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

Par : **Cherif Amadou SOW**

Titre du mémoire : construction d'un zonier en santé en utilisant les nouvelles méthodes de lissage : le krigeage.

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

Membres présents du jury de l'Institut des Actuaires

Secrétariat:

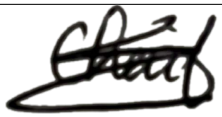
Bibliothèque:

Entreprise : GENERALI

Nom : GENERALI
Signature :
Directeur de mémoire en entreprise :
Nom :
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



Synthèse

Dans ce mémoire, nous nous intéressons à la modélisation d'un zonier technique en santé. La première partie de notre étude a été consacrée au traitement de données. Dans cette étape, nous récoltons dans un premier temps la donnée, ensuite nous nous focalisons sur la compréhension de cette dernière. Une action d'étude de la qualité des données a été menée afin d'identifier les problèmes à corriger ainsi transformer les données brutes pour qu'elles soient propres et exploitables pour la modélisation mise en œuvre dans la suite. L'étude de la qualité de la donnée est une étape importante car elle nous permet de mieux connaître nos données. Une bonne connaissance de ces dernières nous permet de faire des analyses cohérentes, des statistiques fiables. Plus nos données seront fiables plus notre modèle sera pertinent.

Les données utilisées sont fournies par les partenaires de l'entreprise Generali, selon un processus prédéfini. Chaque mois, nous recevons ainsi les données des clients ce que nous appellerons base contrat, les données relatives aux sinistres, aux primes encaissées et aux commissions versées aux délégataires.

Le schéma ci-dessous récapitule notre flux de données mensuelles.

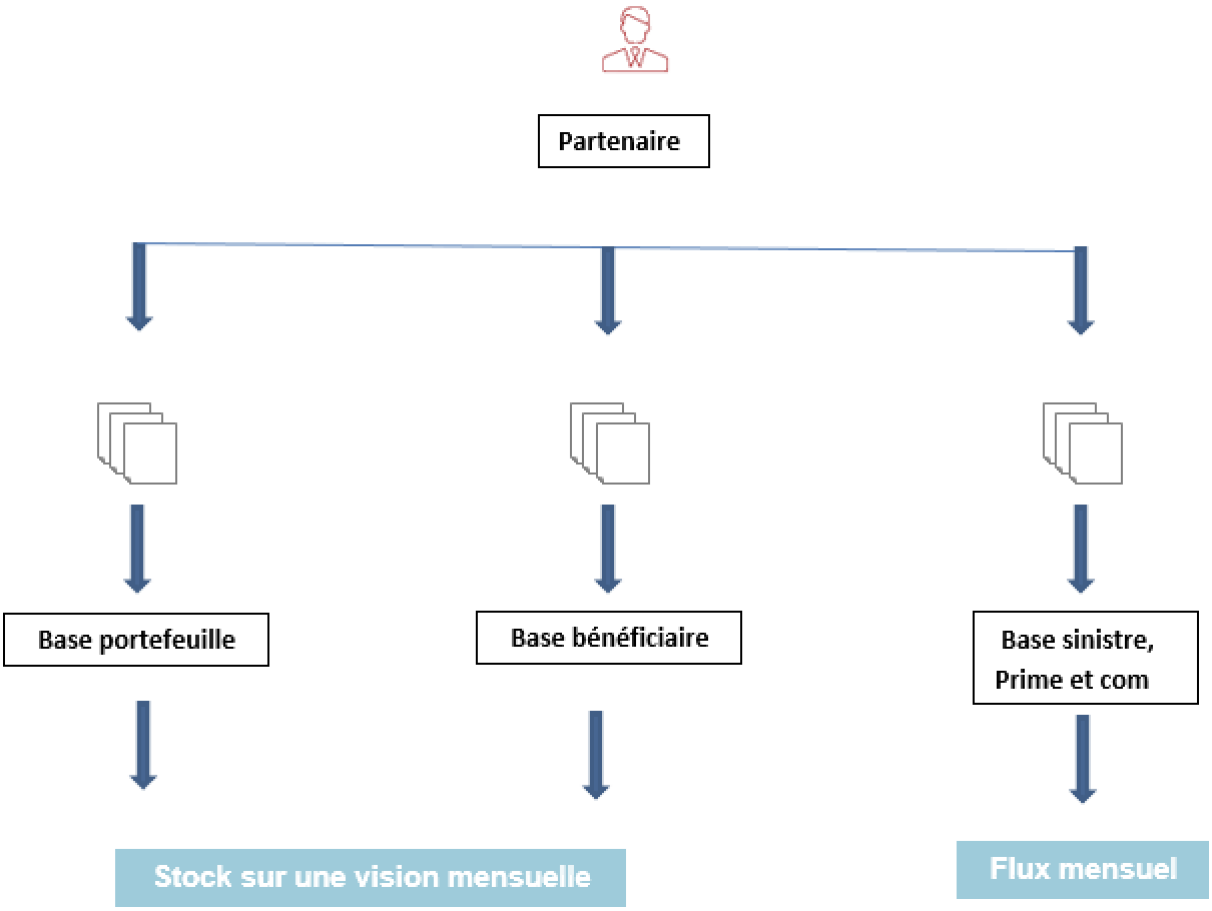


FIGURE 1 – Schéma de données

Nous avons créé une base d'étude à partir des bases mensuelles reçues de nos partenaires. En effet à partir de ces bases nous allons créer une historisation en faisant la concaténation des 3 années d'historique par base de données. Nous obtenons 5 bases historisées :

- une base concaténée contenant le flux des 3 années de sinistres.
- une base concaténée contenant l'information sur les 3 années du portefeuille.
- les autres bases contenant les bénéficiaires, primes et commissions.

Chacune des 5 bases contient une information qui complète l'autre base. Nous devons les joindre afin d'avoir une base unique pour réaliser des études transverses. Nous avons donc effectué la jointure entre les bases pour avoir une seule base technique.

En effet lors de la constitution de cette base technique, nous avons été confrontés à des problèmes de données et pour chaque anomalie détectée, nous avons défini une règle de gestion pour la contourner afin d'obtenir une base propre pour faire nos études. Parmi les corrections effectuées, nous pouvons donc en citer quelques-unes :

- doublons de bénéficiaires, en effet, pour une date donnée nous retrouvions plusieurs états pour un même contrat. Sur certains contrats, le partenaire anticipe l'inactivité du contrat à une date future ce qui fait qu'un contrat peut se retrouver avec plusieurs états possibles (en cours et inactif par exemple). Cette anomalie a été corrigée pour ne pas risquer de doubler les sinistres après la jointure avec la base bénéficiaire.
- nous avons également homogénéisé certains champs de notre base de données notamment les variables tarifaires comme les formules, l'âge et le régime ou encore les postes de garanties.
- nous avons dû exclure la CSP car elle n'est pas bien renseignée par le partenaire. En effet 67% des assurés avait une CSP vide.

En santé, la zone fait partie des variables tarifaires. Pour modéliser un zonier nous enlevons l'information géographique puis nous expliquons les résidus par l'absence d'informations géographiques.

$$Résidus = effet\ géographique \tag{1}$$

Nous ajoutons également des variables externes afin de montrer qu'une partie des résidus est expliquée par l'information géographique externe.

$$Résidus = effets\ géographiques + effets\ variables\ externes \tag{2}$$

Le schéma ci-dessous récapitule la méthodologie appliquée pour la création du nouveau zonier en passant par la modélisation des postes de garanties à la modélisation des résidus jusqu'à la classification des résidus par zone de risque.

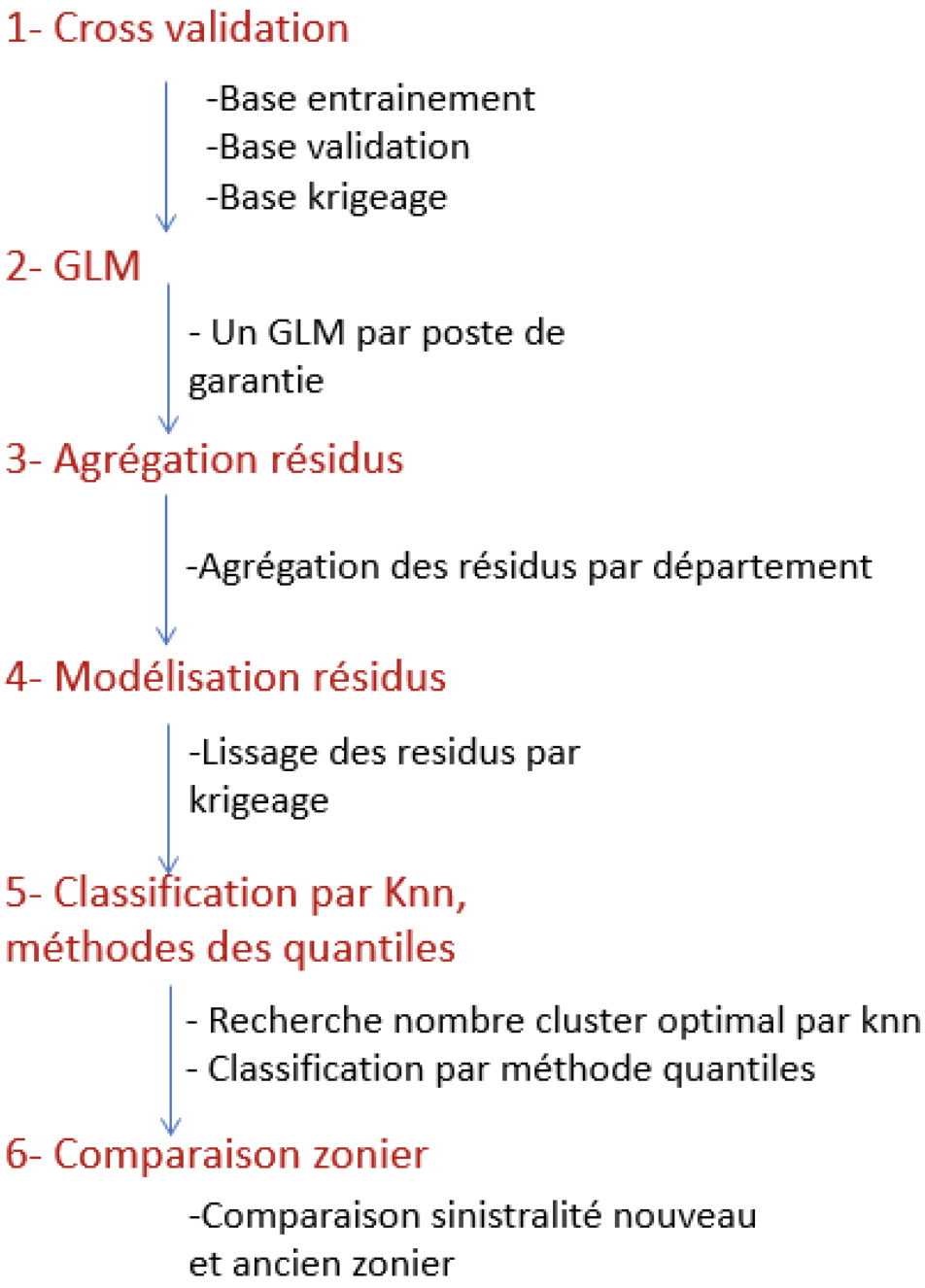


FIGURE 2 – Schéma de construction nouveau zonier

Suite à l'élaboration du nouveau zonier, nous pourrons le comparer au zonier du partenaire ainsi qu'au zonier de Generali hors partenariat. Ceci nous permettra de voir la différence entre le nouveau zonier, le zonier du partenaire et le zonier de Generali hors partenariat.

Comparaison zonier

Dans cette partie, nous comparons le zonier élaboré au zonier déjà existant ainsi qu'au zonier des agents généraux (hors partenariats) afin de déterminer le meilleur zonier.
Il est à noter que le partenaire dispose de 13 produits répartis sur deux zoniers différents. Chaque zonier est appliqué à des produits qui lui sont propres (zonier1, zonier2).
Il est également à noter que le zonier des Agents Généraux (hors partenariats) correspond au zonier des Agents Généraux hors direction des partenariats.
Le graphique ci-dessous illustre la représentation des zoniers sur le territoire de la France.

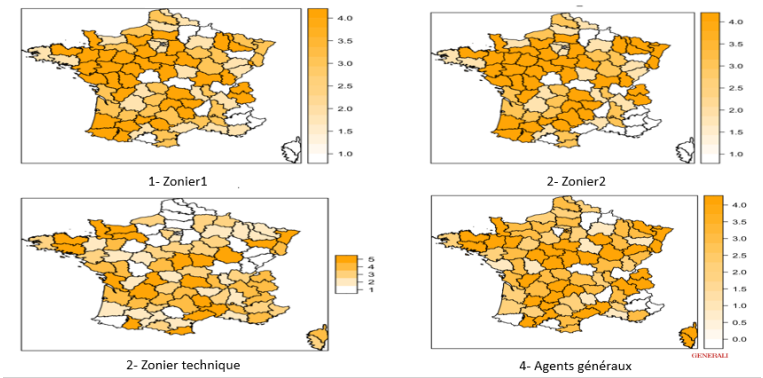


FIGURE 3 – Comparaison zonier

- La cartographie (figure 2) de notre zonier nous permet de voir que nous avons une meilleure répartition du facteur de risque géographique, nous avons une répartition plus homogène par zone de risque. Il prend en compte le risque géographique tandis que le zonier du partenaire est un zonier commercial et il n'est pas discriminant. En effet, certains départements comme le sud et de la France ont été classifiés dans des zones risqueés contrairement au zonier du partenaire qui les classe dans des zones non risquées pour des raisons commerciales.
  Ainsi nous pouvons dire que notre zonier technique prend bien en compte le facteur de risque géographique.
- Le distributeur cible majoritairement les seniors du régime général. Nous constatons que dans le sud de la France la population des seniors est bien représentée, mais aussi cette zone géographique est caractérisée par la présence des offres concurrentielles importantes. Pour être à la fois attractif et compétitif, le distributeur diminue son tarif et propose des tarifs commerciaux.
- Néanmoins, nous notons que notre partenaire dispose d'un portefeuille globalement rentable. Il se caractérise par une mutualisation du risque géographique entre les zones sous-tarifées et les zones sur-tarifées. Il comble une sous-tarification dans les zones où il veut être présent par la sur tarification dans les autres zones.
- Le zonier des Agents Généraux (hors partenariats)(figure 4) est également un zonier commercial. Cependant, nous constatons les mêmes classifications du risque géographique dans certains départements. On note par exemple certains départements du sud-ouest et quelques départements du sud-est notamment Marseille,

communs au zonier technique.

Summary

The first part of our study is devoted to data processing. The study of the quality of the data is a rather important step because it allows us to know our data more thoroughly. A good knowledge of the data allows us to make coherent analyses of reliable statistics and good prediction models, which allows us to better understand our risks. Every month we receive data flows from our partners that comprehend : a contract database, a claim database, a premium database and a commission database. This diagram below summarizes our monthly data flow.

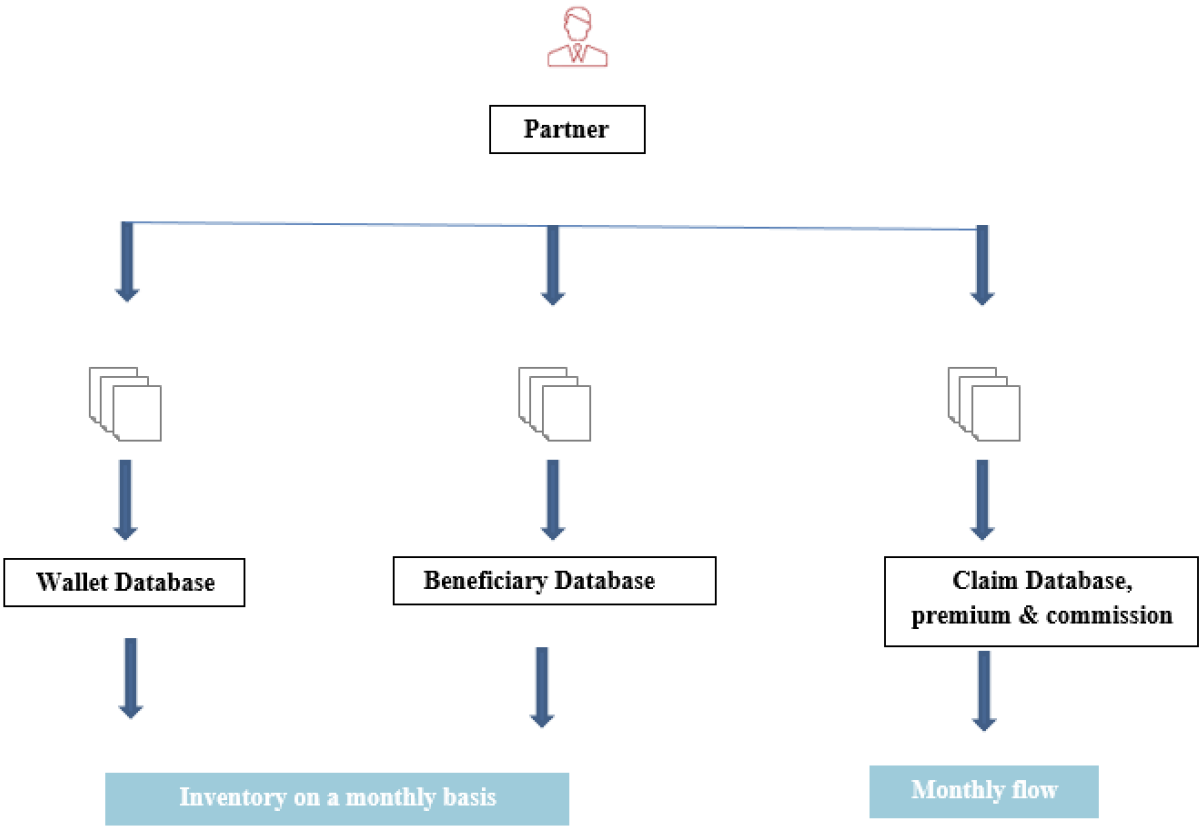


FIGURE 4 – Data schema

We have created a study framework from the monthly databases received from our partners. In fact, from these databases we will create a historical record by concatenating the 3 years of historical data per database. We obtained 5 historical databases :

- A concatenated database containing the flow of the 3 years of claims
- A concatenated database containing the information on the 3 years of the wallet.

- And the other databases containing the beneficiaries, premiums and commission.

Each of the 5 databases contains information that completes the other database, so we have to join them in order to have a single database for cross-sectional studies. We then joined the databases to have a single technical database. (Claim + Contract + premium + beneficiary + commission = final Database).

