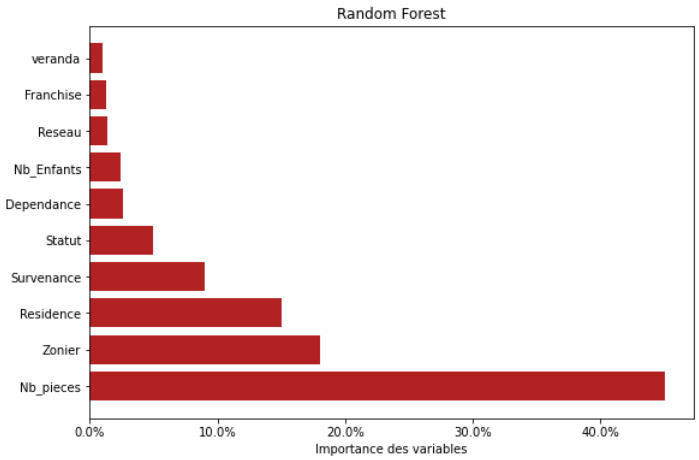


Figure 59 – Importance des variables

La variable qui explique le plus le coût des dommages est le nombre de pièces dans l'habitation suivi du zonier. Les variables les moins contributrices au modèle demeurent le réseau, la franchise ou encore la véranda.

10.4 Modélisation directe vs modélisation par poste

Dans la suite, nous avons comparé les performances de la modélisation directe (tout dommage confondu d'entrée de jeu) à celles obtenues via l'approche de modélisation fine (séparation des postes puis agrégation). Cela étant possible en reconstituant le coût total par somme des coûts relatifs à chaque dommage indemnisé (**Cf** . [ici]). Cela permet d'obtenir une performance globale résultant de l'agrégation des deux modèles. Nous évaluons par la suite les différents indicateurs à savoir l'indice de Gini et le RMSE dans l'échantillon test pour chacune des deux approches. Ci-dessous les résultats obtenus :

Méthode	**Modélisation directe**	**Modélisation par poste**
Gini	18.35%	22.3%
RMSE	1166	873

D'après ces résultats, nous pouvons déjà voir clairement que l'approche de modélisation par poste est plus discriminante au sens du Gini avec presque 4 points en plus. En d'autres termes, cette approche de modélisation par type de dommage segmente mieux le risque qu'une approche de modélisation directe pour le risque étudié. Aussi, au sens du RMSE, l'approche de modélisation fine présente des erreurs de prédiction beaucoup plus faibles. Cela permet de mettre en évidence l'avantage de cette approche modélisation fine dans le cadre de l'évaluation de nos sinistres en dommage aux biens.

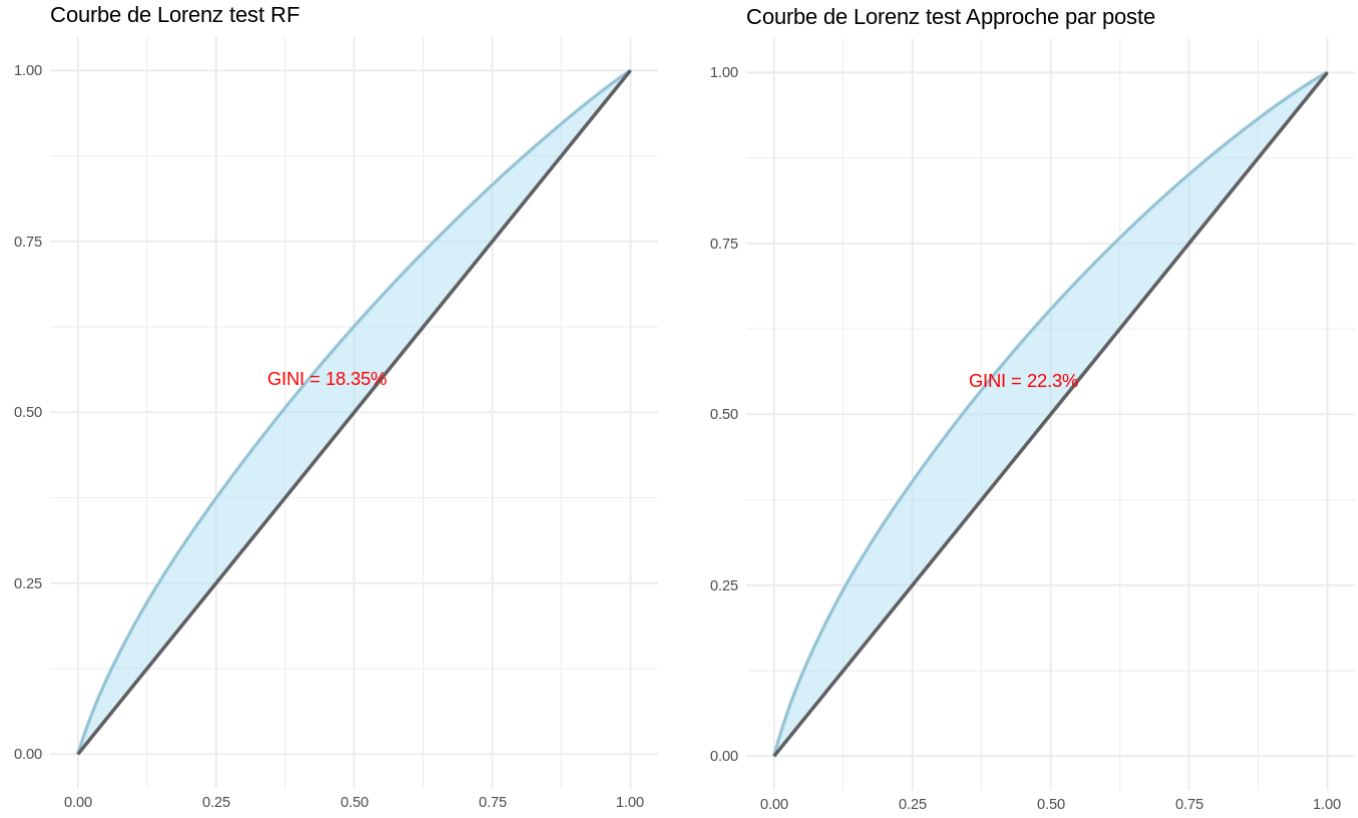


Figure 60 – Courbes de Lorentz par approche de modélisation : modélisation directe (à gauche) et modélisation par poste (à droite)