During the building of this technical database we were confronted with data problems and for each detected anomaly we defined a management rule to bypass it in order to have a clean database to make our studies. Among the corrections made, we can mention :

- Double beneficiaries, Indeed for a given date I found several states of the same contract. On some contracts, the partner anticipates the inactivity of the contract at a future date, which means that we can find a contract with several possible states (current and inactive for example). This constitutes an anomaly that must be corrected otherwise there is a risk of doubling the claims after joining with the profit base
- We have also homogenized certain fields of our database, in particular the tariff variables : formulas, age and plan and guaranteed positions.
- We had to exclude the CSP which is considered a price variable in health because it is not well informed by the partner. In fact, 67% of the insured had an empty CSP

The other concerns were related to the modelling. In health, the zone is part of the pricing variables. To model a zoning, we removed the geographical information and we tried to explain the residuals by the absence of geographical information.

$$Residuals = geographiceffect \tag{3}$$

We also added external variables to show that part of the residuals is explained by the external variables.

$$Residuals = geographiceffect + externalvariableseffect \tag{4}$$

The diagram below summarizes the methodology used to create the new zoning, from the modeling of the guaranteed items to the modeling of the residuals to the classification of the residuals by risk zone.

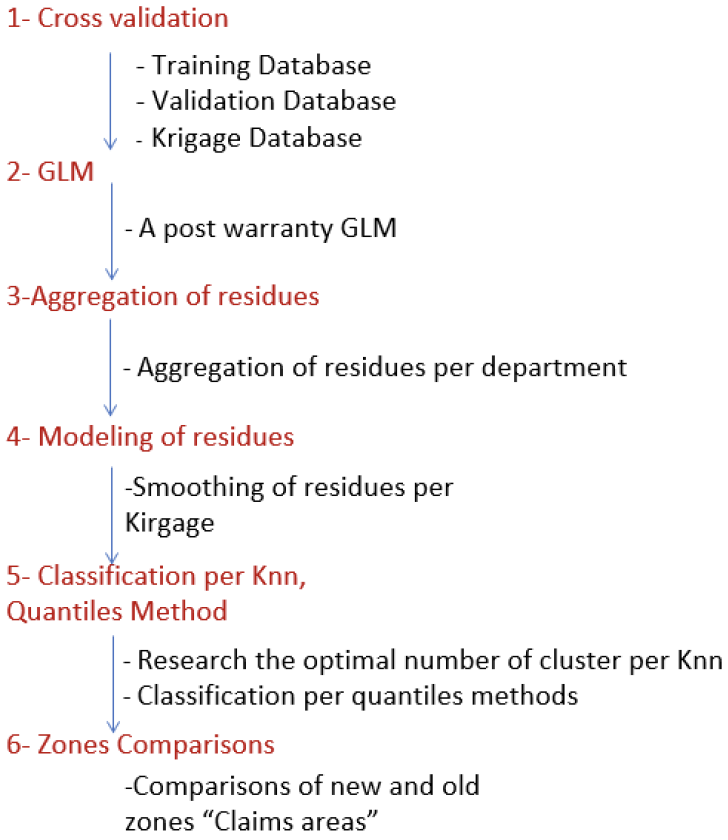


FIGURE 5 – Schéma construction nouveau zonier

Following the development of the new zone, we were able to compare it with the partner's zone and the Generali zone outside the partnership. This will allow us to see the difference between the new zone, the partner's zone and the Generali zone outside the partnership.

- The partner has two different zones called vitality zone and initial zone which he applies in these tariffs. Each zone is applied to its own products.
- The General Agents area (excluding partnerships) corresponds to the General Agents area outside the partnership's management.

Zones Comparison

In this part, I compare the elaborated zone with the already existing zone and the red mark zone in order to see the optimal zone.

The new zoning system that we have developed has a better distribution by department and takes into account the geographical risk, whereas the partner's zoning system is a commercial zoning system that does not sufficiently discriminate. Areas such as the South-East of France are risky areas. However these departments have been classified as non-risky by the partner for commercial reasons.

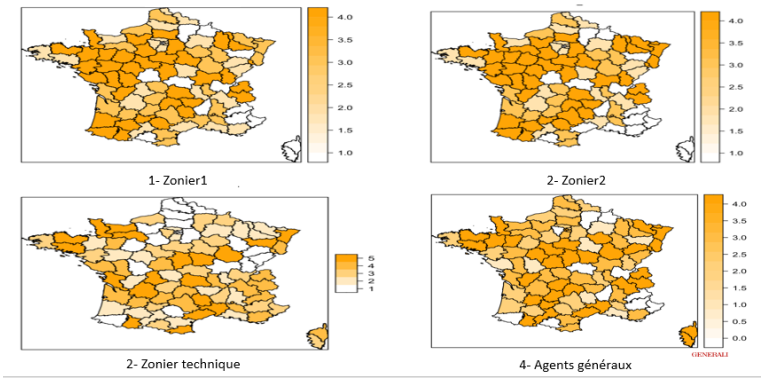


FIGURE 6 – Zoning comparison

The Red Mark zone is also a commercial zone and does not take into account the risk. However, in some areas we have the same risk classifications, the latter is true for some departments in the southwest and some departments in the southeast, notably Marseille

Remerciement

Je tiens à remercier Monsieur François-Xavier DUB, directeur du service Solutions PC et PH Partenariats et Madame Anne Sophie SANQUER, manager du service Modèle et Innovation, ainsi que tous mes collègues de service qui m'ont permis de m'intégrer rapidement au sein de l'équipe.

Je tiens à remercier Madame Anne Sophie SANQUER pour m'avoir accueilli au sein de Generali et dirigé mes travaux d'études durant ce mémoire.

Je remercie également Madame Laetitia GHAMMARTE, et Monsieur Hervé Trinh qui ont suivi mes travaux et ont été très disponibles.

J'exprime également ma gratitude à Monsieur Olivier LOPEZ, mon tuteur pédagogique.

Table des matières

Introduction

L'assurance santé représente une part importante de la branche vie des compagnies d'assurance. Elle est obligatoire et fait ainsi l'objet d'un marché large et très concurrentiel. Dans ce contexte, nous devons maîtriser nos risques afin d'améliorer nos tarifs et de pouvoir répondre à la demande des assurés selon leurs profils de risque. Face à cela, et avec la croissance de la volumétrie de données disponibles, les techniques de modélisation du risque géographique se sont développées et ont pris de l'importance grâce aux méthodes de lissage et de Machine Learning.

L'objectif de ce mémoire est de modéliser un nouveau zonier technique qui prendra en compte le facteur de risque géographique. Le portefeuille étudié concerne la distribution de contrat santé par le plus important courtier de la direction des Partenariats de Generali qui cible majoritairement des profils seniors. Le zonier référencé sur ce portefeuille a donc une orientation très commerciale afin d'obtenir des tarifs attractifs et faire face à la concurrence dans les régions ciblées par le distributeur, à savoir le Sud-est et l'Île-de-France. En effet, la population senior est très présente dans le sud de la France, et qui est également une cible de la concurrence ce qui pousse le distributeur à proposer des tarifs commerciaux pour être attractif. Ce mémoire aura ainsi pour objectif de construire un nouveau zonier technique afin d'analyser le caractère suffisamment discriminant ou non du zonier commercial.

Pour modéliser ce nouveau zonier, nous allons prendre en compte des variables tarifaires que nous allons considérer comme les variables internes puis ajouter des variables géographiques externes afin d'en étudier leur impact. Ces variables externes constituent un complément d'information pour la création de notre nouveau zonier. À titre d'indication nous avons pris en compte des indicateurs environnementaux afin de mesurer leurs impacts sur le nouveau zonier, ce qui est une des nouveautés de ce zonier. D'ailleurs, ces indicateurs environnementaux devraient être pris en compte dans la tarification des contrats santé. Nous avons également pris en compte d'autres indicateurs telle que la qualité de vie par département, la température moyenne ou encore le nombre de centres de santé par département.

Cette étude sera scindée en quatre grandes parties. Dans un premier temps, nous présenterons le cadre de l'étude et ses objectifs métiers. Ensuite, nous détaillerons les étapes clés de la construction de la base de données ainsi que les différents traitements effectués. Puis, nous allons modéliser la fréquence des sinistres en utilisant des modèles linéaires généralisés et enfin, nous lisserons les résidus obtenus de ces modèles linéaires généralisés par le krigeage. Nous allons également étudier la significativité des variables externes en introduisant des méthodes de Machines Learning telle que le Random Fo-

rest puis nous allons classifier les nouveaux résidus par classe de risque géographique en utilisant la méthode du KNN (K plus proche voisins) et la méthode des quantiles. Nous classerons ainsi le facteur de risque géographique du niveau le moins élevé au niveau le plus élevé. Enfin nous allons comparer le nouveau zonier élaboré avec les zoniers du partenaire.

Remarque : ces travaux portent sur un zonier fréquence.

Abstract

Health insurance is an increasingly significant branch of the insurance industry. It is mandatory insurance and is therefore the main focus of a very competitive market. In the current competitive market, it is a necessity to reduce the risks in order to improve the tariffs and meet the demands of insured according to their risk profiles. Faced with the latter and with the growth in the volume of the available data, geographic risk modeling techniques have developed and gained importance thanks to smoothing methods and machine learning.

The objective of this master's thesis is to model a new technical zoning that will take into account the geographical risk factor.

The studied wallet concerns the distribution of health insurance contracts by the most important intermediary of the Partnership Department of Generali. The zoning referenced on this wallet has therefore a very commercial orientation in order to obtain attractive tariffs in the regions targeted by the insurance intermediary, i.e. the South-East and the Ile-de-France. The objective of this thesis will be to build a new technical zoning system in order to analyze whether or not the commercial zoning system is sufficiently discriminating.

To model this new zoning, we will take into account tariff variables that we will consider as internal variables and add external geographic variables to our model in order to study their impact. These external variables constitute additional information for the creation of our new zone. We have taken into account the ecological impact, more specifically the polluting emissions and leisure, and other variables such as the median standard of living per department and the average temperature as well as the number of health centers per department etc.

This study will be divided into four main parts : First we will talk about the scope of the study. Then we will build our technical study base using the monthly flows received from our partner then we will model the frequency of claims using generalized linear models and finally we will smooth the residuals obtained from these generalized linear models by kriging, we will also study the significance of the external variables by introducing machine learning methods such as the Random Forest then we will classify the new residuals from the Random Forest by geographic risk class using the KNN method (K nearest neighbors) and the quantiles method. We classify from the lowest level to the highest level.

Première partie

Cadre de l'étude

Chapitre 1

Présentation de l'entreprise

Dans ce chapitre, nous allons dans un premier temps présenter les secteurs d'activités de Generali dans le monde et en France en particulier, ensuite, nous allons également présenter la direction des partenariats de Generali.

1.1 Présentation des secteurs d'activité de Generali

Generali est une compagnie d'assurance italienne fondée le 26 décembre 1831. Generali est l'une des 50 plus grandes entreprises mondiales selon le classement du magazine Forbes 2015. C'est la troisième compagnie d'assurances au monde, derrière Allianz et Axa. Elle compte 72 millions de clients dans le monde et possède une forte position en assurance-vie.

Aujourd'hui, le Groupe Generali est l'un des principaux assureurs au monde. Son chiffre d'affaires en 2019 s'élève à 69,8 milliards d'euros. Avec 71 936 collaborateurs à travers le monde au service de 60 millions de clients dans plus de 50 pays. Le Groupe figure parmi les leaders sur les marchés d'Europe occidentale, et connaît une forte croissance en Europe centrale et orientale ainsi qu'en Asie.

1.2 Generali France

Fondée il y a plus de 180 ans à Trieste, Generali s'installe dès 1832 en France, sa plus ancienne implantation étrangère. La filiale française a acquis au fil du temps diverses sociétés de l'Hexagone qui se sont regroupées progressivement pour aboutir en décembre 2006 à la création de l'entreprise unique Generali France. Le chiffre d'affaires de Generali France atteint 13,3 milliards d'euros en 2019. La Compagnie s'appuie sur 7000 collaborateurs pour offrir des solutions d'assurances à 7 millions de clients, particuliers ou bénéficiaires de garanties dans le cadre de leur activité, ainsi que 800 000 entreprises et professionnels.

1.3 La direction des partenariats

La direction des partenariats de Generali distribue des contrats à destination du marché des particuliers sur divers risques :

- 40% en Assurance automobile
- 30% en Assurance santé
- 13% en Assurance de Personnes hors santé (prévoyance, obsèques, emprunteurs)
- 13% en Dommages aux biens (MRH, produits Affinitaires...)
- 4% en Protection juridique

Afin d'assurer la robustesse de ce modèle de délégation, la direction des Partenariats s'articule autour de 3 directions.
Le schéma ci-dessous illustre l'organisation de la direction des partenariats.

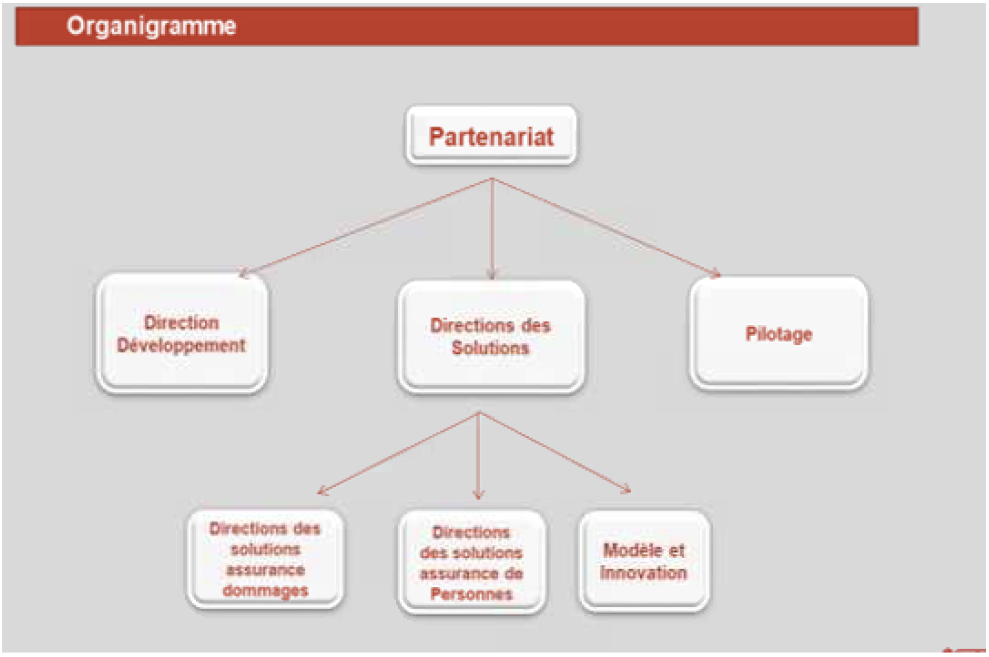


FIGURE 1.1 – Organisation du service

La Direction du Développement est responsable de la prospection et du développement de nouveaux partenariats ainsi que de nouveaux produits. La Direction du Pilotage s'assure de la mise en place des processus de gestion comme les reportings permettant ainsi de suivre et de répertorier les contrats. Il est à noter qu'au sein de cette direction, une équipe qualité de la donnée est responsable de la bonne réception et de la qualité des données de nos partenaires.

La Direction des Solutions qui est responsable des solutions techniques des partenariats tant en dommages qu'en assurance de personnes. Cette direction conçoit les tarifs, estime la rentabilité des produits et assure le suivi des partenaires.

Chapitre 2

Le marché de la santé.

Dans cette partie, allons d'abord présenter la sécurité sociale comme élément de contexte. Puis, nous allons aborder la complémentaire santé et ces mécanismes de remboursement. Ensuite, nous allons définir les principaux postes de garanties en santé. Enfin, nous allons parler de l'enjeu de la santé au sein de la direction des partenariats de Generali.

2.1 Sécurité sociale

En France, la Sécurité Sociale représente un ensemble d'organisations qui concourent à la protection de la population résidente contre tout risque social. Les conditions pour en bénéficier sont l'affiliation du bénéficiaire et de leurs ayants droits dans un régime qui est fonction de la situation professionnelle de l'assuré social. Le système est principalement constitué d'un régime général, d'un régime social des indépendants et d'un régime agricole.

2.2 Complémentaire santé et Mécanisme de remboursement

En plus de la caisse de sécurité sociale, la France dispose d'un autre système d'assurance maladie appelé complémentaire santé. La modalité d'usage est le remboursement partiel des soins.

En raison du désengagement progressif de l'Assurance maladie, il est devenu impératif d'adhérer à une mutuelle de santé afin d'alléger les dépenses sanitaires. Cela explique en partie le fort taux d'adhésion des Français à la couverture d'une complémentaire santé, soit 95%. En effet, les mutuelles prévoient de manière générale le remboursement des frais médicaux, et dans une certaine mesure le recouvrement de soins onéreux mal couverts par l'assurance maladie, notamment le dentaire ou l'optique. Ainsi, les modalités de remboursements diffèrent d'une structure à une autre.

Le principe de remboursement est présenté dans la figure ci-dessous :

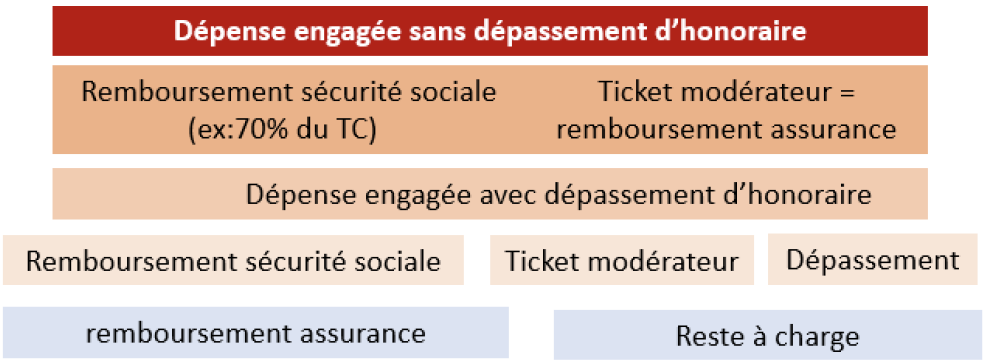


FIGURE 2.1 – Principe de remboursement

2.3 Postes de garanties

Conçus comme un ensemble d'actes médicaux, les postes de garanties s'établissent principalement comme ceci :

- l'hospitalisation englobe les frais de séjour, de chirurgie, d'anesthésie, de transport, d'une chambre particulière, ainsi que le forfait journalier hospitalier.
- le dentaire contient les soins dentaires, les prothèses dentaires, l'orthodontie et certains services non couverts par la Sécurité Sociale comme les implants.
- l'optique regroupe les frais en équipement, les lentilles de contact et la chirurgie réfractive.
- les soins courants englobent les consultations de toute nature : la médecine naturelle, les auxiliaires médicaux, les actes techniques médicaux, l'imagerie médicale, les frais de pharmacie, etc.
- il en existe d'autres postes qui regroupent diverses prestations. Ainsi, chaque mutuelle organise ces postes supplémentaires en fonction de ses modalités de travail.
- l'ensemble des actes médicaux qui ne se retrouvent pas dans les postes précités sont attribués à un poste autre.

2.4 Enjeux de la santé au sein des Partenariats

La santé a un enjeu important dans la Direction des Partenariats. Nous avons des produits divers avec des courtiers présents dans différentes zones et des cibles clientèles différentes. Dans le cadre de ce mémoire, nous nous focaliserons sur le partenaire majoritaire de la Direction des Partenariats qui représente 60% du chiffre d'affaires de la santé.

Deuxième partie

Traitement et présentations des données

Nous rappelons le business model de la direction des partenariats de Generali qui repose sur une distribution déléguée en marque blanche, c'est-à-dire, le porteur de risque est Generali mais, le marketing est au nom du distributeur. Il est à noter que la gestion est également déléguée à des tiers qui disposent de leurs propres tarifs, leurs propres outils de gestion et leurs propres zoniers. Les données étant externalisées, la donnée représente un réel enjeu pour la Direction des Partenariats.

De ce fait, l'objet de mon étude est de travailler en profondeur pour uniformiser les données reçues de nos partenaires, et construire un zonier en passant par la mise en place d'un modèle linéaire généralisé. En santé la sinistralité dépend de la zone, nous avons donc un tarif différencié par zone. L'objectif du zonier est de voir l'influence de la zone géographique sur la sinistralité et de permettre l'application d'un tarif par zone.

Dans un premier essai, nous sommes partis sur la construction d'un zonier pour tous les partenaires de l'équité. Sauf que nous avons été confrontés à un problème majeur qui est la complétude de la donnée collectée au près de nos partenaires. Par exemple, les identifiants clés qui permettent de faire la jointure entre les différentes tables, nous citons ici le cas du champ rang du bénéficiaire qui est une information cruciale qui permet de rattacher correctement le sinistre au bon bénéficiaire. N'ayant pas pu contourner ce problème, nous avons décidé d'orienter ce mémoire sur le partenaire qui représente plus de 60% du chiffre d'affaires en santé.

Dans cette partie, nous allons expliciter toutes les étapes nécessaires de notre traitement de données ainsi que les traitements effectués ainsi que les différentes transformations effectuées et les règles de gestion d'anomalies définies pour améliorer la qualité de la donnée du portefeuille retenu. Le contrôle de la qualité de la donnée est une étape importante, car elle nous permet de mieux connaître nos données, une bonne connaissance de celles-ci nous permet de faire des analyses pertinentes des statistiques fiables ainsi de réaliser de meilleurs modèles de prédictions afin de mieux appréhender nos risques.

Dans un environnement très concurrentiel et sinistré, il est impensable de travailler sur des données qui ne sont pas de qualité. Lors de ce traitement, nous avons été confrontés à diverses anomalies de données. Pour chaque anomalie rencontrée, nous avons défini une règle de gestion pour la contourner.

À la fin de ce chapitre, nous nous retrouvons avec une base de données que nous estimons propre à 99%. Ainsi, nous pourrons faire nos statistiques puis commencer la modélisation de la fréquence des sinistres.

Chapitre 3

Construction base étude

Les données utilisées pour cette étude sont issues des bases transmises par nos partenaires dans le cadre du suivi mensuel. Tous les mois, les délégataires nous envoient 5 bases de données à savoir : contrat, bénéficiaire, sinistre prime et commission.

La figure suivante synthétise notre schéma de données.

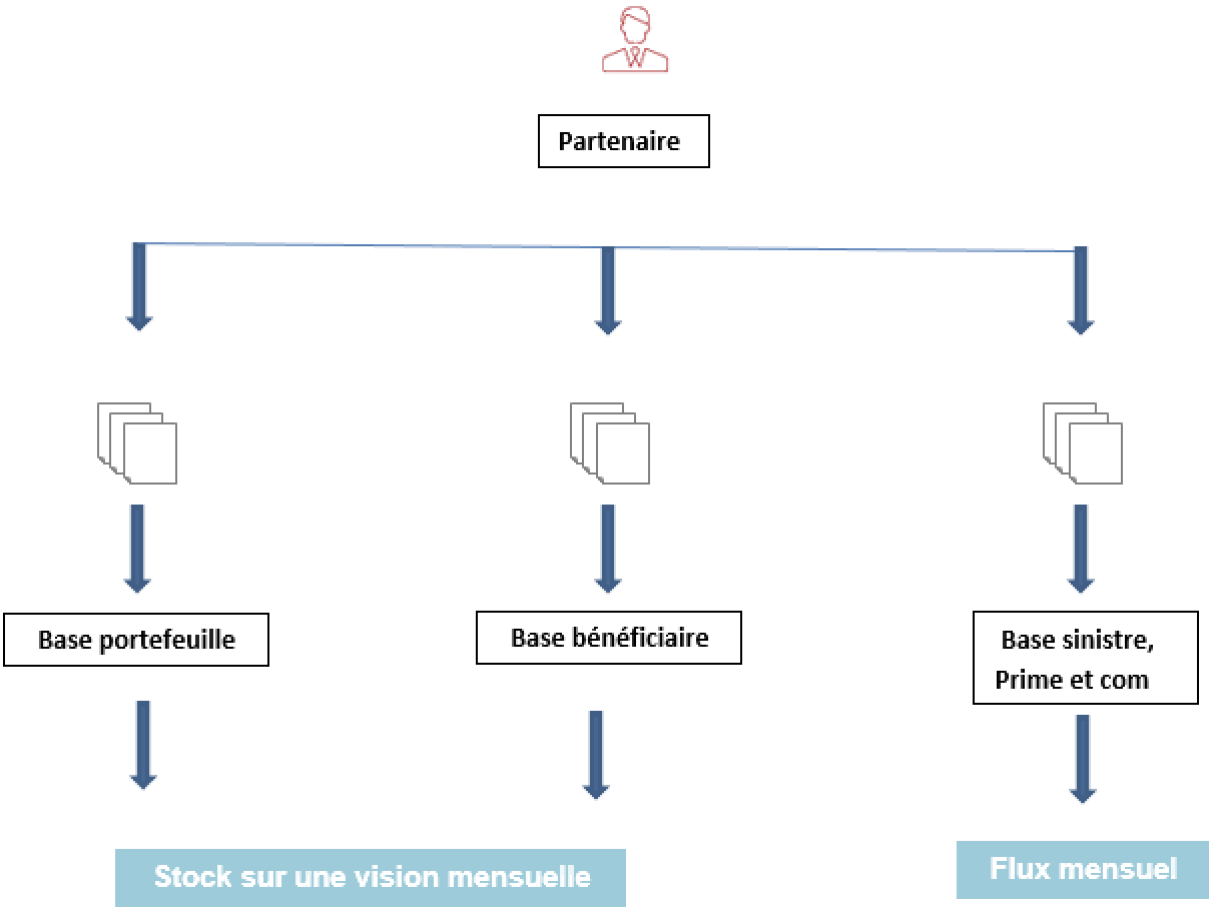


FIGURE 3.1 – Schéma de données

- La base sinistre permet de connaître le flux de paiement des sinistres par bénéficiaire. Elle contient l'ensemble des informations sur le sinistre notamment le

montant des remboursements, le taux de remboursement ainsi que les postes de garanties consommées etc.

- La base prime contient toutes les primes reçues des assurés. Au sein de cette base nous retrouvons les informations sur les primes versées par l'assuré ainsi que la période de couverture.
- La base commission contient les informations sur les commissions versées au délégataire.
- La base contrat quant à elle, est une photo du portefeuille et permet de dénombrer les contrats avec des données précises sur l'état des contrats (en cours, suspendu, résilié ou sans effet). Elle contient également les informations géographiques de l'assuré, la formule souscrite ainsi que le niveau de garantie, ou encore la date d'effet et la date de fin d'effet du contrat qui sera utile pour le calcul de l'exposition. On y retrouve également certaines variables tarifaires telles que le régime, la formule souscrite.
- La base bénéficiaire contient toutes les informations relatives aux bénéficiaires du contrat (âge, période de couverture. . .). On attribue à chaque bénéficiaire un rang.
  - Le rang 10 correspond au souscripteur
  - Le rang 20 correspond au conjoint
  - Les rangs 30, 31,32 . . . correspondent aux enfants du bénéficiaire.

  Le rang du bénéficiaire sera utile pour faire la liaison avec les sinistres. En effet, chaque sinistre est rattaché à son bénéficiaire.

Chaque base contient une information importante afin de pouvoir faire des statistiques descriptives, différentes études transverses et principalement construire notre zonier selon une approche de modélisation des modèles linéaires généralisés par poste de garantie. Pour ce faire nous devons rattacher le sinistre et la prime au bon bénéficiaire et ensuite l'associer au bon souscripteur.
Dans un premier temps, nous avons concaténé tous les flux mensuels pour créer une base qui recense un historique profond des informations des assurés. Ensuite, nous contrôlons la qualité de ces informations et nous repérons les différentes anomalies à retraiter. Ainsi, nous avons mis en place un système de règles de gestion qui permet d'optimiser et améliorer les données dans le but d'avoir une base propre et bien structurée.

3.1 Reconstitution historique de données

L'objectif de cette partie est de créer un historique de données à partir des données mensuelles reçues du partenaire.
Dans le cadre de ce mémoire nous avons reconstitué 3 ans d'historique de janvier 2018 à décembre 2020. En effet, la durée moyenne de renouvellement d'un portefeuille santé est de 3 ans, ce qui justifie notre décision. Nous estimons alors, qu'une profondeur de 3 ans va nous permettre d'avoir des profils de risques variés pour notre étude.
Ce travail a nécessité dans un premier temps de comprendre le data warehouse de l'équipe (l'entrepôt des données) et de comprendre la structure globale et les formats des bases envoyées par le partenaire.

Partant de 2018 jusqu'à 2020, nous avons développé un script en quatre étapes :

- récupérer le fichier brut fournit par le partenaire
- preprocessing : exécution d'un algorithme permettant de formater et adapter les données
- homogénéisation des bases pour pouvoir les traiter par la suite de manière équivalente
- concaténation des différentes bases issues des étapes précédentes
- Mise en place règles de gestion et correction des anomalies.

Le défi étant le nombre des fichiers important depuis 2018, mais également le nombre de produits dont dispose notre partenaires.
Spécificité de notre portefeuille : ce partenaire a 13 produits toutes cibles et seniors, ce qui fait la particularité entre ces produits c'est le BtB [1] et le BtC [2]. Ce qui diffère entre ces produits c'est le niveau de garantie proposé (précisé dans les tableaux de garanties), le zonier et le tarif appliqué.
Le script ci-dessus est un programme automatique qui tourne tous les mois pour la création de la base qui sera utilisée pour les études techniques, reportings, calcul des indicateurs de sinistralité, etc...

3.2 Mise en forme des bases en base cible

Dans cette partie, nous avons repris les bases historisées pour les mettre au format cible que nous avons défini. Chaque base a son format :

- pour la base contrat : nous ne sélectionnons que les informations nécessaires du souscripteur et nous définissons les noms des variables selon un cahier de charges.
- pour la base bénéficiaire : nous reprenons les informations nécessaires du bénéficiaire.
- pour les bases prime, sinistre et commission : nous reprenons les variables clés définies qui seront utiles dans les études transverses que nous allons mener.

Les bases contiennent un grand nombre de variables donc il est nécessaire de passer par une étape de "features selection" afin de sélectionner les variables qui pourraient rapporter le plus de signal pour nos modèles.

3.3 Contrôle de la cohérence des données

À ce stade, nous allons effectuer des tests de cohérence de la donnée dans l'objectif amélioration et fiabilisation de la qualité de données mais aussi, éviter que les erreurs opérationnelles biaisent notre étude. L'idée donc est de nettoyer la base des anomalies. De prime abord, nous contrôlons la cohérence des clés qui permettent de fusionner les différentes tables à savoir :

- **le numéro de contrat** correspond au numéro de contrat de l'assuré.

1. BtB : ils vendent les produits via des courtiers détaillants
2. BtC : ils vendent directement aux clients

- **le numéro de police mère** permet de retrouver le produit.
- **la date de traitement** correspond à la date d'observation du sinistre.
- **le rang de bénéficiaire** correspond au rang du bénéficiaire.

Les quatre clés ci-dessus sont présentes dans les différentes tables, elles nous permettent de faire la liaison entre elles. Contrôler la cohérence des clés consiste à vérifier d'abord la complétude de ces informations ensuite d'uniformiser le format de celles-ci. Par exemple si un contrat A est présent dans la base contrat, on vérifie s'il existe dans la base bénéficiaire etc... Par analogie, nous avons vérifié la présence des clés sur toutes les tables. À partir de ces 4 clés nous définissons une clé commune pour pouvoir agréger les bases afin d'avoir une base unique. Ces contrôles de cohérence effectués sur les clés nous permettent de ne pas perdre d'information après la jointure des différentes bases de données.

- **Date de traitement**
  Normalisation du format de date comme suit yyyy/mm/dd.
- **Police Mère**
  Vérification de la présence des mêmes produits.
- **Numéro de Contrat**
  Vérification de la présence des mêmes numéros de contrat.
- **Rang du bénéficiaire**
  Vérification de la présence des mêmes bénéficiaires.

Dans ce qui suit nous détaillerons l'ensemble des traitements effectués dans les tables en vue de corriger certaines anomalies.

3.3.1 Base bénéficiaire

- Nous avons constaté que sur certains contrats le partenaire anticipe l'inactivité du contrat à une date future ce qui fait qu'un contrat peut se retrouver avec plusieurs états possibles (en cours et inactif par exemple) ce qui constituait une anomalie qu'il fallait corriger sinon nous risquons de doubler les sinistres après la jointure avec la base bénéficiaire.

- Nous avons défini une règle de gestion pour supprimer tous les contrats qui sont en anomalies afin d'avoir une base retraitée propre pour mieux faire les jointures ceci nous permettra de ne pas perdre d'information.

- Nous avons également constaté que certains bénéficiaires étaient absents de la base de données alors que leurs sinistres étaient présents. Nous avons donc recréé ces lignes de bénéficiaires en reprenant les images c'est-à-dire en reprenant les informations du bénéficiaire disparu à partir des images précédentes.

3.3.2 Base Contrat

- Nous avons retraité les variables tarifaires puis nous avons calculé l'exposition au sinistre que nous allons expliquer dans la partie suivante.

- Nous avons constaté que 67% de la variable catégorie professionnelle était vide. Des actions sont en cours auprès de nos partenaires pour améliorer la qualité de la donnée. Nous avons mis en place un chantier avec nos partenaires pour améliorer la qualité de la donnée et les sensibiliser sur les variables que nous jugeons primordiales pour toute étude actuarielle.

3.3.3 Base Sinistre

- Nous avons vérifié si chaque sinistre était attribué à un bénéficiaire et que le numéro de contrat était présent dans la base contrat.
- Nous avons constaté que certains contrats n'étaient pas présents dans la base bénéficiaire. Nous avons dû redéfinir ces lignes de contrats dans la base bénéficiaire en reprenant l'image du contrat au mois précédent.
- Nous avons également constaté que certains rangs de bénéficiaires étaient vides dans la base bénéficiaire. Nous avons donc repris les informations du contrat des mois précédents pour corriger les rangs vides afin de pouvoir attribuer le sinistre au bénéficiaire.

3.4 Homogénéisation des champs

Dans cette section nous avons homogénéisé tous les champs notamment les variables tarifaires et les postes de garanties afin d'améliorer et de consolider la qualité de notre base de données. Cette étape consiste aussi à traiter les valeurs manquantes.

a) **Poste de garanties**
Mapping des codes Actes. Nous rappelons : les postes de garanties sont conçus comme un ensemble d'actes médicaux comme définit dans la section 2.3. Pour homogénéiser les postes de garanties, nous devons d'abord retraiter les actes garantis par partenaire, ensuite, nous classerons ces derniers par postes de garanties.
Le démarche suivie consiste à faire un mapping de tous les codes actes du partenaires que nous avons croisés avec les codes actes de la sécurité sociale que nous avons récupérés du site web officiel Noémie [3].
A cette étape, nous nous attendons à avoir des codes actes identiques à ceux de la sécurité sociale, sauf que le partenaire disposait de sa propre transcodification.

3. Code acte de la sécurité sociale

Pour palier à cette limite, nous avons isolé les codes actes du partenaire identique de la sécurité sociale ensuite isolé ceux qui ne sont pas de la sécurité sociale pour faire le croisement afin de retrouver le code de sécurité sociale correspondant. Ce slide ci-dessous résume la méthodologie adoptée pour le mapping des codes actes.

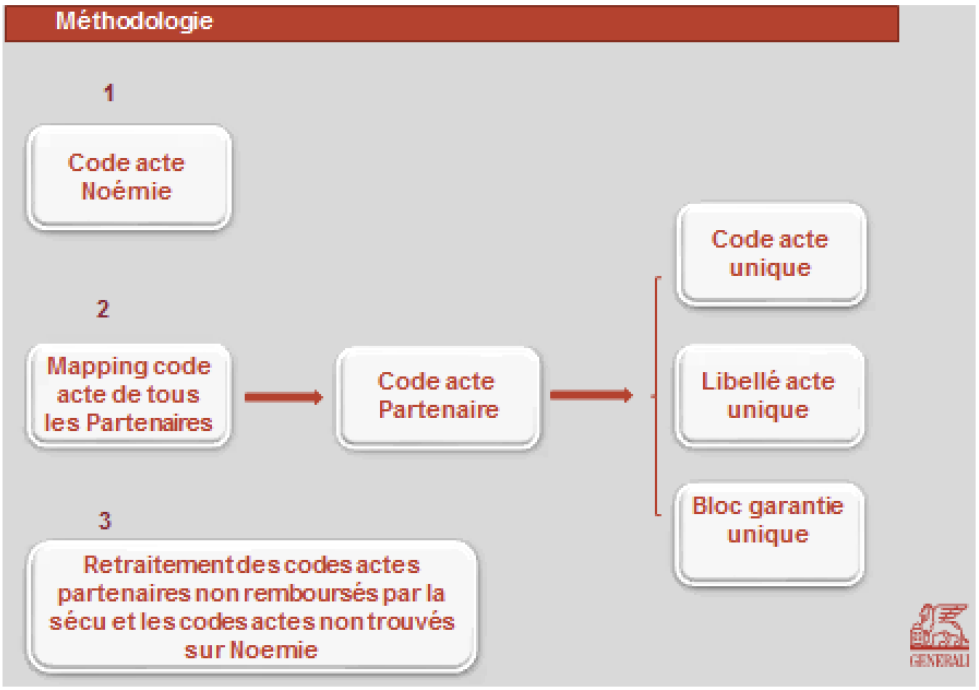


FIGURE 3.2 – Mapping code actes

Grâce au retraitement des codes actes, nous avons pu obtenir finalement les libellés actes qui caractérisent la classification commune des actes médicaux. Le regroupement des libellés en classe nous à parmi d'obtenir 6 postes de garanties définies ainsi : soins courants, hospitalisation, optique, dentaire, pharmacie etc...