10.5 Backtesting :

Afin de valider la fiabilité du modèle, nous aurions bien voulu réaliser un backtesting de ce dernier. Mais pour cela il aurait fallu récupérer les rapports de 2022 qui ne sont pas disponibles à date, lancer l'outil pour extraire les nouvelles données et mettre en place une nouvelle base . Cela aurait permis d'appliquer le modèle à un échantillon de sinistres déjà clos et évaluer les écarts de modélisation dans le temps.

10.6 Limites de l'étude et axes d'amélioration:

Volumétrie des données:

Une des difficultés concernant l'application d'une approche de modélisation fine réside au niveau de la faible volumétrie des données que peuvent présenter certains postes dans le cadre des estimations. Nos travaux sont principalement concentrés sur la modélisation des dommages relatifs à l'embellissement et à l'immobilier nous permettant d'expliquer environ 94% des coûts indemnisés au titre des sinistres expertisés pour cette garantie. En effet, notre échantillon ne présente pas à ce jour suffisamment de données nous permettant d'implémenter des modèles pour expliquer le coût moyen du mobilier ou encore des autres indemnisations portant des libellés spécifiques. Cela est à relativiser avec le fait que ces deux derniers postes sont très peu sinistrés dans la base d'étude pour cette garantie. De ce fait, leur modélisation individuelle a été testée sur une volumétrie très restreinte présentant une variance assez élevée. D'ailleurs, un des comportements nous paraissant pathologique de la modélisation de ces postes ("Mobilier"et "Autres") est que les modèles sur-apprennent et ne permettent donc pas d'aboutir à des résultats concluants. Une autre limite à nos travaux est que l'étude a été restreinte aux rapports d'expertise du cabinet saretec faisant que nous ne disposons que d'une partie des rapports relatifs à notre périmètre d'étude. La prochaine étape de l'étude (actuellement en cours) sera d'appliquer la même démarche avec les rapports d'experts du cabinet Texa. Cela permettra d'obtenir un gain en terme de volumétrie et de consolider cette étude sur un volume de rapports plus conséquent.

Qualité des données :

La base risque présente certaines variables avec un faible taux de remplissage. Ces champs ne sont pas exploitables du fait de ce nombre important de valeurs manquantes dont le retraitement est susceptible d'introduire un biais aux nos modèles. Pourtant ce sont des variables qui paraissent pertinentes pour expliquer le coût suite à un dégât des eaux à savoir "l'étage", "la présence de cheminée" ou encore "l'année de construction". Mais aussi, auraient pu permettre de mieux approfondir les études. Ainsi, des améliorations sont à réaliser concernant le remplissage de ces champs. Cela pourra certainement permettre d'envisager l'usage de ces variables dans les modèles.

Conclusion :

Les sinistres expertisés représentent près des deux tiers de la charge en dégât des eaux ces dernières années. De ce fait, il est important pour un assureur de consacrer une étude spécifique à ces types de sinistre dans le cadre de la compréhension et du pilotage de l'exercice. Comme mentionné en introduction, l'enjeu de ce mémoire était double. D'une part, Generali souhaite alimenter ses bases pour enrichir ses données en DAB et lancer différentes études. Ceci, à l'aide de données non structurées issues des rapports d'experts qui s'avèrent être une source précieuse d'informations. D'autre part, l'objectif était par la suite de proposer une approche de modélisation par poste dans le cadre de l'évaluation du coût moyen des sinistres expertisés en dégât des eaux du marché de l'habitation. C'est dans ce sens que différentes étapes ont été abordées : extraction puis mise en cohérence des données, retraitement des données et modélisation du coût moyen.

Dans un premier temps, nous nous sommes focalisés sur la mise en place de la base des données d'expertise. La première difficulté de ces travaux était de réussir à généraliser la lecture des données sur l'ensemble des rapports. Cela nous a amenés à lire une centaine de rapports puis à automatiser un algorithme d'extraction des données sur un volume conséquent de rapports. Une fois cette phase finalisée, l'étape suivante consistait à retraiter la base. En effet, il a fallu d'abord identifier et retraiter les différents clés de jointure dans le cadre du rapprochement des bases. Mais aussi, nous étions amenés à traiter les cas de figure où un sinistre était rattaché à plusieurs rapports, notamment via un processus d'identification et de sélection de la dernière vision des rapports dans le cas où cette dernière retrace l'historique des indemnisations. Dans un scénario contraire, où la dernière vision n'annule et remplace pas les précédents, il a fallu analyser puis reconstituer le coût total des prestations. Par la suite, l'étape d'après consistait à fiabiliser les données. L'idée était d'analyser les écarts et de vérifier si les données d'expertise reflètent toujours la réalité des indemnisations de Generali. La base mise à disposition permet aujourd'hui à Generali France d'exploiter ces nouvelles données dans le cadre de différentes études. Mais aussi, cela a surtout permis de disposer d' une base structurée présentant une vision détaillée des prestations allouées par Generali et d'atteindre ainsi une granularité très fine dans le cadre de l'analyse détaillée du coût des sinistres.

Dans un second temps, au-delà de ces retraitements préliminaires de la base d'étude, la qualité de la modélisation est très liée au nettoyage des données notamment le traitement des données manquantes des données aberrantes ainsi que la discrétisation des variables. Ainsi, certaines variables explicatives ont été complétées en partie par des données rapports à l'issue d'une extraction dans la mesure du possible, avant d'appliquer la méthode des K-plus proches voisins pour assurer la complétion totale. C'est le cas par exemple de la variable nombre de pièces qui s'est avérée très discriminante pour les modèles étudiés.