En santé les postes garanties ne sont pas homogènes, ils présentent des couvertures différentes par conséquent des niveaux de consommation non homogènes. Il est donc important de les modéliser séparément pour avoir une meilleure modélisation.

b) **Formule**

Le contenu des produits d'assurance santé est principalement dépendant des garanties et des options souscrites. Ces garanties et options se combinent au travers des formules pour donner le produit tel qu'inscrit dans le contrat. Il existe alors une multitude de garanties permettant au souscripteur de personnaliser son assurance santé au point de quasiment bénéficier d'une couverture sur-mesure, ce qui permet aussi par ailleurs de définir le niveau de remboursement des soins.

La formule correspond au niveau de garanties souscrites et détermine le niveau de remboursement des soins. C'est la raison pour laquelle elle est considérée comme une variable tarifaire. Nous distinguons principalement 6 formules en santé :

- **Formule 1** : correspond au niveau de remboursement du ticket modérateur.
- **Formule 2** : elle correspond à un niveau d'entrée de gamme.
- **Formule 3 et 4** : elles ont des niveaux de remboursement de milieu de gamme, le niveau 3 étant inférieur au niveau 4.
- **Formule 5 et 6** : ce sont les niveaux de remboursement haut de gamme niveau 5 étant inférieur au niveau 6.

Notre partenaire dispose d'une gamme de produits étendue, pour chacun des produits nous avons une formule spécifique. D'où la nécessité d'homogénéiser cette variable en 6 formules uniquement. Ce travail va nous permettre de calibrer un GLM sur tous les produits. L'avantage d'avoir des formules homogènes est que la réalisation d'un nouveau modèle ne sera pas nécessaire en cas d'ajout d'un nouveau produit, nous pourrons directement utiliser le zonier pour ce nouveau produit.

c) **Regime**
Le régime est considéré comme une variable tarifaire. La variable de régime étant imparfaitement renseignée, pour traiter les données manquantes au niveau de cette variable, nous avons récupéré des visions antérieures ayant l'information.

d) **Âge**
Idem pour l'âge du bénéficiaire, il s'agit d'une variable tarifaire. Afin d'avoir des profils de risque homogènes, nous avons décidé de retraiter l'âge en le classant par catégories.
Nous obtenons les tranches suivantes :

- **Enfant** : moins de 6 ans
- **Mineur** : compris entre 6 et 18 ans
- **Adulte** : compris entre 18 et 60 ans
- **Sénior** : supérieur à 60 ans

e) **Département, Ville et Région**
Nous avons constaté que pour certains assurés l'information sur leurs départements étaient vides, nous avons ainsi utilisé l'information de la commune ou libellé du département pour compléter les départements vides. Nous avons repris l'information sur le site de l'INSEE pour ensuite ajouter la région correspondante.

f) **Calcul de l'exposition**
Nous rappelons aussi qu'en assurance, l'exposition représente la part de présence de l'assuré dans l'année où le contrat est exposé au risque.
Elle se calcule grâce à la formule suivante :

$$Exposition = \frac{date\ de\ fin\ du\ contrat - date\ de\ début\ du\ contrat}{365} \tag{3.1}$$

Une exposition de 1 signifie que le contrat est exposé toute l'année d'assurance souscrite. Par contre, une exposition de 0 peut être liée à l'état du contrat dans le portefeuille souvent, ce sont les contrats sans effet, résiliés ou inactifs. Par conséquent, il est important de retirer ces contrats de la modélisation pour ne pas biaiser.

g) **Calcul de la Fréquence et coût moyen d'un sinistre**
Le nombre de sinistres correspond au nombre d'actes consommés par un assuré. Nous avons créé un algorithme qui nous permet de calculer le nombre de sinistres. Cet algorithme prend en compte la date de soin, l'acte consommé et le bénéficiaire. Tous les actes consommés le même jour par le même bénéficiaire seront considérés comme étant un seul sinistre. Par la suite, nous utilisons le nombre de sinistres pour calculer la fréquence des sinistres. En effet, la fréquence des sinistres est le nombre d'actes moyen consommé par un assuré sur la période d'un an.

$$Fréquence\ sinistre = \frac{Nombre\ de\ sinistre}{Exposition} \tag{3.2}$$

Nous rappelons également que le coût moyen est la somme moyenne dépensée par un assuré pour un acte de soin.

$$Coût\ moyen = \frac{Montant\ sinistre}{Nombre\ de\ sinistre} \tag{3.3}$$

3.5 Jointure entre les différentes bases

L'objectif est d'agréger les 5 bases afin d'avoir une base unique pour faire nos études. Sur chaque base on retrouve des informations essentielles sur la mise en place d'un zonier, ces bases de données se complètent mutuellement.

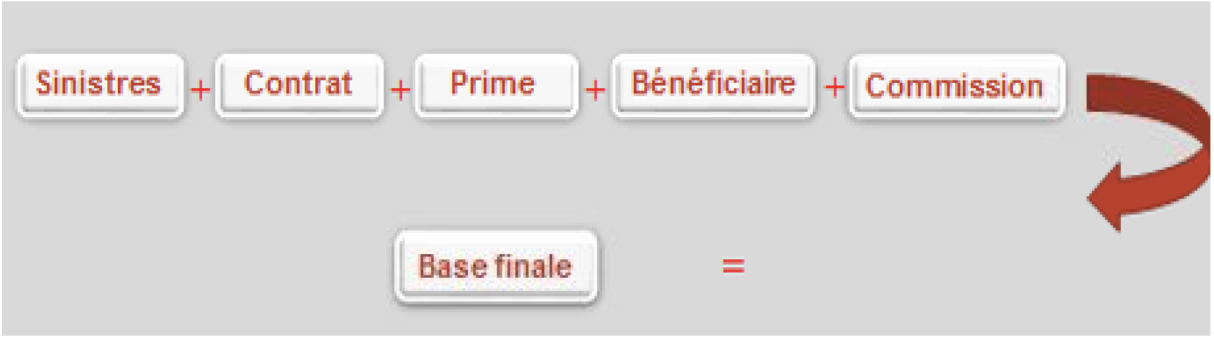


FIGURE 3.3 – Schéma construction base finale

Aprés avoir appliqué la jointure nous obtenons une base de 46 millions de lignes et 80 colonnes.

3.6 Présentation des données base cible

Pour information cette liste n'est pas exhaustive nous avons choisi les variables communes entre les différentes bases ainsi que les variables tarifaires.

Le tableau ci-dessous nous donne une liste de variables qui seront présentes dans la base uniformisée :

Champs	Description	Base
DATE_TRAITEMENT	Date d'observation du sinistre	Toutes les bases
N_POLICE	Numéro de police mère	Toutes les bases
N_CONTRAT	Société Anonyme	Toutes les bases
BENEF_RANG	Rang du bénéficiaire	Base bénéficiaire
BENEF_SEXE BENEF_AGE	Les informations du bénéficiaires	Base bénéficiaire
BENEF_ETAT	Etat du contrat s'il est en cours ou résilié	Base bénéficiaire
EQUI_REGL	Montant Remboursement Equité	Base sinistre
BLOC_GAR	Poste Garantie	Base sinistre
ADRESSE_SOUSC DEP_SOUSC	Les coordonnées du souscripteur	Base contrat
FORMULE	Niveau de garantie	Base contrat
RO_SOUSC	Régime obligatoire	Base contrat

TABLE 3.1 – Variables base unique

Nous avons utilisé de nouvelles technologies telles que Pyspark, Hadoop pour construire notre base de données. Notre base de travail contient 900 000 lignes et 17 colonnes : Âge, Poste garantie, les variables tarifaires, la fréquence des sinistres etc. . .
Dans toute la suite de cette étude nous allons utiliser cette base pour faire les statistiques descriptives ainsi que nos modèles.

3.7 Présentation et Traitement des variables externes

On part de l'hypothèse que la fréquence des sinistres ne dépend pas que des caractéristiques propres à l'assuré. C'est-à-dire que les composantes du risque d'un assuré ne dépendent pas exclusivement des variables tarifaires : âge, régime et la formule souscrite. On suppose l'hypothèse qu'habiter dans certaines zones influe directement sur la fréquence des sinistres. Ajouter les variables externes à notre étude constitue un complément d'informations pour la création de notre nouveau zonier.
Les variables externes que nous avons utilisées pour ce mémoire sont issues des conclusions publiées dans le cadre d'études gouvernementales ou de l'INSEE.
Nous décrivons dans ce qui suit les traitements mis en place pour une bonne qualité des données externes qui nous servira pour obtenir les résultats souhaités.

Le principal traitement consiste à l'agrégation des différentes sources externes par département.
Pour cela, il est nécessaire de télécharger les données relatives à la même période de temps, ici nous avons utilisé les données de 2018, 2019 et 2020. Aussi, pour les dépenses d'assurance santé remboursées par les CPAM, nous avons utilisé les dépenses hors prestations hospitalières (en raison de contraintes de Mémoire de nos PC). Les dépenses ont ensuite été calculées en moyenne par personne, permettant de savoir combien coûte en moyenne par personne les soins de santé remboursés dans chaque département.
L'agrégation par département a permis ainsi de pouvoir fusionner les bases de données externes (pollution, parcours sportifs et de santé, et dépenses d'assurance santé, température etc..). La base externe ainsi constituée a été fusionnée avec la base interne en utilisant comme clé de fusion les départements où vivent les assurés.
Pour justifier la prise en compte de la pollution, nous estimons que les assurés vivant dans les départements particulièrement assujettis à la pollution sont susceptibles d'avoir une santé moins bonne que les assurés vivant dans des départements peu ou pas pollués. Pour ces assurés, la fréquence des sinistres sera donc plus élevée que les assurés dans les zones moins polluées.
Le même raisonnement reste valable pour les assurés vivant dans les départements dotés de parcours sportifs et de santé. Nous souhaitons étudier si la disponibilité de tels équipements de loisirs et de sport dans un département offre une réelle amélioration des conditions de vie aux habitants du département en termes de lutte contre le stress et des autres pathologies liées au manque d'exercice physique et d'espace de distraction.
Pour ce qui est du nombre de centres de santé, et du niveau de vie médian ainsi que toutes les autres variables externes, nous estimons que ces variables ont un impact sur la sinistralité par département.

Liste des variables externes utilisées :

Variables	**Définition**
NB_D101	Établissement santé court séjour
NB_D103	Établissement santé long séjour
NB_D104	Établissement psychiatrique
NB_D108	Centre de santé
NB_D109	Structures psychiatriques en ambulatoire
NB_D110	Centre médecine préventive
NB_D112	Hospitalisation à domicile
NB_D113	Maison de santé pluridisciplinaires
NB_D301	Pharmacie
NB_D302	Laboratoire d'analyses et de biologie médicales
NB_D303	Ambulance
NB_D304	Transfusion sanguine
NB_F109	Nombre Parcours sportif/santé par département
qt_emi_pol_kg	Quantité pollution
Niv_vi_A_median	Niveau de vie annuel médian
temp_moy	Température moyenne

TABLE 3.2 – Listes des variables externes

Représentation de quelques variables externes par département.

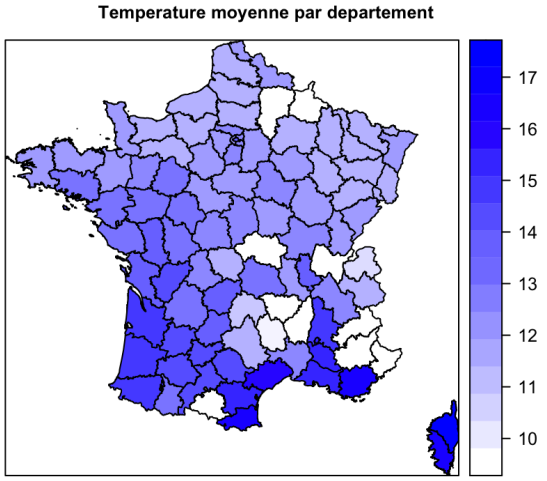


FIGURE 3.4 – température moyenne par département

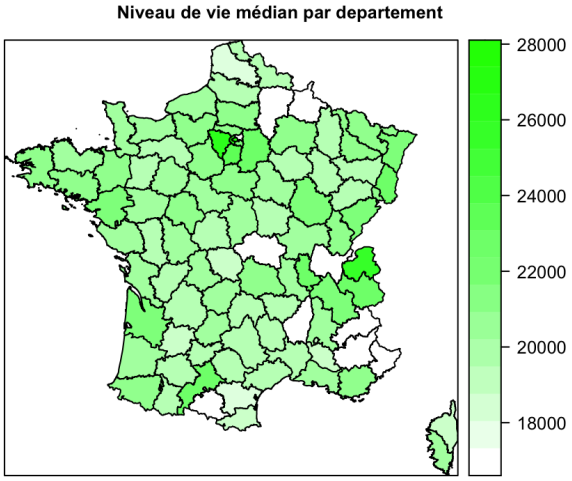


FIGURE 3.5 – Niveau de vie moyen par département

Nous constatons que l'Île-de-France est le département ayant le niveau de vie le plus élevé. Par ailleurs, on observe que les températures dans le sud de la France sont les plus élevées par rapport au reste du territoire. Dans la suite de ce mémoire, nous modélisons les résidus par l'ensemble des caractéristiques géographiques des communes correspondants. En prenant en compte la nature et la structure de ces données, nous considérons que les modèles de Machine Learning sont plus adaptés pour ce type de modélisation.

Chapitre 4

Statistiques descriptives

La partie sur les statistiques descriptives est une partie importante de notre étude, elle nous permet d'avoir une meilleure connaissance de notre portefeuille ainsi que le comportement de nos assurés selon l'âge, la formule souscrite, le régime et l'information géographique. Nos statistiques seront plus centrées sur le portefeuille, les variables tarifaires et les statistiques sur les informations géographiques.

4.1 Analyse statistique du portefeuille

4.1.1 Répartition des données par années

Le partenaire étudié représente plus de 60% du chiffre d'affaires de notre portefeuille. Les données utilisées sont comprises entre janvier 2018 et décembre 2020.
Le graphique ci-dessous représente la répartition des bénéficiaires de notre portefeuille par année.

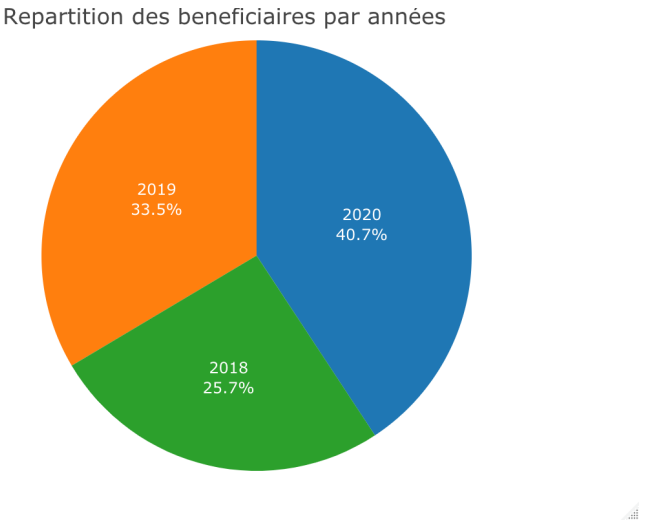


FIGURE 4.1 – Répartition des bénéficiaires par année

- Plus de 40% des bénéficiaires sont rattachés à l'année 2020.

- 33,5 % des bénéficiaires sont rattachés à l'année 2019.
- Le reste des bénéficiaires se concentrent sur l'année 2018.

On constate une bonne répartition de nos données d'étude sur les trois années. Cependant, on a un portefeuille qui grandit chaque année avec une production de plus de 80.000 affaires nouvelles par année.

4.1.2 Répartition des assurés par âge et par sexe

Cette pyramide des âges ci-dessous représente la répartition par sexe de notre portefeuille entre 2018 et 2020. Elle est constituée de deux histogrammes juxtaposés, un pour chaque sexe, par convention les hommes à gauche et les femmes à droite. Sa forme révèle les caractéristiques de la population.

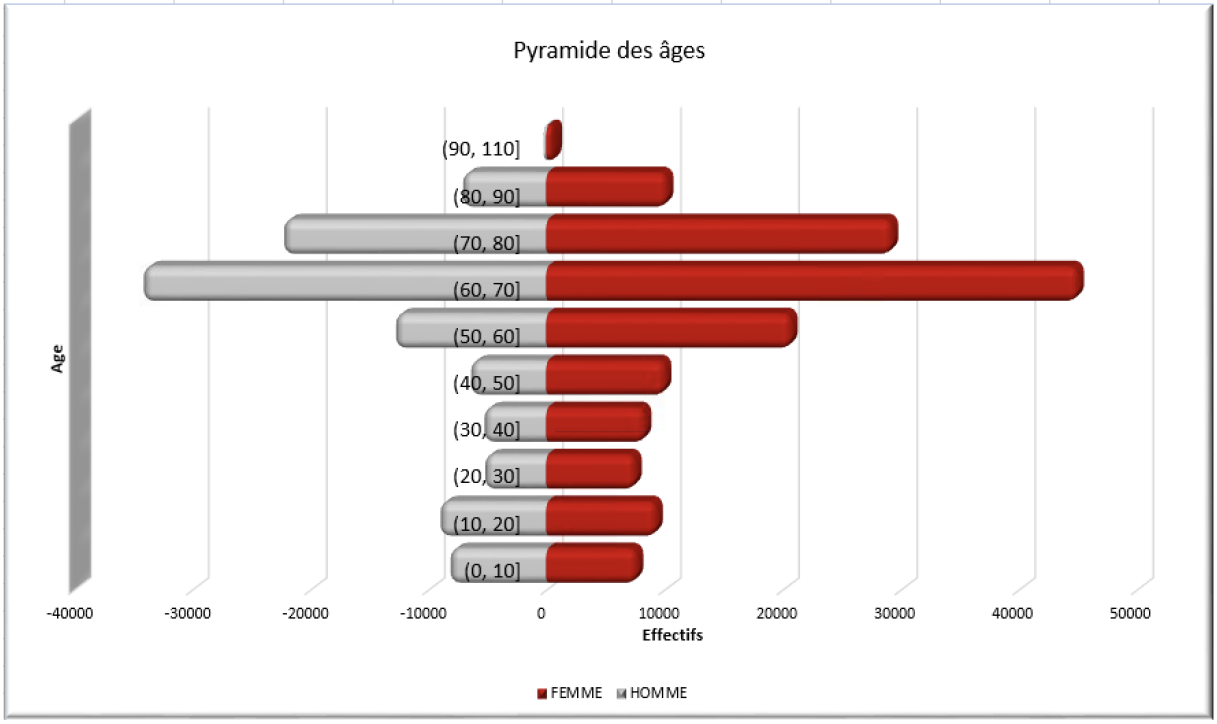


FIGURE 4.2 – Pyramide des âges

L'allure de la pyramide à hauteur large nous montre que la population est majoritairement composée de seniors. En effet, le portefeuille étudié contient des produits plus orientés sur les seniors. Nous constatons aussi que le nombre de femmes est supérieur au nombre d'hommes. Nous constatons également que le pic se situe entre 60 et 70 ans.
Cette pyramide reflète la stratégie commerciale du distributeur dédié qui cible majoritairement des profils seniors. Les profils non-seniors sont la population active disposant en général d'une mutuelle, ils n'ont pas besoin de souscrire à une nouvelle mutuelle. Ceux qui n'en ont pas souhaitent prendre une mutuelle qui rembourse le mieux. En effet, la majorité des salariés bénéficient d'une complémentaire santé dans le cadre

d'un contrat de groupe signé avec leur entreprise, qui en assume en partie le coût. Mais une fois à la retraite, ils doivent payer seul la cotisation.
Les seniors sont les retraités du régime général, en général ils ne disposent pas d'une mutuelle, ils font recours à des complémentaires santé pour mieux être remboursés. C'est la raison pour laquelle le distributeur dédié cible principalement les seniors.
La population senior est exposée à risque élevé de maladie, ceci dit un nombre d'actes médicaux beaucoup plus important par rapport aux jeunes. En tant qu'assureur, nous devrons être capables de faire face à un défi de surconsommation en ajustant nos tarifs de façon à permettre à cette population de seniors une meilleure garantie de vie.