Nous avons également réalisé quelques statistiques descriptives des différentes variables disponibles afin d'avoir un à priori visuel de la tendance du coût moyen. Cette étape a permis d'identifier les potentielles incohérences au niveau de leur distribution pour procéder à des ajustements. C'est dans ce sens que, dans le cadre de la classification, certaines variables présentaient beaucoup trop de modalités ou certaines catégories avaient un effectif trop faible. Des méthodes de classification telles que la méthode des KMeans ou de regroupements par avis d'expert ont été challengées.

Une fois ces étapes de création, de mise en cohérence et de fiabilisation de la base finalisées, nous avons abordé dans un dernier temps la phase de modélisation. C'est dans ce sens que, nous avons en premier lieu implémenté un GLM qui est l'approche actuarielle "traditionnelle" pour chacun des dommages modélisés que nous avons par la suite comparé avec le Random Forest. Il s'en est suivi que le GLM présente de meilleures performances pour la modélisation de l'embellissement. Toutefois, le Random Forest performe mieux dans le cadre de la modélisation du poste immobilier. Concernant la modélisation de l'immobilier, un léger phénomène de sur-estimation locale est identifié notamment au niveau de quelques modalités où ils n'ont pas suffisamment appris. Toutefois, nous retrouvons des prédictions satisfaisantes sur les modalités avec assez d'observations. Plus de données pourraient certainement permettre de mieux capturer les effets risques pour ces classes à faible volumétrie et d'affiner le modèle. Mais aussi, l'application d'un modèle tel que le gradient boosting n'a pas été testée, mais sa mise en oeuvre pourrait probablement fournir de meilleurs résultats.

Finalement, nous avons comparé les résultats d'une approche de modélisation directe (tout dommage confondu) par rapport à une approche de modélisation fine (séparation par poste puis agrégation). Il s'en est suivi que cette dernière fournit de meilleurs résultats au regard des différents indicateurs. D'abord, au sens du Gini, la segmentation du risque est de meilleure qualité. De même, côté RMSE, les erreurs de prédiction sont moins élevées. Ces premiers résultats illustrent l'avantage que pourrait offrir cette approche de modélisation fine dans le cadre de l'évaluation du coût des sinistres.

Ce qui fera l'objet d'un développement ultérieur et qui sera bientôt en cours, est l'application de la même démarche avec les autres cabinets afin de consolider les informations d'expertise et d'augmenter la volumétrie des données disponibles. Cela permettra de renforcer la robustesse des modèles et de disposer d'assez de données pour compléter la modélisation de l'ensemble des postes.

Références

[1] Arbres CART et Forêts aléatoires, Importance et sélection de variables Robin Genuer, Jean-Michel Poggi. https://hal.archives-ouvertes.fr/hal-01387654v2/document

[2] Actuariat IARD - ACT2040 Partie 4 - modèles linéaires généralisés. https://freakonometrics.hypotheses.org/files/2013/10/slides-2040-4.pdf

[3] S.Gazzola et A. Rorato, Réglementation des assurances, Cours dispensé à l'ISUP, 2021.

[4] Andreas C.Muller et Sarah Guido. Le machine learning avec python

[5] [Breiman, 2001] Breiman L. (2001). Random forests

[6] CHARPENTIER A., DUTANG C., (2012), L'actuariat avec R. https://cran.r-project.org/doc/contrib/Charpentier_Dutang_actuariat_avec_R.pdf

[7] MAUD Thomas. (2021) Econométrie de l'assurance non-vie. Cours ISUP Master 2 Actuariat

[8] MAUD Thomas. (2021) Econométrie de l'assurance non-vie. Cours ISUP Master 2 Actuariat

[9] Tarification IARD : Introduction aux techniques avancées, F. Planchet, A. Miseray.

[10] Wikistat.sélection de modèles avec R. https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-scenar-reg-penal-prostate.pdf

[11] The university of sydney.Variable Selection Stepwise, AIC and BIC.https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Lecture/lecture11_2020JC.html#19

[12] Mémoire d'actuariat, Création de zoniers en assurance habitation à l'aide de variables externes et de méthdhodes de Data Science, Antoine PESNAUD, 2020.

[13] FFA, L'assurance habitation en 2020. https://www.franceassureurs.fr/nos-chiffres-cles/assurance-de-dommages-et-responsabilite/lassurance-habitation-en-2020/ [13]

ANNEXES

1 - Test d'indépendance du χ^2 :

Le test d'indépendance du χ^2 permet à partir d'un échantillon l'hypothèse de dépendance ou d'indépendance entre 2 variables qualitatives. En considérant X et Y 2 variables qualitatives présentant chacune les modalités respectives k et r. Le tableau de contingence se déclinera comme suit :

	Y_1	Y_2	...	Y_r	Total
X_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,r}$	$n_{1,.}$
X_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,r}$	$n_{2,.}$
...	...	...	...	...	...
X_k	$n_{k,1}$	$n_{k,2}$	...	$n_{k,r}$	$n_{k,.}$
Total	$n_{.,1}$	$n_{.,2}$	...	$n_{.,r}$	n

De ce fait le χ^2 se définira de la manière suivante :

$$\chi^2 = \sum_{i,j} \frac{(n_{i,j} - \frac{n_{i,.}n_{.,j}}{n})^2}{\frac{n_{i,.}n_{...}}{n}}$$

2 - Présentation des approches de modélisation :

Dans ce mémoire, deux approches de modélisation ont été testées et comparées :

— **Approche de modélisation directe** : Il s'agit de l'approche de modélisation traditionnelle consistant à modéliser directement le coût total relatif aux différent postes qui seront agrégés d'entrée de jeu pour servir de données d'entrée aux modèles.