4.1.3 Répartition bénéficiaire par produit

Le graphique ci-dessous montre la répartition de bénéficiaires par produit de notre portefeuille.

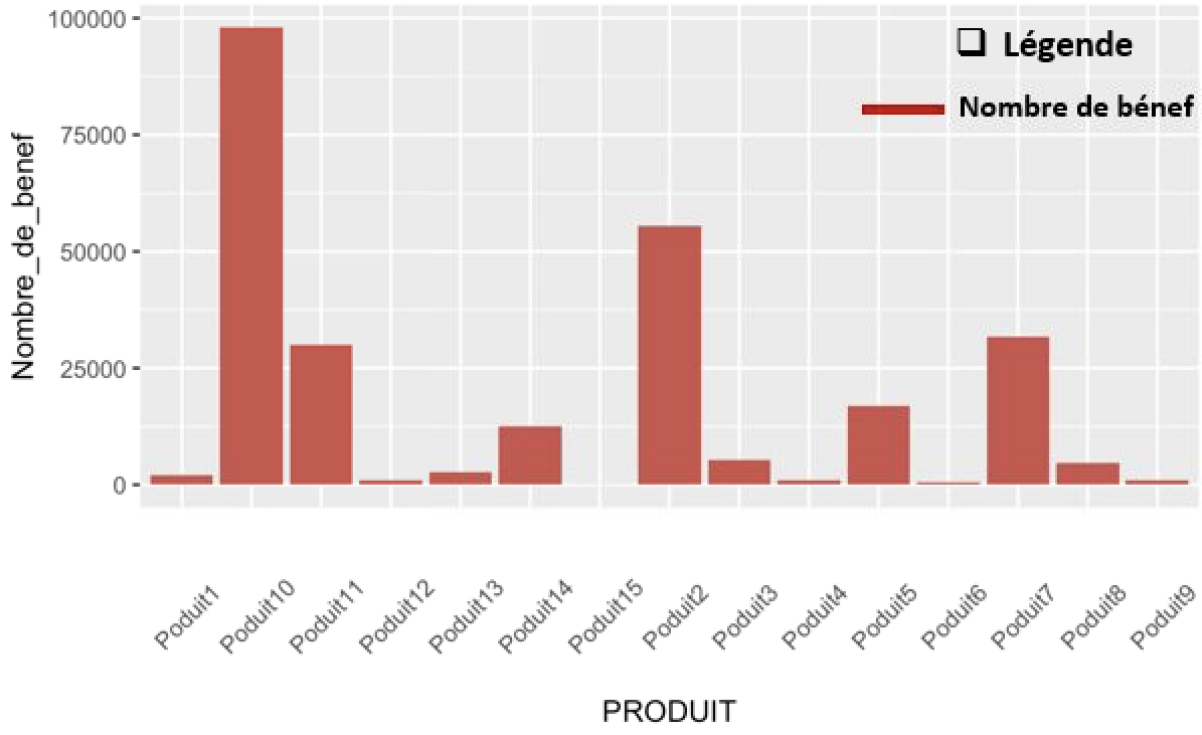


FIGURE 4.3 – Répartition des bénéficiaires par produit

Nous remarquons que les produits 10, 2 et 7 sont les plus gros produits du portefeuille retenu. Ils représentent plus de 70% des bénéficiaires.

4.2 Impact de la covid sur la sinistralité

Dans cette partie, nous nous intéressons à l'année 2020, il s'agit d'une année particulière au vu de la crise sanitaire internationale covid. L'objectif est de voir l'intérêt d'intégrer les données de l'année 2020 sur notre étude. Pour ce faire, nous analyserons l'évolution de la sinistralité par mois et par année.
Ce graphique nous montre l'évolution de la sinistralité sur les trois années.

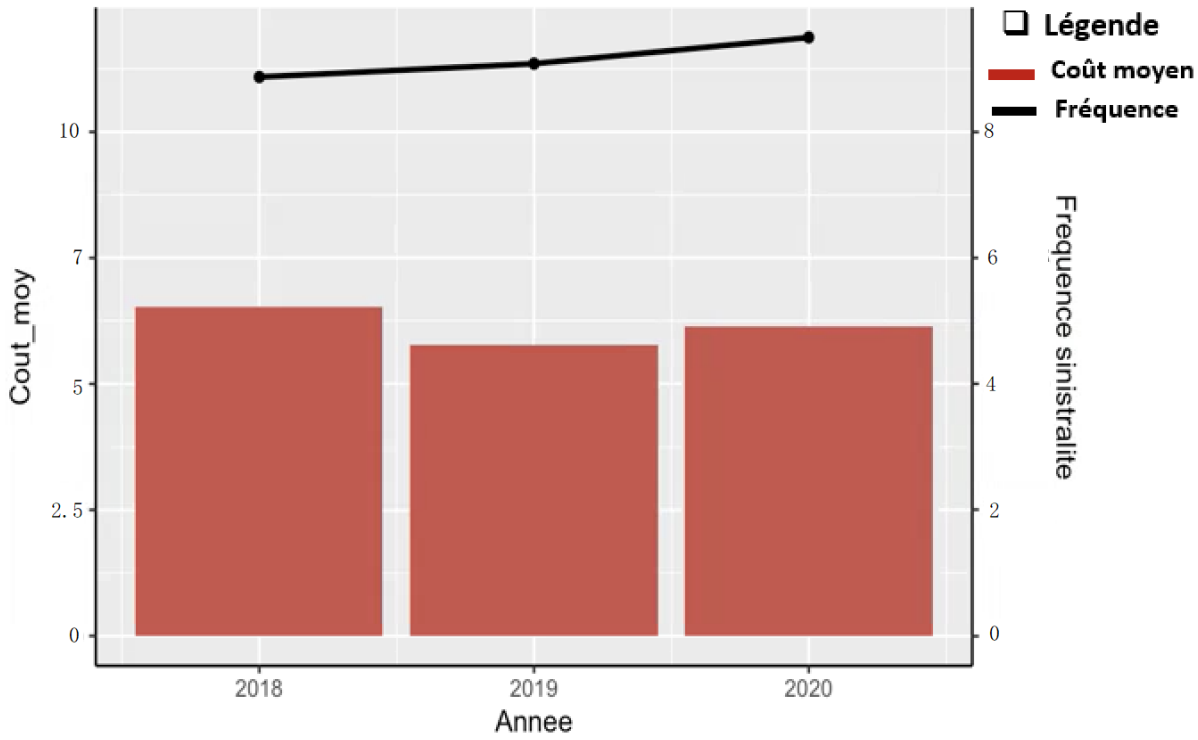


FIGURE 4.4 – Fréquence et coût moyen par année

La répartition de la sinistralité est assez homogène entre les trois années 2018, 2019 et 2020. Nous constatons que la fréquence des sinistres croit tous les ans.

Le graphique ci-dessous montre la fréquence et l'exposition par mois entre janvier 2018 et décembre 2020.

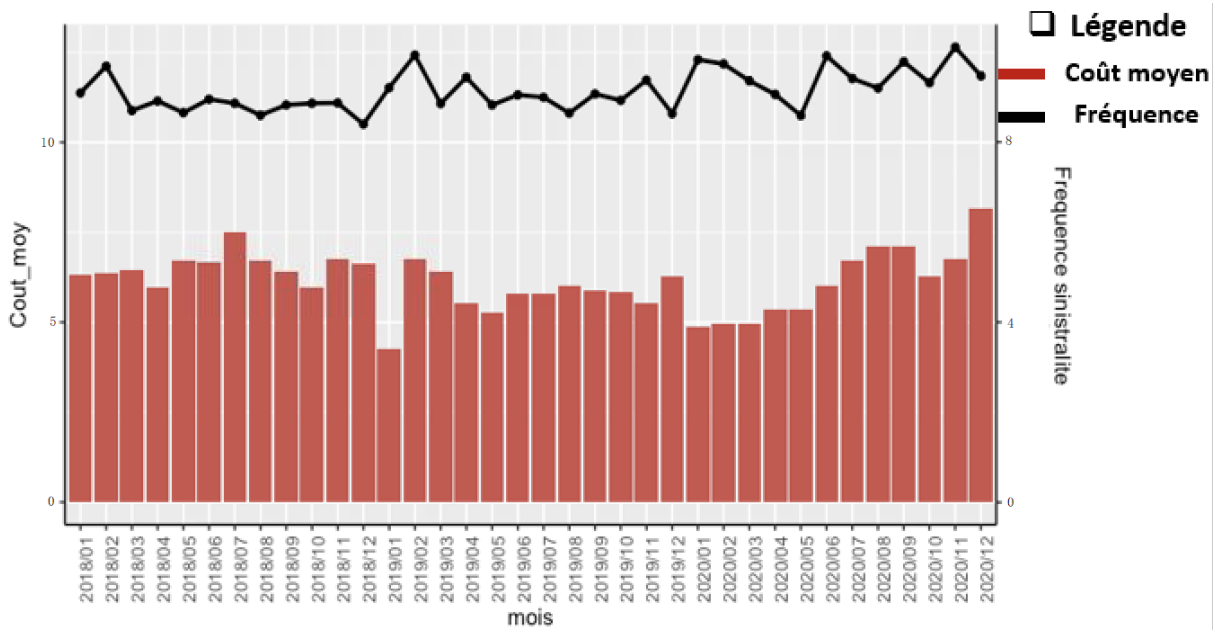


FIGURE 4.5 – Fréquence et coût moyen par année

Globalement, nous n'observons pas de tendance particulière sur les mois d'observations. Nous n'observons pas une sous-sinistralité sur l'année 2020 ou d'augmentation des sinistres sur cette dernière.
Cependant, nous pouvons noter une baisse de la fréquence des sinistres sur les mois de janvier, février, mars et avril jusqu'à atteindre un pic en avril. Cette période coïncide avec le début du confinement beaucoup de soins ont été reportés ou annulés. Toutefois, nous observons un rattrapage des sinistres en fin d'année.

Impact de la Covid par département

Ces graphiques ci-dessous illustrent la représentation de la fréquence des sinistres par département et par année.

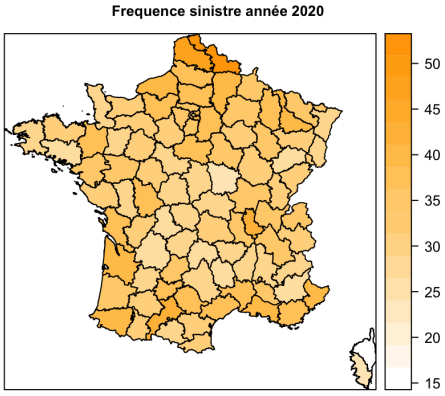


FIGURE 4.6 – Sinistre moyen par département sur l'année 2020

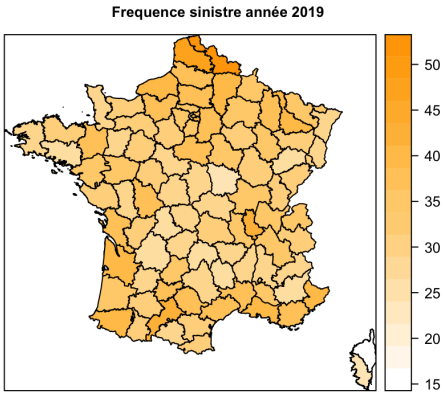


FIGURE 4.7 – Sinistre moyen par département sur l'année 2019

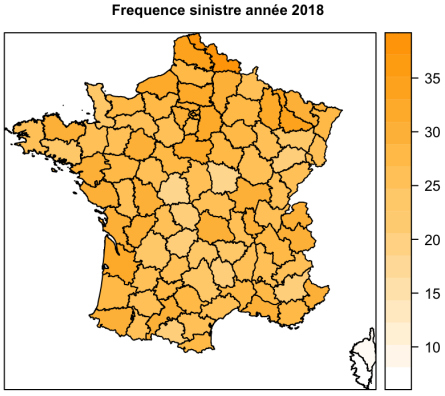


FIGURE 4.8 – Sinistre moyen par département sur l'année 2018

L'analyse montre qu'il n'y a pas de sur-sinistralité par département.
Globalement, nous constatons une homogénéité de la répartition des sinistres et nous pouvons donc intégrer les données de l'année 2020.
Cependant, nous observons un fort rattrapage sur la fin de l'année qui continue sur l'année 2021. Il serait intéressant d'étudier l'impact du Covid sur l'année 2021 avant de faire des études complémentaires. Ceci permettra également de constater les nouvelles formes de maladies qui verront le jour à cause des actes médicaux qui n'ont pas été respectés ou au manque d'activités sportives.

4.3 Statistique variable tarifaire

4.3.1 Statistique par poste garantie

La modélisation de la sinistralité du portefeuille en vue de construction du zonier nécessite de disposer d'un ensemble de variables tarifaires importantes. L'intérêt d'avoir suffisamment de variables est d'être sûr de modéliser fidèlement le risque non géographique avant l'extraction des résidus pour le zonier. Dans le graphique suivant, nous avons tracé la fréquence et l'exposition des sinistres par postes de garanties.

Ce graphique traduit la fréquence et l'exposition des sinistres par poste garantie

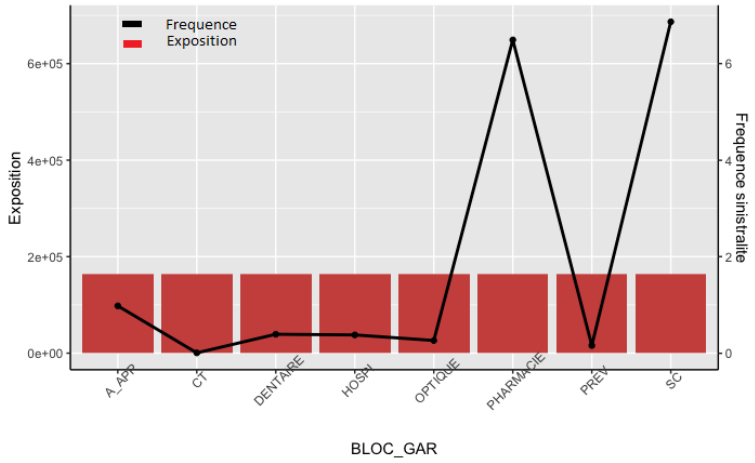


FIGURE 4.9 – Fréquence et exposition par poste de garaties

Nous constatons que la pharmacie et les soins courants portent la sinistralité. En effet ce sont les postes de garanties les plus sinistrés. Cependant le poste hospitalisation fait partie des postes les moins sinistrés.

4.3.2 Statistiques des variables tarifaires

- Âge

  L'Âge est considéré comme une variable tarifaire. En effet, plus on est âgé plus on est exposé au besoin de soin, c'est pourquoi l'âge intervient comme une variable tarifaire. Nous avons ainsi regroupé les âges par catégorie d'âge : mineur, enfant, adulte et senior afin d'avoir des classes d'âge homogène de risque.

  Ce graphique ci-dessous montre la fréquence et l'exposition par âge :

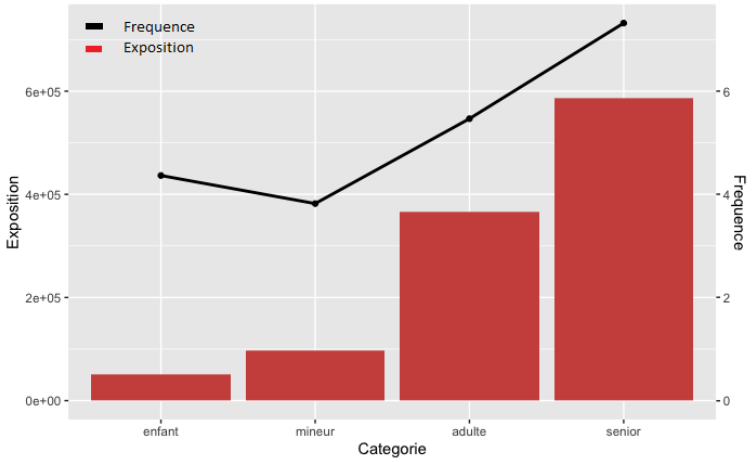


FIGURE 4.10 – Fréquence et exposition par âge

Nous avons un portefeuille qui est majoritairement composé de seniors, de plus nous constatons que la fréquence des sinistres augmente avec l'âge de l'assuré. Plus on vieillit, plus on est exposé aux sinistres.

- Formule

  La formule correspond au niveau de garanties détermine le niveau de remboursement en cas de sinistre. C'est la raison pour laquelle elle est considérée comme une variable tarifaire.

  Ce graphique ci-dessous montre la fréquence et l'exposition par formule :

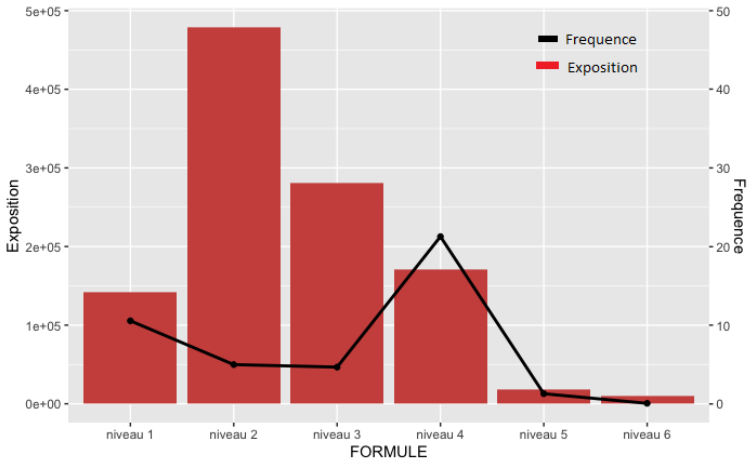


FIGURE 4.11 – Fréquence et exposition par formule

Nous constatons que la formule 2 est la formule la plus souscrite cependant la formule 4 est la plus sinistrée. Il s'en suit la formule 1 et 3. Les formules 5 et 6 sont les formules les moins exposées et les moins sinistrées.