— **Approche de modélisation fine** : Cette approche consiste à modéliser séparément les différents postes de sorte à disposer de plusieurs modèles distincts. Ces derniers vont par la suite faire l'objet d'une agrégation en vu de restaurer la performance globale du modèle.

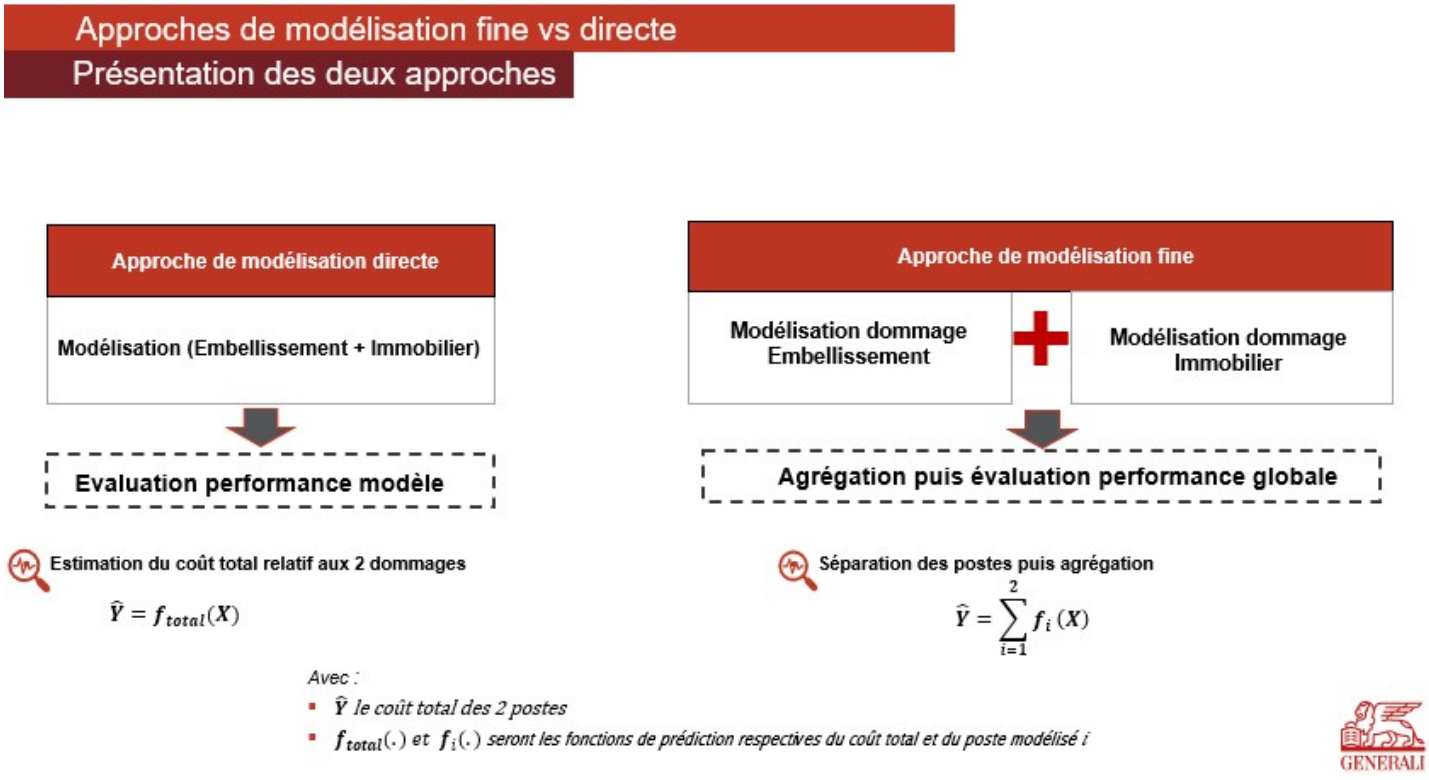
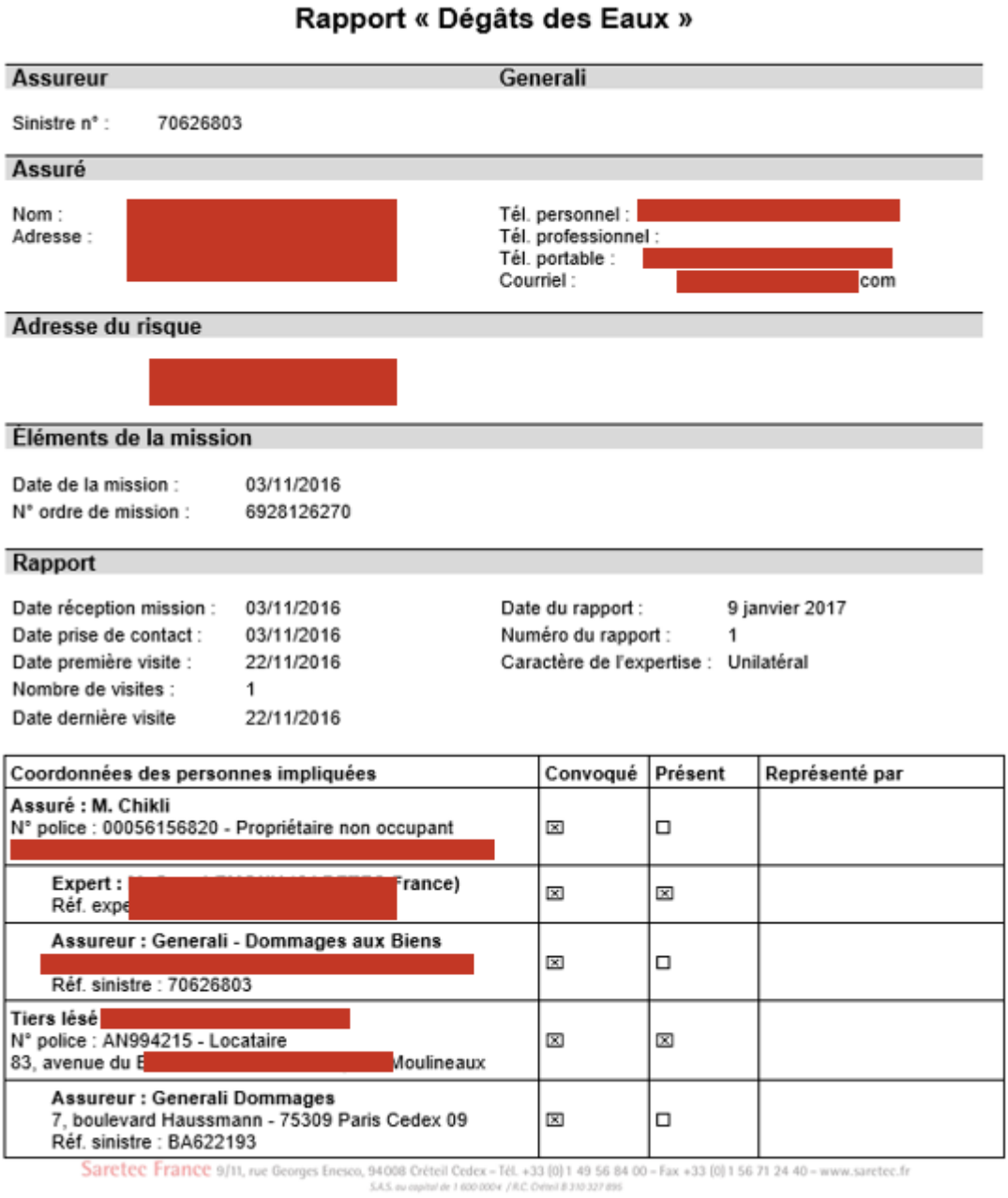