Ce graphique ci-dessous montre le nombre de bénéficiaires par formule.

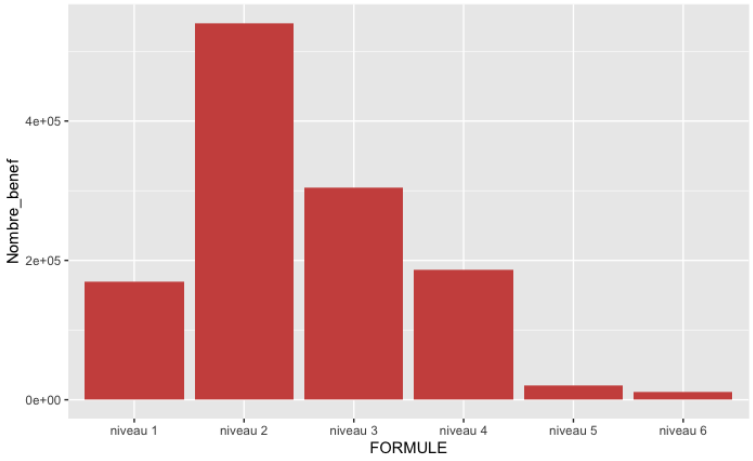


FIGURE 4.12 – Nombre de bénéficiaire par formule

Les formules 1, 2, et 3 sont en général les formules les plus souscrites car ce sont les formules d'entrée de gamme et proposent moins de garanties que les formules haut de gamme. Contrairement aux autres formules, les formules 5 et 6 sont les formules les moins souscrites car étant les formules de haut de gamme. Elles

peuvent être considérées comme étant des formules chères avec une couverture plus importante.

- Régime

  Le régime est considéré comme une variable tarifaire.
  Nous distinguons 4 régimes :
  régime général : considéré comme étant le régime des salariés, Il couvre les salariés, les retraités du secteur privé et les fonctionnaires.
  régime agricole : le régime agricole est un régime de protection sociale obligatoire qui couvre l'ensemble des salariés et non-salariés du domaine agricole.
  régime alsace Moselle : c'est un régime particulier de sécurité sociale ils bénéficient de tarifs plus avantageux.
  Régime TNS : travailleurs non-salariés

  Ce graphique ci-dessous illustre la fréquence et l'exposition par régime.

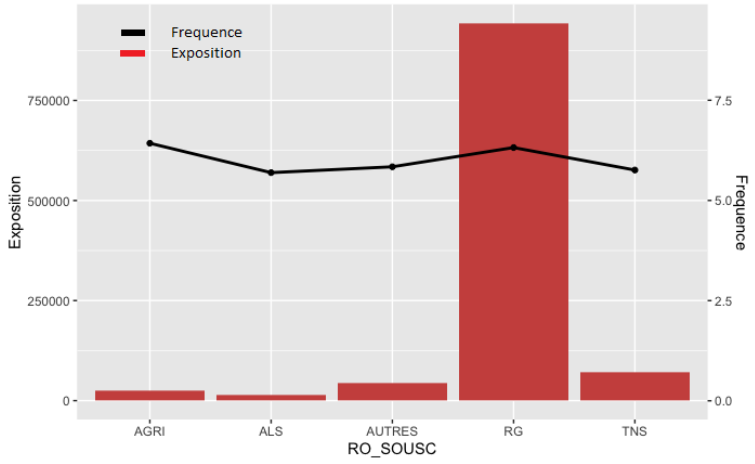


FIGURE 4.13 – Fréquence d'exposition par régime

Nous constatons que le régime général est plus exposé. Cependant, la fréquence des sinistres par régime est presque la même. En effet, nous constatons que les bénéficiaires sont majoritairement affiliés à ce régime.

Ce graphique ci-dessous nous montre le nombre de bénéficiaires par régime.

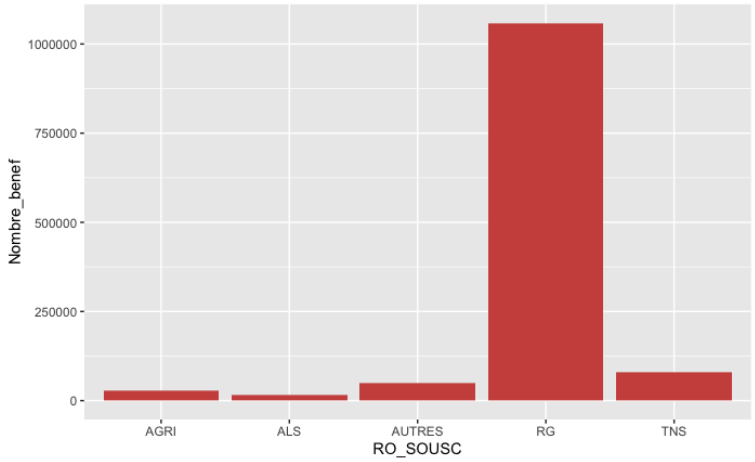


FIGURE 4.14 – Nombre de bénéficiaire par régime

Nous constatons que notre portefeuille est composé majoritairement des salariés du régime général. Les autres régimes sont moins représentés sur notre portefeuille ce qui montre que notre portefeuille est majoritairement composé de salariés du secteur privé et de fonctionnaires.

Les statistiques réalisées dans cette partie nous permettent de voir que notre portefeuille est majoritairement composé des seniors du régime général qui sont en général des retraités et qui souscrivent des formules bas de gamme. Ceci nous permet de conclure que nous travaillons sur un portefeuille avec un CSP [1] moyen.

1. CSP : catégorie socioprofessionnelle

4.3.3 Statistique ancien zonier

Nous rappelons que l'objectif de cette étude est de mettre en évidence l'importance d'une bonne segmentation géographique pour avoir un tarif d'assurance santé compétitif dans un marché extrêmement concurrentiel. À noter que notre partenaire utilise deux anciens zoniers pour tarifer ses produits. Les deux graphiques ci-dessous montrent la fréquence des sinistres selon les deux anciens zoniers.

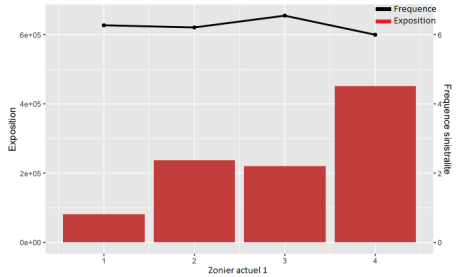


FIGURE 4.15 – Fréquence exposition ancien zonier1

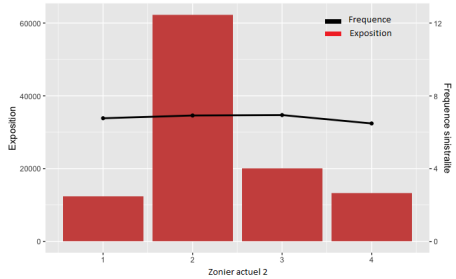


FIGURE 4.16 – Fréquence exposition ancien zonier2

En ce qui concerne le zonier 1, nous constatons que la zone 4 est la zone la plus exposée. Concernant le zonier 2, nous constatons que la zone 2 est la zone la plus exposée. L'évolution de la fréquence des sinistres est similaire entre les deux zoniers de plus, nous constatons presque la même tendance entre les 4 zones.

Chapitre 5

Aspect théorique

Dans ce présent chapitre, nous allons aborder tous les aspects techniques utilisés pour la calibration de notre zonier technique. Nous allons présenter les modèles linéaires généralisés que nous allons utiliser pour modéliser la fréquence des sinistres. Nous allons également introduire les méthodes de Machines Learning telles que le Random forest que nous allons utiliser pour étudier la significativité des variables externes. Nous allons également présenter les méthodes de classifications supervisées notamment le KNN (K plus proche voisins).

5.1 Méthode de validation croisée

La validation croisée est une méthode statistiques permettant de mesurer les performances d'un modèle prédictif sur de nouveaux ensembles de données.
L'approche des ensembles de validation consiste à partitionner les données de manière aléatoire en N sous-ensembles : un ensemble est utilisé pour tester le modèle et les autres N-1 sous-ensembles pour entraîner le modèle.
La performance du modèle est donnée par la moyenne des performances obtenues sur chaque validation croisée.
Ce schéma ci-dessous nous montre l'exemple d'un 4-fold Cross Validation pour l'indicateur du MSE.

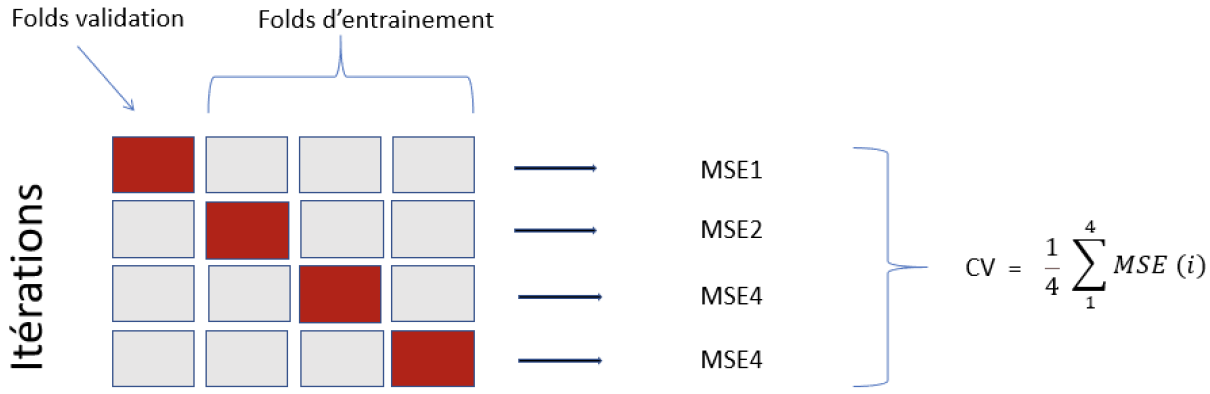


FIGURE 5.1 – 4 fold Cross validation

Métriques de performance du modèle

Le MSE [1] est la moyenne arithmétique des carrés des écarts entre prévisions du modèle et observations.

le RMSE [2] mesure l'erreur de prédiction moyenne faite par le modèle pour prédire le résultat d'une observation. C'est-à-dire la différence moyenne entre les valeurs de résultat connues observées et les valeurs prédites par le modèle. Plus le RMSE est bas, meilleur est le modèle.

La MAE [3] est une alternative à la RMSE qui est moins sensible aux valeurs aberrantes. Elle correspond à la différence absolue moyenne entre les résultats observés et prévus. Plus la MAE est basse, meilleur est le modèle.

5.2 Mesure de dépendance

Pearson

Mesure la corrélation linéaire entre deux variables quantitatives. Le coefficient de corrélation linéaire entre X_1 et X_2 est défini par :

$$r(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{VARX_1}\sqrt{VARX_2}} \tag{5.1}$$

Kendal

Disposant de deux couples (X_1, X_2) et (X_1', X_2') et identiquement distribués τ de Kendall se définit comme la probabilité de concordance moins la probabilité de discordance. Le τ de Kendall associé au couple (X_1, X_2) de variable aléatoire possédant des fonctions de répartition marginales continues est définie par :

$$\tau(X_1, X_2) = P[(X_1 - X_1')(X_2 - X_2') > 0] - P[(X_1 - X_1')(X_2 - X_2') < 0] \tag{5.2}$$

où (X_1', X_2') est indépendant de (X_1, X_2) et possède la même loi que ce dernier.

Spearman

Considérons le couple $X = (X_1, X_2)$ de fonction de répartition marginale F_1 et F_2 continues et définissons (X_1', X_2') une version indépendante de X (i.e X_1' admet comme fonction de répartition jointe F_1F_2).

Le ρ de Spearman est :

$$\rho(X_1, X_2) = 3(P[(X_1 - X_1')(X_2 - X_2') > 0] - P[(X_1 - X_1')(X_2 - X_2') < 0]) \tag{5.3}$$

C'est-à-dire que le tripe de la différence des probabilités de discordance de X_1 et X_1'.

5.3 Random Forest

Random Forest effectue à la fois l'échantillonnage en ligne et en colonne avec l'arbre de décision comme base. Les modèles h1, h2, h3, h4 sont plus différents qu'en ne faisant que l'ensachage en raison de l'échantillonnage sur colonne.

1. MSE : Mean Square Error, MCE : Moyenne des Carrés des Erreurs
2. RMSE : erreur quadratique moyenne
3. MAE : Erreur absolue moyenne

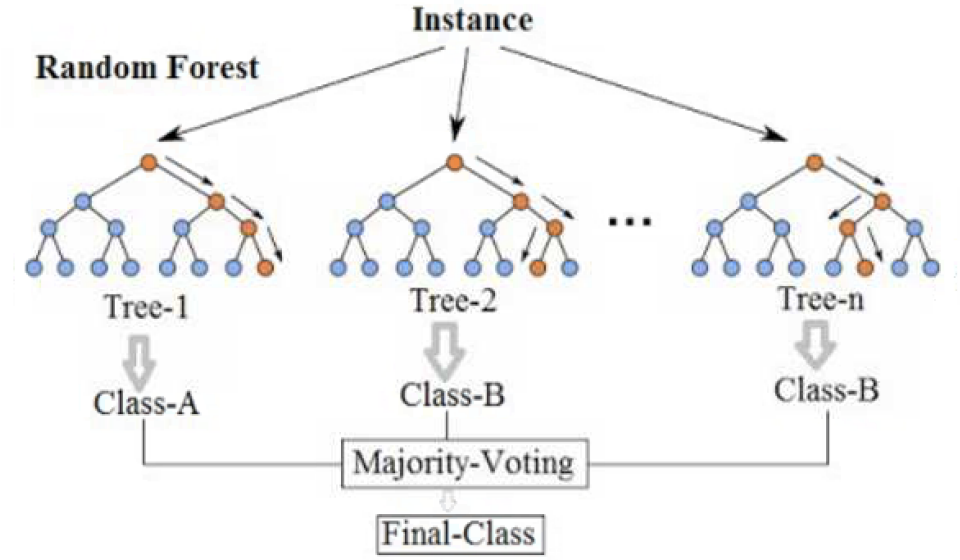


FIGURE 5.2 – Random forest

À mesure que vous augmentez le nombre d'apprenants de base (k), la variance diminuera. Lorsque vous diminuez k, la variance augmente. Mais le biais reste constant pendant tout le processus k peut être trouvé en utilisant la validation croisée

Forêt aléatoire = DT (apprenante de base) + ensachage(échantillonnage en ligne avec remplacement) + ensachage de caractéristique(échantillonnage e colonne) + agrégation(moyenne / médiane, vote majoritaire)

(5.4)

Ici, nous voulons que notre apprenant de base ait un biais faible et une variance élevée, alors nous entraînons le DT sur toute la longueur. Nous ne sommes pas préoccupés de la profondeur, nous les laissons croître car à la fin la variance diminue dans l'agrégation. Pour le modèle h1, les jeux de données (D-D ') non utilisés dans la modélisation sont hors du jeu de données. Ils sont utilisés pour la validation croisée du modèle h1.
Regardons les étapes prises pour implémenter une forêt aléatoire.

1. Supposons qu'il y ait N observations et M entités dans l'ensemble de données d'apprentissage. Tout d'abord, un échantillon de l'ensemble de données d'apprentissage est prélevé au hasard avec remplacement.
2. Un sous-ensemble de M caractéristiques est sélectionné au hasard et la caractéristique qui donne la meilleure division est utilisée pour diviser le nœud de manière itérative.
3. L'arbre est devenu le plus grand

4. Les étapes ci-dessus sont répétées et la prédiction est donnée sur la base de l'agrégation des prédictions à partir de n nombre d'arbres. Complexité du train et de l'exécution

$$Tempsd'entraînement = O(\log(nd) * k) \tag{5.5}$$

$$Tempsd'exécution = O(profondeur * k) \tag{5.6}$$

$$Espace = O(stockerchaqueDT * K) \tag{5.7}$$

À mesure que le nombre de modèles de base augmente, le temps d'exécution de la formation augmente, nous utilisons donc toujours la validation croisée pour trouver l'hyperparamètre optimal.

5.4 KNN (K plus proches voisins)

En intelligence artificielle, plus précisément en apprentissage automatique, la méthode des k plus proches voisins est une méthode d'apprentissage supervisé.
La méthode des KNN [4] fait partie des méthodes les plus simples d'apprentissage supervisé pouvant être utilisée pour les cas des classifications.

5.4.1 Complexité algorithmique de KNN

L'algorithme naïf de recherche de voisinage consiste à passer sur l'ensemble des n points de A et à regarder si ce point est plus proche ou non qu'un des plus proches voisins déjà sélectionné, et si oui, l'insérer. On obtient alors un temps de calcul linéaire en la taille de A : O(n) (tant que k « n). Cette méthode est appelée la recherche séquentielle ou recherche linéaire.

5.4.2 Avantages et inconvénients de l'algorithme KNN

Avantages

— L'algorithme est simple et facile à mettre en œuvre.
— Aucune hypothèse sur les données (linéaires, affines).
— L'algorithme est polyvalent. Il peut être utilisé pour la Classification, la régression.

Inconvénients

— Pas efficace pour des jeux de données larges.
— L'estimation de ce modèle devient de mauvaise qualité quand le nombre de variables explicatives est grand.

4. KNN : K plus proches voisins

5.5 Les modèles linéaires généralisés

L'objectif de la modélisation est d'expliquer et prévenir sur un phénomène.
De même l'objectif des modèles linéaires généralisés est d'expliquer une variable à expliquer appelé variable réponse par ces variables explicatives.
Les variations d'une "variable réponse" sont expliquées par des facteurs. Les valeurs susceptibles d'être prises par la variable réponse sont prédites grâce aux variables explicatives. La modélisation est constituée d'un comportement d'une grandeur naturelle qui se compose à la fois d'une partie déterministe et d'une partie aléatoire. La première permet de décrire le comportement moyen et le second constitue le différentiel de la vraie valeur de la variable à la partie déterministe. La modélisation est fondée à cet effet sur deux plans : déterministe qui ajuste la forme mathématique à la variable et du phénomène aléatoire qui attribue une forme de variabilité du phénomène autour de sa moyenne, notamment une forme au hasard.

- On modélise pour un comptage de sinistres par la loi de poisson ou la loi binomiale négative ensuite on choisit le meilleur modèle en utilisant le modèle qui minimise les AIC.
- On modélise les coûts moyens, montant des sinistres par la loi normale ou la loi gamma ensuite on choisit le meilleur modèle en utilisant le modèle qui minimise les AIC.

Notons (Y_i),1 $\leq$ i $\leq$n, l'ensemble des variables aléatoires à expliquer. Afin de pouvoir employer un modèle GLM, il nous faut poser les hypothèses suivantes :

- $(Y_1,..., Y_n)$ définit une famille de variables aléatoires indépendantes qui suivent une distribution appartenant à la famille exponentielle.
- Les prédicteurs $(X1,..., X_n)$ correspondent aux composants déterministes du modèle sous la forme de combinaison linéaire.
- Pour tout i$\in$ 1,. . . ,n, loi de Y_i est supposée appartenir à une famille de distribution dont les paramètres dépendent des variables explicatives à travers une fonction de lien, g strictement monotone.