Figure 61 – Approche comparative : Modélisation fine vs Modélisation directe [ici]

3 - Exemple de rapport d'expertise :

Ci-dessous un exemple de rapport d'expertise. Dans un soucis de confidentialité, certains éléments ont été masqués. [ici]



Rapport « Dégâts des Eaux »

Assureur Generali

Sinistre n° : 70626803

Assuré

Nom : [redacted]
Adresse :

Tél. personnel : [redacted]
Tél. professionnel :
Tél. portable : [redacted]
Courriel : [redacted]com

Adresse du risque

[redacted]

Éléments de la mission

Date de la mission : 03/11/2016
N° ordre de mission : 6928126270

Rapport

Date réception mission : 03/11/2016
Date prise de contact : 03/11/2016
Date première visite : 22/11/2016
Nombre de visites : 1
Date dernière visite 22/11/2016

Date du rapport : 9 janvier 2017
Numéro du rapport : 1
Caractère de l'expertise : Unilatéral

Coordonnées des personnes impliquées	Convoqué	Présent	Représenté par
Assuré : M. Chikli N° police : 00056156820 - Propriétaire non occupant [redacted]	☒	☐	
Expert : [redacted]rance) Réf. expe[redacted]	☒	☒	
Assureur : Generali - Dommages aux Biens [redacted] Réf. sinistre : 70626803	☒	☐	
Tiers lésé [redacted] N° police : AN994215 - Locataire 83, avenue du E[redacted]Moulineaux	☒	☒	
Assureur : Generali Dommages 7, boulevard Haussmann - 75309 Paris Cedex 09 Réf. sinistre : BA622193	☒	☐	

Saretec France 9/11, rue Georges Enesco, 94008 Créteil Cedex – Tél. +33 (0)1 49 56 84 00 – Fax +33 (0)1 56 71 24 40 – www.saretec.fr
S.A.S. au capital de 1 600 000 € / R.C. Créteil B 310 327 895

Figure 62 – Page 1/3 rapport d'expertise

Contrat

Type de police :	Multirisques Habitation	Numéro de contrat :	00056156820
Date effet :	22/08/2011	Dénomination de la police	MRH MULT DOMICILE PNO D4
Franchise :	Non communiquée €	Indice de base : 875,2	Indice applicable : 932,7

Vérification du risque

Qualité de l'assuré :	Propriétaire non occupant		
Type de risque :	Maison individuelle		
Nombre de pièces principales :	7	Risque vérifié :	Oui
Surface :	150 m²	Conformité du risque :	Oui

Sinistre

Date réelle du sinistre :	18/04/2016		
Suppression de la cause :	Oui	Réparation faite :	Oui
Convention applicable :	Aucune	Réserve sur la garantie :	Non
Nature garantie :	Dégâts des eaux	Mesure de prévention :	Non

Observations et réserves sur le contrat ou sur le risque :

Le risque assuré est un pavillon de type R+2 et R-1 d'une surface habitable de 150 m², édifié (construction traditionnelle en parpaing + couverture en tuiles mécaniques) en périphérie du centre-ville d'ISSY LES MOULINEAUX.

Ce pavillon comporte 7 pièces principales dont 1 double séjour, 2 chambres à l'étage + salle de bains, ainsi que 2 chambres au second étage + salle de bains. Un niveau de sous-sol non aménagé de a été constaté.

Les menuiseries extérieures en bois sont toutes munies de volets bois; la porte d'entrée est équipée d'une serrure à 3 points de fermeture.

GENERALI

Figure 63 – Page 2/3 rapport d'expertise

Causes et circonstances du sinistre

Point de départ du sinistre

Pavillon de Monsieu

Causes et circonstances

Le présent sinistre résulte d'une fuite sur canalisation d'alimentation privative non accessible sous la baignoire du pavillon de ocataire de Monsieur infiltrations ont occasionné des dommages aux embellissements d'origine dudit logement.

Lors de notre passage les supports endommagés étaient en voie d'assèchement.

Localisation des dommages :

Escalier et chambre enfant à l'étage

Description des dommages :

Embellissements d'origine : dégradation de la peinture sur les murs et au plafond

Estimation des dommages et calcul indemnitaire

Bénéficiaire : Monsieur

Taux de TVA : Voir en annexe le tableau Evaluation des dommages

Nature estimation : Estimation en indemnisation pécuniaire

Tableau de règlement

Désignation	Indemnité immédiate	Indemnité différée	Indemnité totale
Embellissements en indemnisation pécuniaire	2 967,69 €	523,71 €	3 491,40 €
Total avant application de la franchise	2 967,69 €	523,71 €	3 491,40 €
Montant des indemnités nettes	**2 967,69 €**	**523,71 €**	**3 491,40 €**

Accord assuré : Nous avons obtenu un accord verbal de l'assuré sur le montant des dommages.

Observations : Sans objet

Le montant des dommages est détaillé dans le tableau de chiffrage en fin de rapport (annexe 1).

Conclusion

Les dommages aux embellissements d'origines sont à prendre en charge par vos soins dans cette affaire.

Figure 64 – Page 3/3 rapport d'expertise

4 - Liste des autres variables de la base construite :

Ci-dessous la liste des autres variables de la base de données. Ces informations servirons à enrichir la base de données et seront utiles dans le cadre de différentes études actuarielles.

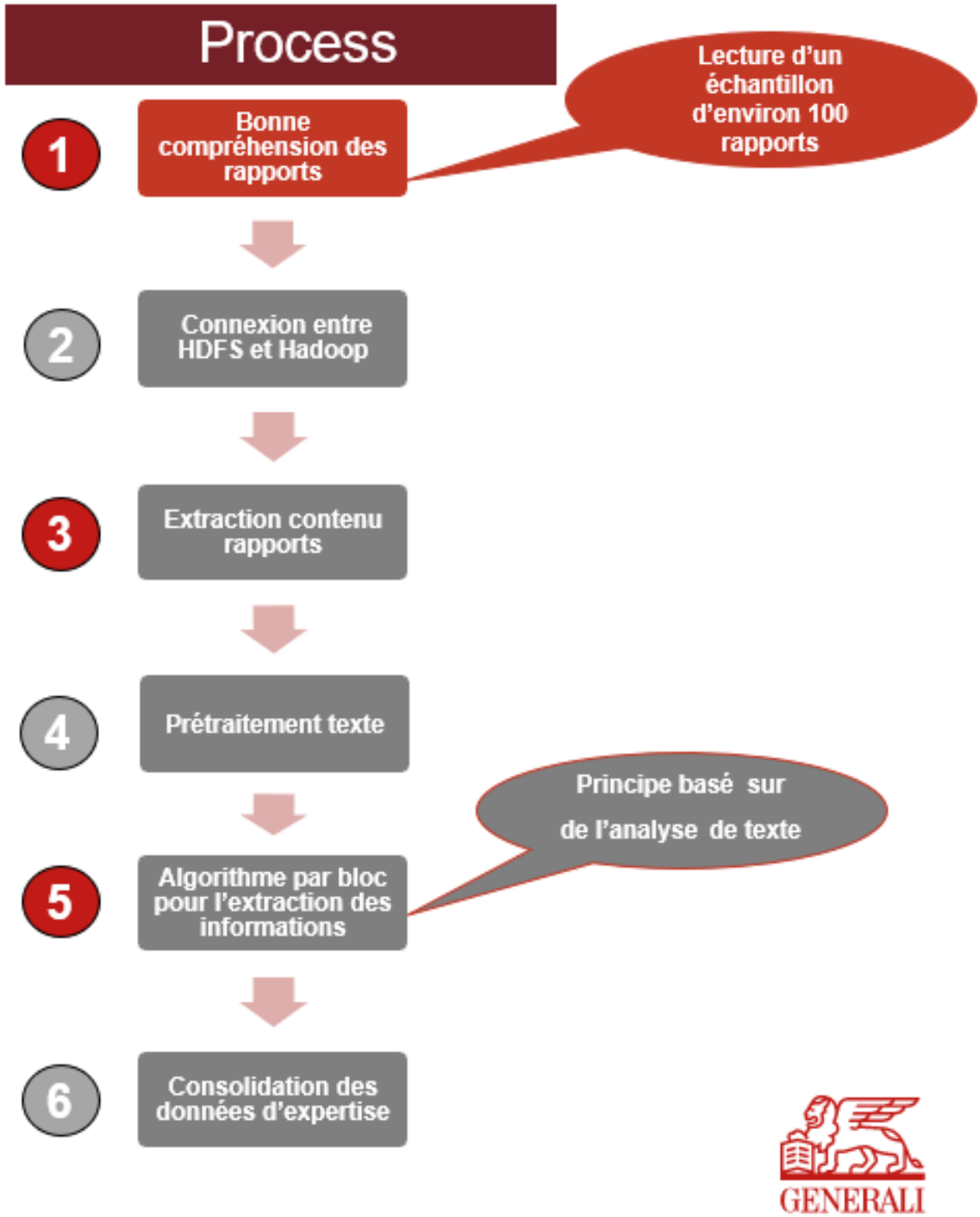
[ici]