Fonctions de liens
Les résultats rendus par l'ajustement d'un modèle GLM dépendent de la fonction de lien employée. Lors de cette étude, la fonction de lien utilisée est définie par :
g :]0, 1] $\rightarrow$ R
g(x) = ln(x). Ainsi le modèle sera multiplicatif et s'utilise de la façon suivante, On a alors,

$$\log(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j.x_{(i,j)} \Leftrightarrow E[Y_i] = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j.x_{(i,j)}) \tag{5.8}$$

on a alors,

$$E[Yi] = \exp(\beta_0)\exp(\beta_1 x_{(i,1)})\exp(\beta_p x_{(i,p)}) \tag{5.9}$$

Les coefficients calculés étant tous positifs, cela écarte la possibilité d'avoir une prime pure négative. De plus, le choix de logarithme comme fonction de lien, en générant un modèle multiplicatif, permet de voir facilement l'effet de chaque modalité d'un

critère de tarification sur la prime de référence. Dans ce qui suit on présente la forme générale des distributions exponentielles ainsi que des distributions particulières de cette famille.

Famille exponentielle

Soit Y une variable aléatoire. Alors Y suit une loi de la famille des distributions exponentielles si et seulement si sa densité peut être exprimée sous la forme suivant :

$$f_(\theta, \phi) = \exp(\frac{y\theta - b\phi}{a(\phi)}) + c(y, \phi) \tag{5.10}$$

1. θ le paramètre de la moyenne,
2. ϕ le paramètre de dispersion lié à la variance,
3. a une fonction définie sur R non nulle,
4. b une fonction définie sur R au moins deux fois dérivable, avec une dérivée seconde positive.
5. c une fonction définie sur R2. La moyenne et la variance d'une variable aléatoire dont la densité est de la forme exponentielle sont définies de la façon suivante,

$$E[Y] = b'(\theta) \; et \; Var(Y) = b''(\theta)a(\phi) \tag{5.11}$$

6. Il existe pour chaque loi de la famille exponentielle une fonction de lien qui permet de faire le lien entre l'espérance et le paramètre θ de la loi. Cette fonction est appelée la fonction de lien canonique, notée g_c, et relie l'espérance, usuellement notée μ, au paramètre θ de la manière suivante :

$$\theta = g_c(\mu) = g_c(E[Y]) = g_c(g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_i)) \tag{5.12}$$

Cette égalité intervient dans l'estimation des paramètres du modèle. Quand la fonction de lien est la fonction canonique, le paramètre naturel θ devient donc la combinaison linéaire des variables explicatives,

$$\theta = \beta_0 + \beta_1.x_1 + \ldots + \beta_p.x_p \tag{5.13}$$

Les composantes déterministes seront estimées à partir de la méthode de maximum de vraisemblance.

Estimation des coefficients par la méthode du maximum de vraisemblance

Notons $(y_1, \ldots, y_n)$ un échantillon aléatoire de taille n indépendantes et identiquement distribuées, où leur loi appartient à la famille exponentielle. Alors la fonction de vraisemblance peut s'écrire de la forme suivante :

$$L(\theta, \phi, y_1, ..., y_n) = \prod_{k=1}^{n} [\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(\phi, y_i)] \tag{5.14}$$

avec θ le paramètre canonique inconnu, le paramètre de dispersion ϕ supposé connu

$$E[Y] = b'(\theta) \; et \; Var(Y) = b''(\theta)a(\phi) \tag{5.15}$$

$\beta = (\beta_1, \ldots, \beta_p)$, ainsi l'estimation des coefficients du GLM se fait en cherchant les β qui maximisent la vraisemblance, c'est-à-dire qu'ils vérifient les conditions suivantes :

$$\frac{\partial L(\theta, \phi, y_1, ..., y_n)}{\partial \beta} = 0 \; et \; \frac{\partial L(\theta, \phi, y_1, ..., y_n)}{\partial \beta^2} < 0 \tag{5.16}$$

Pour la plupart des modèles linéaires généralisés, les équations qui déterminent les paramètres au sens du maximum de vraisemblance sont non linéaires. En pratique, il faut recourir à des méthodes itératives pour maximiser la fonction de vraisemblance. L'algorithme de Newton-Raphson est la méthode de résolution itérative la plus courante pour estimer les prédicteurs du GLM. L'algorithme approxime le logarithme de la fonction de vraisemblance dans un voisinage du paramètre initial par une fonction polynomiale qui a la forme d'une parabole concave. Elle a la même pente et la même courbure dans les conditions initiales que la log-fonction de vraisemblance. Il est facile de déterminer le maximum de ce polynôme d'approximation. Ce maximum fournit la seconde étape du processus d'estimation et l'on reprend la procédure décrite précédemment. Les approximations successives convergent rapidement vers les estimations au sens du maximum de vraisemblance.

Distributions

En probabilité et en statistiques, les distributions Tweedie appartiennent à la classe des modèles de dispersion exponentielle, célèbres pour leur rôle dans les modèles linéaires généralisés. C'est une famille de distribution de probabilité qui comprend des distributions continues telles que la distribution Normale et Gamma et des distributions discrètes comme la distribution de Poisson. Nous allons juste présenter la loi de Poisson et la loi binomiale négative.

La loi de Poisson

La loi de Poisson est une loi discrète dépendant d'un paramètre d'intensité, un nombre réel noté Λ. Soit Y une variable aléatoire suivant une loi de poisson de paramètre réel positif $\Lambda, \Lambda -> P(\Lambda)$.

La loi de Y est alors : $\forall k \in N, P(y = k) = \frac{exp(-\Lambda)\Lambda^k}{k!}$

La loi Poisson a la particularité d'avoir une moyenne et une variance qui sont égales $E[Y] = Var(Y) = \Lambda$

Troisième partie

Modélisation de la fréquence des sinistres par la méthode des Glm

Chapitre 6

Modélisation des postes de garanties par la méthode des Glm

Les modèles linéaires généralisés nous donnent des modèles interprétables et permettent de comparer facilement les modèles entre eux.
Le but est d'expliquer le nombre de sinistres ou montant appelé variable réponse par les variables explicatives, ensuite sélectionner les variables significatives et définir un modèle optimal ou toutes les variables expliquent le nombre de sinistres.
Dans cette partie du mémoire, nous allons appliquer des modèles linéaires généralisés aux données de sinistralité des postes de garanties pour avoir la fréquence. A l'issu de cette modélisation, nous obtenons des résidus qui seront ensuite expliqués par des variables externes. Finalement, nous arriverons aux regroupements des résidus modélisés pour construire le zonier fréquence.

6.1 Dispersion variable

Pour modéliser la fréquence des sinistres, nous utilisons souvent pour ce type de modélisation des lois de comptage. En assurance, les lois de comptage communément utilisées sont la loi de poisson et la loi binomiale négative.

- La loi de Poisson : l'hypothèse principale est la forme très particulière de la loi de probabilité, pour la loi de Poisson la moyenne est égale à la variance ce qui veut dire que nos données sont équi-dispersées.
- En ce qui concerne la loi binomiale négative la variance est supérieure à la moyenne, on dit que les données sont sur-dispersées.

Nous allons supposer l'hypothèse d'équi-dispersion de nos données puis nous utiliserons le critère basé sur le rapport variance/moyenne pour faire une première sélection de loi. Un rapport supérieur à 1 par exemple soupçonne l'existence d'une sur-dispersion, donc induisant à écarter immédiatement la possibilité que la fréquence suit une loi de Poisson.

Ce tableau ci-dessous montre la variance et la moyenne par catégories d'âges :

Âge	**Moyenne**	**Variance**
Adulte	8,77	28,94
Enfant	7,94	32,83
Mineur	8,02	32,38
Senior	9,48	24,51

TABLE 6.1 – Moyenne et variance par catégorie

On rejette l'hypothèse d'équi-dispersion des données. Nos données sont sur-dispersées. En effet, la variance est supérieure à la moyenne. La sur-dispersion renvoie à la loi binomiale négative, on devrait calibrer nos données par la loi binomiale négative. Pour valider notre hypothèse, nous allons voir un peu plus loin dans ce mémoire, un calibrage de la loi en tenant compte de l'ensemble des informations apportées par les variables explicatives. Nous allons lancer pour la loi proposée un modèle avec toutes les variables explicatives. Ensuite, nous allons sélectionner les variables pertinentes via la procédure stepwise basé sur le critère AIC.

6.2 Modèle fréquence $*$ coût moyen

L'une des hypothèses du modèle fréquence $*$ coût moyen est l'indépendance entre fréquence et le coût moyen, c'est à dire nous pourrons modéliser séparément la fréquence des sinistres et le coût moyen. Pour valider cette hypothèse nous utilisons les copules de dépendance et les mesures de dépendance telles que le coefficient de corrélation de Pearson, Spearman ou de Kendall pour montrer que la fréquence et le coût moyen sont indépendants. Nous comparons trois choses :

- la distribution fréquence, coût moyen et copule d'indépendance par poste de garanties.
- les lignes de niveau fréquence $*$ coût moyen ainsi que les lignes de niveau de la copule indépendance par poste de garanties.
- les coefficients de corrélation entre la fréquence et le coût moyen par poste de garanties.

Premièrement nous allons tenter de valider l'hypothèse d'indépendance entre la fréquence et le coût moyen en utilisant des copules.

Copule indépendance

Les copules sont des outils qui permettent de représenter la dépendance entre des variables aléatoires de façon réaliste. Dans notre cas nous allons utiliser la copule bivariée pour modéliser la dépendance fréquence coût moyen.

Une copule est la fonction de répartition d'un vecteur de variables aléatoires U = (U_1,U_2) dont toutes les composantes U_i (i = 1, 2) obéissent à la loi uniforme (0,1).

Une copule de dimension 2, C(U_1,U_2) est une fonction de $[0,1] * [0,1]] -> [0,1]$ telle que :

— Marginales uniformes pour tout u et v dans [0, 1],
C(U_1,0)=0, C(U_1,0)= U_1 et C(0,U_2)= U_2

— Croissante quel que soit U_1,U_2, U_3 dans [0, 1], tels que $U_1 < U_2$ et $V_1 < V_2$

$$C(U_2, V_1) - C(U_2, V_2) - (C(U_1, V_1) - C(U_1, V_2)) > 0 \tag{6.1}$$

Ce graphique ci-dessous traduit la représentation de la copule estimée et la copule indépendance.

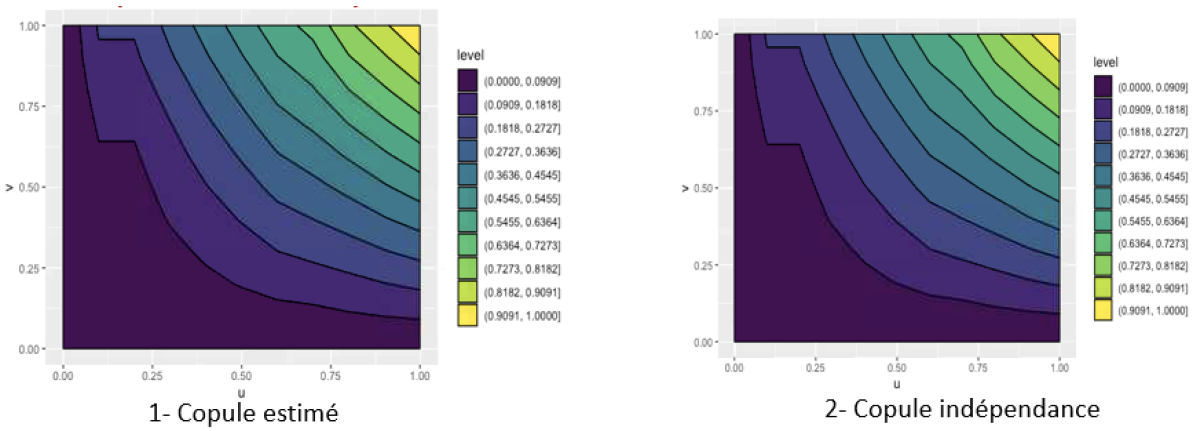


FIGURE 6.1 – Représentation copule Indépendance et copule estimée

La copule estimée sur la fréquence et le coût moyen a la même distribution que la copule d'indépendance, les lignes de niveau sont également les mêmes que celui de la copule d'indépendance. De manière analogue nous avons comparé la copule indépendance avec une copule empirique estimée pour chaque poste de garanties. On obtient le même graphique que la copule d'indépendance.

Les tableaux suivants récapitulent les coefficients de corrélation de Pearson, Spearman et Kendall :

Poste de garantie	Spearman	Kendall	Pearson
HOSPITALISATION	0,159	0,112	0,069
SOINS COURANTS	-0,002	-0,001	-0,015
OPTIQUE	-0,041	-0,030	-0,024
CURES THERMALES	0,015	0,012	-0,043
PHARMACIE	0,022	0,033	-0,005
AUTRES	-0,034	-0,028	-0,002
DENTAIRE	0,033	0,022	0,029
PREV ET BIEN ÊTRES	-0,069	-0,044	-0,053
AUTRES PROTHÈSES	0,075	0,053	0,018

TABLE 6.2 – Corrélation de Pearson, Spearman et Kendall

Nous constatons une faible corrélation entre la fréquence et le coût moyen par poste. Nous pouvons conclure ici que la fréquence et le coût moyen sont indépendants. L'hypothèse d'indépendance étant validée, nous pouvons modéliser séparément la fréquence et le coût moyen.

6.3 Étude d'indépendance entre les variables explicatives

Une hypothèse fondamentale à la convergence des modèles linéaires généralisés en particulier et les modèles économétriques au global est l'absence de la colinéarité entre les variables explicatives. En effet, cela impact la variance des modèles, l'existence d'une colinéarité entre les variables explicatives augmente la variance du modèle.

V de cramer

Nous utilisons le V de cramer pour calculer le coefficient de corrélation entre nos variables il se base sur le test d'indépendance du chi 2 entre les variables. Le test du chi 2 permet d'évaluer l'indépendance entre deux variables qualitatives la relation qui existe entre les variables qualitatives.

Le V de cramer se définit comme suit : L'indice V de Cramer

$$V = \frac{D^2}{(N.inf((r-1);(c-1)))^1/2} \tag{6.2}$$

— $0 \leq V \leq 1$

— N : effectif total (N est la somme des cellules de la table contingence formée par les variables étudiées).

— r : nombre de lignes de la table contingence.

— c : le nombre de colonnes de la contingence

— D^2 est en fonction de $E_(i,j)$ ou $E_(i,j)$ est la fréquence théorique en cas d'indépendance de deux variables $O_(i,j)$ et fréquence observée

L'analyse de corrélation est souvent utilisée en statistiques descriptives. Elle est très utile pour mesurer la corrélation entre deux variables.
L'objectif est de voir la relation de dépendance deux à deux entre les variables. Dans le graphique généré par un corrélogramme, les corrélations négatives sont en rouge foncée et les corrélations positives en bleue foncée. L'intensité de la couleur est proportionnelle au coefficient de corrélation, donc plus la corrélation est forte (c'est-à-dire plus proche de -1 ou 1), plus les cases sont foncées. La légende des couleurs sur le côté droit du corrélogramme montre les coefficients de corrélation et les couleurs correspondantes.

Le V de cramer appliqué à nos variables tarifaires nous permet d'avoir le corrélogramme suivant :

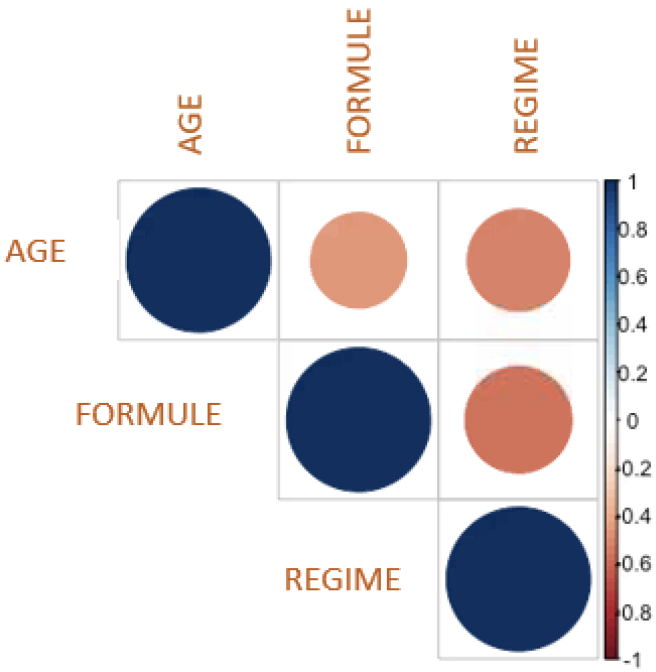


FIGURE 6.2 – Corrélogramme des variables tarifaires

Les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation. A droite du corrélogramme, la légende de couleurs montre les coefficients de corrélation et les couleurs correspondantes. Le corrélogramme nous montre qu'il existe une faible corrélation entre les variables tarifaires. Par conséquent, on peut garder toutes les variables tarifaires dans la modélisation. Il est important dans un modèle de supprimer les variables fortement corrélées. La forte corrélation des variables explicatives peut entraîner un problème d'estimation des coefficients et l'interprétation d'un modèle.

6.4 Choix des postes à modéliser

En santé, les principaux postes de soins sont les suivants : Hospitalisation, Soins courants, Pharmacie, Dentaire et optique. Ils présentent des couvertures différentes et, en conséquence, des niveaux de consommation non homogènes. Il est donc important de les modéliser séparément pour avoir une meilleure modélisation.
Pour calibrer un GLM, il faut avoir suffisamment de données par poste de garanties, certains postes n'avaient pas beaucoup de données pour calibrer un GLM. Par conséquent nous avons décidé de ne pas faire de modèle GLM sur les postes qui ont une faible consommation. Nous avons exclu de la modélisation tous les postes dont la fréquence et les règlements représentent moins de 2% de la consommation totale.

Tableau proportion des sinistres par postes de garanties :

POSTE DE GARANTIES	**Nombre sinistres**	**Règlement**	**Fréquence**
AUTRES PROTHÈSES	6,31%	6.6%	6,31%
CURES THERMALES	0,04%	0,02%	0,4%
DENTAIRE	2,52%	9,7%	2,52%
HOSPITALISATION	2,43%	19,5%	2,43%
OPTIQUE	1,69%	8,7%	1,69%
PHARMACIE	41,78%	26,7%	41,78%
PRÉV ET BIEN ÊTRE	1,02%	1%	1,02%
SOINS COURANTS	44,20%	27,6%	44,20%

TABLE 6.3 – Tableau proportion des sinistres

Nous constatons que la fréquence des sinistres est portée principalement par la pharmacie et les soins courants. Ils représentent successivement 41,78% et 44,20% de la fréquence totale des sinistres, cependant l'hospitalisation a une faible fréquence, mais représente de 19,5% de la charge totale. Finalement après analyses de ces indicateurs, nous avons retenu les postes suivants pour la modélisation : hospitalisation, soins courants, dentaire, optique, pharmacie et autres prothèses et appareillages.