Index	Liste_Des_Variables	Description
1	Numero_sinistre	Reference du dossier sinistre
2	Numero police	Reference du contrat
3	Type de risque	Maison, Appartement, Hôtel,...)
4	Adresse du risque	
5	Reserve sur garantie	Garantie remise en cause (oui/non)
6	Mesure de prevention	Mesures de prévention mises en place (oui/non)
7	Vérification du risque assuré	Oui/Non
8	Nature Garantie	Nature de la garantie applicable
9	Caractère de l'expertise	Unilatéral/Contradictoire etc
10	Type de recours	(exercé/subi)
11	Suppression de cause	Cause du sinistre a été supprimé (oui/non)
12	Qualité de l'assuré	Type d'assuré (locataire, propiétaire,...)
13	Type d'indemnisation	Autoréparation, Indemnisation pécuniaire
14	Statut indemnisation	Assuré a été indemnisé (oui/non)
15	Date du sinistre	Date de survenance du sinistre
16	Date d'effet	Date d'effet du contrat
17	Date Visite	Date de visite de l'expert
18	Code postal	Code postal
19	Montant recours	Montant de recours à exercer ou à subir
20	Produit	MRH/MRI/MRC/etc...
21	Capitaux	Montant des capitaux propres
22	Franchise	Montant de la ranchise
23	Surface	Surface de l'habitation en m^2
24	Nb pièces	Nombre de pièces disponible
25	Localisation des dommages	Cuisine, chambre, salle de bain, bureau ...
26	Localisation dommages 2	piscine, plafonds, dalle, plancher, arbre, mûr,...
27	Causes et circonstances	Zone de commentaires sur les causes et circonstances du sinistre
28	Convention applicable	Aucune, IRSI, CIDRE, ...
29	Tranche convention	Tranche de la convention applicable (1,2,..)
30	Type de rapport	Type de rapport (expertise sur site, télé expertise, reconnaissance, ...)
31	Type de dommage	embellissement/immobilier/mobilier..)
32	Montant de dommage	Montant des dommages indemnisés (indemnité immédiate, différée, totale)
33	Destinataire du règlement	Assuré, Tiers, Reneurs, ...

Index	Quelques variables sécheresses	Description
1	Nature plancher	dallage, bois, ...
2	Charpente	Bois,
3	Couverture	Ardoise, toiture,tuiles...
4	Arrêté CatNat	Oui, Non
5	Date arrêté	Date
6	Transfert de propriété	Oui/Non
7	Tranche mobilière souscrite	Montant
9	Nombre d'étage	1,2,3 ...
10	Conformité du risque	Risque vérifié (Oui,Non)
11	Couverture	Toiture, Tuile,...
12	Suppression de cause	Cause du sinistre a été supprimé (oui/non)
13	Nature sol	béton, carrelage, parquet...
14	Topographie terrain	Vallonné, en pente, plat...
15	Nature arbre	arbre, chêne, sorbier, saule, cerisier...

5 - Récapitulatif processus extraction des données d'expertise :

Le processus de mise à disposition d'un outil permettant d'extraire et de consolider une base centralisant les données contenues dans les rapports d'expertise se décline comme suit : [ici]



6 - Dictionnaire des mots clés : Dans le cadre de la récupération des données montants en zone tableau, il était nécessaire dans un premier temps de mettre en place un dictionnaire de référence alimenté par les mots clés relatifs aux dommages habituellement indemnisés suite à un sinistre.

Dictionnaire de référence :

```
list_of_type_dommage=dict({"immobili":"Immobilier",'embel':"Embellissement","contenu":"Contenu","mobili":"Mobilier"
                           ,"immateriel":"Immaterielles","annex":"FraisAnnexes","affer":"FraisAfferents"
                           ,"fuit":"RechercheFuite","nettoyag":"Nettoyage"})
```

7 - Extraction de la variable franchise :

Franchise

```
def extract_info_franchise(path):
    try:
        text_1_type_franchise=re.split("franchise",path)[1]
        f=text_1_type_franchise.split()[:7]

        if 'differee' in f:
            f = 'NR'
        elif '€' in f[0]:
            if f[0].replace('€','').isnumeric():
                f = f[0].replace('€','')
        elif f[1] == '€':
            if f[0].isnumeric() or f[0].replace(',','').isnumeric():
                f = f[0]
        elif f[2] == '€':
            if f[1].replace(',','').isnumeric():
                if len(f[1]) >=3:
                    if f[0].isnumeric():
                        f = f[0] + f[1]
                elif f[0].isnumeric() and f[1].isnumeric():
                    f = f[0] +','+ f[1]

        ### Tableaux
        else:
            text_2_franchise=re.split("tableau de reglement|tableau pour|tableau de",path)[1]
            text_3_franchise=re.split("franchise",text_2_franchise)[1]
            test = re.split("€",text_3_franchise)[0]
            if len(test) < 80:
                text_4_franchise=re.split('€',text_3_franchise)[1].replace(' ','')
                if len(text_4_franchise) < 10 and len(text_4_franchise) != 0 and text_4_franchise.replace(',','').isnumeric():
                    f = text_4_franchise


        if type(f) == list:
            f = 'NR'

    except:
        f="NR"


    return(f)
```

8 - Extraction de la variable numéro de sinistre :

Numéro de sisnistre :

```
def extract_info_text_refsin(path):
    try:
        text_1_assure=re.split("sinistre n|reference du sinistre|reference client|ref. sinistre |ref. client
                               |reference assureur|numero sinistre",path)[1]
        doscodc=text_1_assure.split()[0]
        if doscodc=="on":
            text_1_assure=re.split("reference  assureur",path)[1]
            doscodc=text_1_assure.split()[0]
        if len(doscodc)<=5:
            text_1_assure=re.split("reference compagnie|sinistre n|n  sinistre",path)[1]
            doscodc=text_1_assure.split()[0]
    except:
        doscodc="NR"
    return(doscodc)
```

9 - Extraction de la variable capitaux :

Capitaux

```
def extract_info_capitaux(path):
    try:
        l=[]
        liste = []
        compteur = 0
        text_1_type_capitaux=re.split("capitaux",path)[1]
        c=text_1_type_capitaux.split()[:7]

        if c[0] == 'concernes':
            c1 = text_1_type_capitaux.split('verification du risque')[0]
            c2 = c1.split('\n\n')
            for e in c2:
                if e == ' ':
                    break
                else:
                    l.append(e)
            for e in l:
                if '€' in e:
                    compteur += 1
                    if e.replace('€','').replace(',','').replace(' ','').isnumeric():
                        liste.append(e.replace('€','').replace(' ',''))
            if len(liste) != compteur or len(liste) == 0 or len(liste) > 10:
                c = 'NR'
            else:
                c = liste
        else:
            c = 'NR'

        if type(c) == list:
            for i in range(len(c)):
                c[i] = float(c[i].replace(',','.'))
            c = str(np.round(np.sum(c),2)).replace('.',',')
    except:
        c="NR"


    return(c)
```

10 - Illustration création d'une partie de la base de données :

Création bases

```
def creation_csv_SARETEC(csv_path,csv_name,min1,max1):
    Liste_franchise = list()
    Liste_surface = list()
    Liste_Type_rapport = list()
    Liste_Capitaux = list()
    Liste_nom = list()
    Liste_numero = list()

    for elem in bdd['Texte_RE_modif'][min1:max1]:
        Info_franchise=extract_info_franchise(elem)
        Info_surface=extract_info_surface(elem)
        Info_Type_rapport=extract_info_type_rapport(elem)
        Info_Capitaux = extract_info_capitaux(elem)
        Info_numero = extract_info_text_refsin(elem)
        Nblignes=len(bdd['Texte_RE_modif'][min1:max1])



        try:
                Liste_franchise.append(Info_franchise)
                Liste_surface.append(Info_surface)
                Liste_Type_rapport.append(Info_Type_rapport)
                Liste_Capitaux.append(Info_Capitaux)
                Liste_numero.append(Info_numero)

        except:
                pass

        df = pd.DataFrame(list(zip(Liste_numero,Liste_Type_rapport,Liste_franchise,Liste_surface,Liste_Capitaux)),
                    columns =['Liste_numero_sinistre','Liste_Type_rapport','Liste_franchise','Liste_surface','Liste_Capitaux'])
    #df=df.drop_duplicates([])
        df.to_csv(csv_path+csv_name+".csv",index=False,sep=";")
```

11 - Extraction de la variable recours :

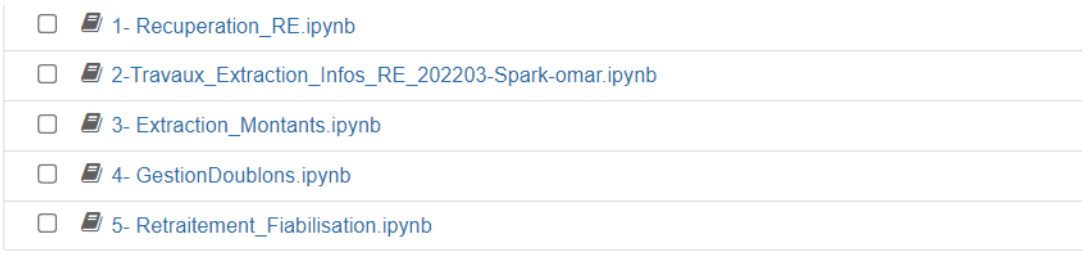
Recours à exercer :

```
# ok mais à verifier encore
def extract_info_text_recours_exercer(path):
    try:
        text_1_recours_exc=re.split("recours à exercer",path)[1]
        recours_exc=text_1_recours_exc.split()[0]
    except:
        recours_exc="NR"
    if len(recours_exc)==1 :
        try:
                    text_1_recours_exc=re.split("annexe 1",path)[1]
                    text_2_recours_exc=re.split("recours à exercer",text_1_recours_exc)[1]
                    recours_exc=text_2_recours_exc.split()[0]
        except:
                     recours_exc="NR"

    return(recours_exc)
#Recours à exercer :
```

12 - Synthèse outil automatisation extraction des données rapports :

Ci-dessous la liste des différents notebooks python à exécuter dans l'ordre dans le cadre de la consolidation de la base des données d'expertise.



13 - Illustration d'un bout de la base créés à partir des données rapports :

Ci-dessous une visualisation d'une partie de la base centralisant quelques données provenant des rapports d'expertise et qui permet aujourd'hui à Generali France d'atteindre une segmentation beaucoup plus fine dans le cadre de la ventilation de la charge sinistre.

Liste_Nom_RE	doscodc	Police	Garantie	IMMOBILIER	EMBELISSSEMENT	CONTENU	AUTRES	TOTAL
IARD-R198594305.docx	00BA456530	66058321	Dégat des eaux	0	5985,05	0	0	5985,05
IARD-R213790900.docx	70414445	54694822	Dégat des eaux	5495,09	1412,4	0	0	6907,49
IARD-R198528043.docx	48578324	56257922	Dégat des eaux	0	925,28	0	0	925,28
IARD-R239259354.docx	48622346	56257922	Dégat des eaux	929,5	0	0	0	929,5
IARD-R217312910.docx	39046014	000AH543812	Dégat des eaux	1926,66	0	0	0	1926,66
IARD-R214574581.docx	38339569	54914838	Dégat des eaux	0	763,98	0	0	763,98
IARD-R199317959.docx	48580520	000AA093450	Dégat des eaux	0	1069,5	0	0	1069,5
IARD-R214328940.docx	38338895	000AD547278	Dégat des eaux	0	309,36	0	0	309,36
IARD-R215284432.docx	70611430	000AH452646	Dégat des eaux	1177	0	400	0	1577
IARD-R214369000.docx	40432717	000AH482665	Dégat des eaux	0	4355,57	0	0	4355,57
IARD-R215297001.docx	66110465	000AL026617	Dégat des eaux	0	741,51	0	0	741,51
IARD-R213518679.docx	66105705	56335951	Dégat des eaux	574,47	1026,2	0	0	1600,67
IARD-R214564456.docx	38339888	000AD784859	Dégat des eaux	0	784	0	0	2681,6
IARD-R213570311.docx	40432112	000AL602525	Dégat des eaux	0	0	1039,2	0	1039,2
IARD-R215393209.docx	65004835	000AL871982	Dégat des eaux	0	2040,85	0	0	2040,85
IARD-R199470100.docx	48581809	17149276	Dégat des eaux	642	696,76	0	0	1338,76

List of Figures