6.5 Modélisation de la fréquence des sinistres

Le but du GLM est d'expliquer la fréquence des sinistres par les variables explicatives ou tarifaires. Nous voulons expliquer la fréquence des sinistres donc nous allons calibrer notre modèle avec le GLM [1] sur la fréquence de poisson ou le GLM binomial négatif. On obtient :

— une idée sur la loi des déviances résiduelles

— l'estimation des coefficients

1. GLM : Modèle linéaire généralisé

— la déviance du modèle estimé et du modèle nul

— le critère AIC [2]

Le niveau de consommation étant différent par poste, nous avons calibré un modèle par poste de garanties.
Nous avons utilisé la méthode de cross validation pour valider notre modèle.
La méthode de cross validation consiste à faire trois tirages aléatoires sans remise sur notre base de données :

1. le premier tiers des données pour la base entraînement.
2. le second tiers des données pour base de validation.
3. le dernier tiers données pour base test qui est aussi la base d'entraînement krigeage.

1. Sur la base d'entraînement nous avons réparti la base par poste de garanties, nous obtenons une base d'entraînement par poste de garanties. Ensuite nous construisons un modèle sur l'ensemble des données d'entraînement réparties par poste de garanties. On obtient :
   — Une base d'entraînement Hospitalisation qui sera utilisée pour la création du modèle hospitalisation.
   — Une base de données d'entraînement pour les soins courants qui sera utilisée pour créer le modèle pour les soins courants. De manière analogue nous avons créé les modèles pour le reste des postes de garanties optique, pharmacie etc.... Sur la base de validation nous avons scindé une base par poste garantie également.
2. Les modèles étant créés, nous appliquons ces modèles à l'ensemble de données de validation pour prédire le résultat de nouvelles observations. Cette étape permet de tester l'efficacité du modèle sur l'échantillon. Nous validons les modèles créés à partir des bases d'entraînement sur les bases de validations. Nous avons créé une base de validation par poste de garanties :
   — Une base de validation Hospitalisation cette base de données sera utilisée pour la validation du modèle hospitalisation
   — Une base de validation pour les soins courants qui sera utilisée pour valider le modèle pour les soins courants. De manière analogue, nous avons créé les bases de validation pour le reste des postes de garanties : optique, pharmacie etc.... Cette étape de validation des données nous permet également de quantifier l'erreur de prédiction en utilisant des métriques de performance d'un modèle comme la MSE qui est l'erreur quadratique moyenne et MAE qui peut être défini comme étant l'erreur absolue moyenne. Ces métriques seront utiles pour le choix du meilleur modèle. Les modèles de poissons et les modèles binomiaux étant créés nous allons passer à l'étape de choix du meilleur modèles entre ces derniers.
3. La base d'entraînement krigeage est la base qui sera utilisée pour lisser les résidus.

2. AIC :Le critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique.

6.5.1 Comparaison modèle

1. **Critère AIC**
   Pour comparer nos modèles, nous allons nous baser sur le critère AIC. Le critère d'information AIC s'applique aux modèles estimés par une méthode du maximum de vraisemblance. Le meilleur modèle est celui qui minimise le AIC. Dans notre cas de figure les modèles binomiaux négatifs présentent les AIC les plus faibles, par conséquent ils sont considérés comme étant nos meilleurs modèles.
   Le tableau ci-dessous montre les AIC obtenus par postes de garanties.

Poste de garanties	Binomiale	Poisson
HOSPITALISATION	150 633	310 185
OPTIQUE	143 287	183 441
PHARMACIE	157 977	290 895
SOINS COURANT	460 052	1 460 361
DENTAIRE	510 649	1 500 746

Table 6.4 – Comparaison AIC

   Nous constatons que les modèles GLM binomiaux minimisent les AIC.

2. **Indicateur MSE et MAE**
   Les tableaux ci-dessous récapitulent les MSE obtenus :

Poste de garanties	MSE_train Poisson	MSE_train Binomiale
HOSPITALISATION	7,12	2,84
OPTIQUE	2,89	1,93
DENTAIRE	5,15	2,39

Table 6.5 – Comparaison MSE base d'entraînement

Poste de garantie	MSE_val Poisson	MSE_val Binomiale
HOSPITALISATION	6,98	2,28
OPTIQUE	2,69	2,83
DENTAIRE	4,82	1,83

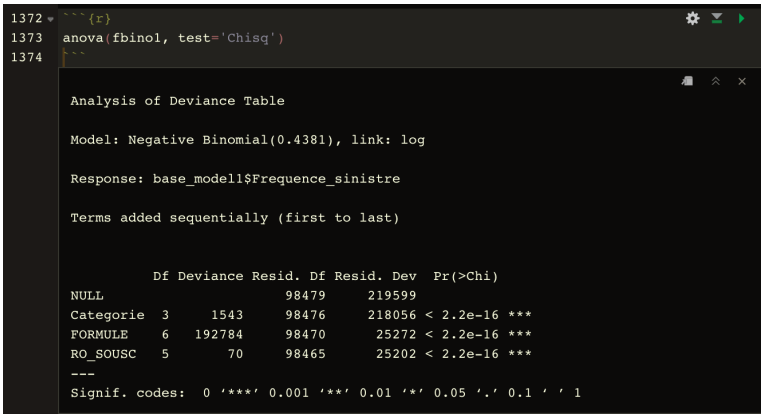
Table 6.6 – Comparaison MSE base de validation

   Nous constatons que les modèles GLM binomiaux minimisent MSE et MAE. Lors de la comparaison entre deux modèles, le meilleur modèle est celui qui minimise le MSE et le MAE. Les modèles GLM binomiaux sont les meilleurs modèles.

Les modèles GLM binomiaux minimisent les AIC de plus ils minimisent le MSE et le MAE.
On conclut que les modèles GLM binomiaux sont nos meilleurs modèles.
Dans la suite de ce mémoire nous allons utiliser les modèles GLM binomiaux négatifs.

6.5.2 Sélection des variables

On va tester la significativité de nos variables pour ensuite utiliser la méthode descendante. Dans ce modèle, nous prenons le modèle où toutes les variables sont significatives. Le but de cette méthode est de voir si, en retirant certaines variables, cela améliore la qualité du modèle. Pour ce faire, on peut utiliser la fonction Anova pour la sélection des variables.



```
1372 ```{r}
1373 anova(fbino1, test='Chisq')
1374 ```

Analysis of Deviance Table

Model: Negative Binomial(0.4381), link: log

Response: base_model1$Frequence_sinistre

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     98479     219599
Categorie  3     1543     98476     218056 < 2.2e-16 ***
FORMULE    6   192784     98470      25272 < 2.2e-16 ***
RO_SOUSC   5       70     98465      25202 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 6.3 – Sortie Anova

Après l'étude de la sortie d'Anova, on déduit que toutes les variables sont significatives. Par conséquent, on conserve pour ce modèle toutes les variables explicatives. Nous constatons que toutes les variables sont significatives pour tous les modèles que nous avons créés.

6.5.3 Validation du modèle

Pour finir nous voulons valider le modèle. Pour ce faire, nous allons nous baser sur une validation par étude de la déviance (on va comparer le log-vraisemblance du modèle estimé avec celle du modèle parfaitement ajusté dit saturé). La déviance est définie par :

$$Delta = 2 * (logv(estimé) - logv(saturé)). \tag{6.3}$$

Le modèle est adéquat si la déviance est faible.
On pose H_0 : "le modèle est adéquat" et H_1 : "le modèle n'est pas adéquat".
On rejette donc H_0 si Delta est plus grande qu'un chi^2_{n-p}.
Donc on calcule la p-value :

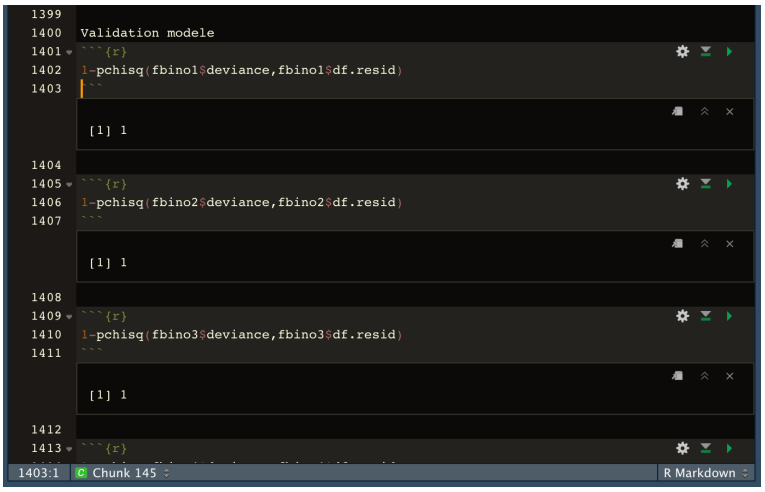


FIGURE 6.4 – Sortie validation modèle

On ne rejette pas l'hypothèse nulle que le modèle ajuste correctement les données on accepte l'hypothèse H_0 selon laquelle le modèle est adéquat.
Les modèles binomiaux ajustent correctement nos données.

Quatrième partie

Mise en place du zonier

Chapitre 7

Mise en place du zonier

Dans la section précédente, nous avons modélisé la fréquence des sinistres. Ce qui nous a permis d'avoir des résidus par département. Cependant, nous n'avons pas de résidus dans tous les départements de la France. Cette partie est le cœur de notre étude, l'objectif dans un premier temps est de lisser les résidus par la méthode de Krigeage afin de lisser les résidus sur tous les départements de la France.
Ensuite, nous pourrons ajouter les variables externes à notre modèle en utilisant des méthodes de Machine Learning, notamment le Random forest et enfin, nous pourrons classifier nos résidus par département en utilisant les KKN et la méthode des quantiles. On les classe du risque le moins élevé au risque le plus élevé.

7.1 Agrégation des résidus

Dans l'étape précédente, nous avons modélisé nos postes garanties par des GLM. Ce qui nous a permis de calculer les résidus issus de la prédiction :

$$Résidus = Fréquence\ observée - Fréquence\ prédite \tag{7.1}$$

Nous obtenons des résidus pour chaque poste de garanties puis nous cumulons les résidus pour avoir des résidus totaux que nous allons ensuite agréger à la maille département. Ils seront ensuite lissés par krigeage. L'objectif de ce dernier est de réduire la forte variabilité entre deux départements proches.
Le krigeage permet aussi de faire l'interpolation des résidus aux départements qui n'étaient pas représentés dans les données. Après le lissage, nous aurons une répartition homogène des résidus sur tous les départements de France.

Représentation des résidus par département. Sur la figure suivante, nous avons la carte qui représente les résidus par département sur toute la France. Les zones à couleur foncée sont les départements où l'absence d'information géographique est très marquée.

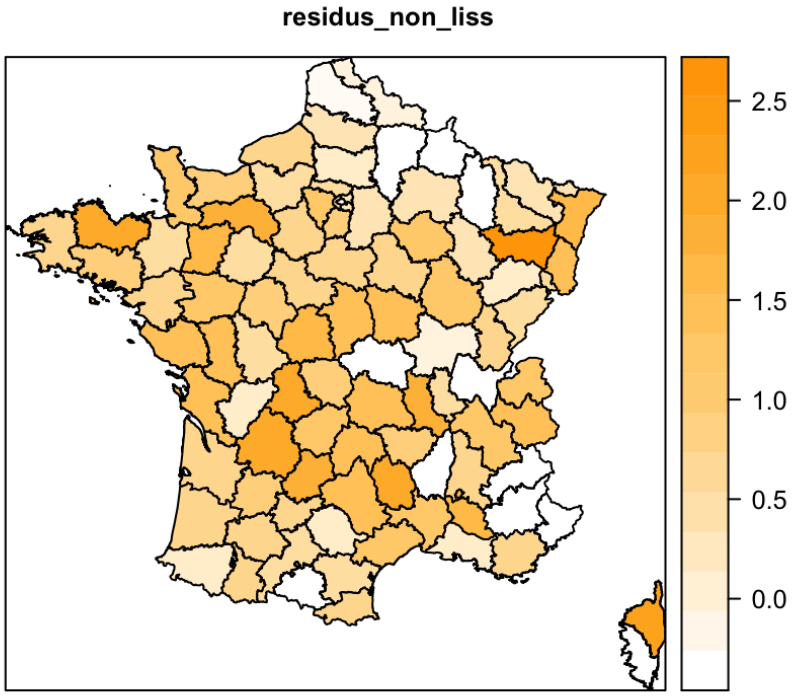


FIGURE 7.1 – Représentation des résidus non lissés

La carte illustre aussi le fait que les résidus non lissées sont non homogènes par département, d'où la nécessité de faire un lissage avant de modéliser ces résidus. Ce qui va nous permettre d'interpoler les départements non observés dans notre portefeuille.

7.2 Lissage : krigeage

Le krigeage a été développé par Danie Gerhardus Krige pour la prospection minière, initialement il était conçu pour le forage de puits pétroliers.
Depuis, ses applications ont largement été étendues dans plusieurs domaines notamment, dans la mise en place d'un zonier dans le monde de l'assurance.
Le krigeage peut être défini comme étant la capacité de prédire une variable cible en l'occurrence les résidus, à partir des résidus des autres départements. En cas d'absence d'un département dans notre portefeuille, le krigeage permet de lisser les résidus de cette zone en prenant en compte toutes les caractéristiques des zones voisines.
Le krigeage est un prédicteur linéaire. Il est considéré comme un BLP : Best Linear Predictor, c'est-à-dire le meilleur prédicteur linéaire entre les observations non biaisées. Il tient compte non seulement de la distance entre les résidus mais également du caractère de dépendance géo-spatiale entre les résidus.

Le krigeage permet de faire une modélisation d'une donnée non observée par moyennisation pondérée et interpolée des caractéristiques des résidus observés : si un département n'est pas présent dans notre portefeuille, il permet de prédire ou de lisser les résidus de ce département à partir des caractéristiques des autres départements aux alentours, c'est l'avantage du krigeage.

$$r_i^* = \sum r_j w^* \tag{7.2}$$

— r_i^* résidus prédits

— r_j résidus des départements aux alentours

— w^* prend en compte les caractéristiques des départements aux alentours.

La théorie krigeage prend en compte la structure et le caractère de dépendance géospatiale des résidus. Il apporte en plus un cadre mathématique puissant permettant l'étude et l'analyse de la robustesse d'un tel lissage.
Il existe deux types de krigeage : le **krigeage simple** et le **krigeage ordinaire.**
Le krigeage simple :
il s'agit d'un modèle de krigeage global qui ne prend pas en compte les variations locales de la tendance déterministe : on considère dans ce modèle que les variations restent constantes sur l'ensemble de la France.
Le krigeage ordinaire :
le krigeage ordinaire considère que la tendance est constante mais seulement par morceau. Autrement dit, la tendance est constante au niveau d'un voisinage et non plus sur l'ensemble de la France. Dans ce mémoire, nous allons utiliser le krigeage ordinaire.
Autocorrélation spatiale :
L'Auto-corrélation spatiale permet de mesurer la dépendance géospatiale des résidus. En présence d'auto-corrélation spatiale, on observe que la valeur d'une variable pour une observation est liée aux valeurs de cette même variable pour les observations voisines. L'analyse de l'auto-corrélation spatiale des résidus nous permet de faire une analyse quantifiée de la structure spatiale du risque de sinistralité non captée par les variables tarifaires utilises dansé le GLM.
Semi-variogramme
Le semi-variogramme est un outil qui permet alors de décrire la corrélation spatiale entre observations spatiales.

La carte de la figure suivante représente les résidus lissés par krigeage.

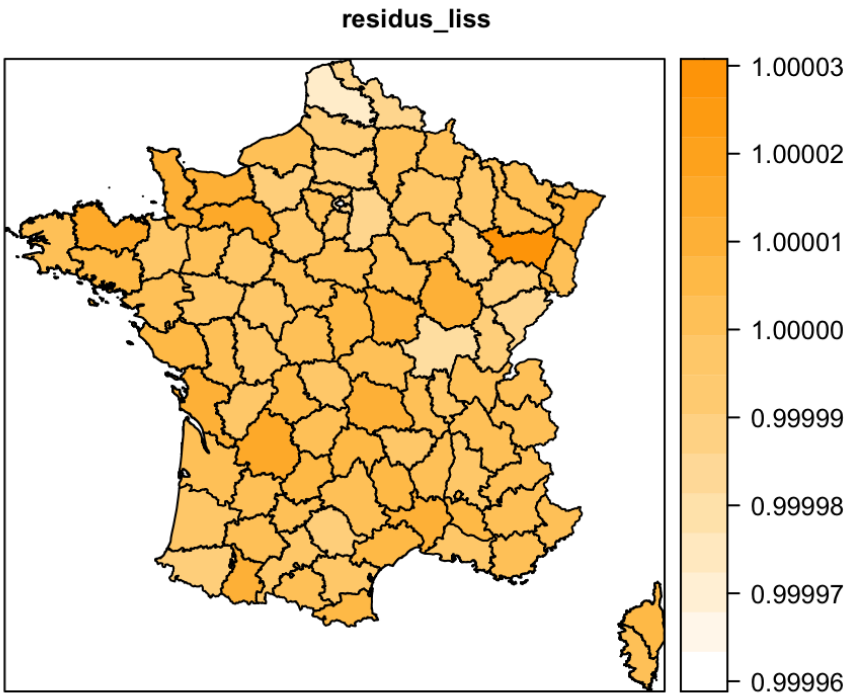


FIGURE 7.2 – Représentation des résidus lissés

On a une répartition plus homogène des résidus après le lissage par krigeage. Nous n'observons plus de très fortes variations entre départements proches.

7.3 Variables externes

Les résidus étant lissés dans la section précédente, dans cette partie, nous allons modéliser ces résidus et étudier l'importance des variables externes en utilisant les méthodes de Machines Learning qui sont des modèles plus adaptés pour faire ce type de modélisation. Nous allons utiliser le Random Forest qui est particulièrement efficace et souvent utilisé pour ce type de modélisation. Le Random Forest va nous permettre de classer les variables explicatives en fonction de leurs liens avec la variable à expliquer, c'est un modèle robuste, rapide et simple à utiliser. Il utilise des arbres de classification, ce qui permet d'obtenir une prédiction fiable, grâce à son système de forêt d'arbres décisionnels. Ainsi nous pourrons directement voir le classement des variables qui sont significatives à notre modèle.

Sous R nous avons implémenté un modèle de Random forest sur nos résidus en utilsant comme variables explicatives les variables externes. Le graphique ci-dessous nous permet de voir l'importance des variables explicatives externes. Le Random forest appliqué à nos données nous permet d'avoir le classement décroissant des variables externes définies dans la section 3.8 au sens de leur influence.

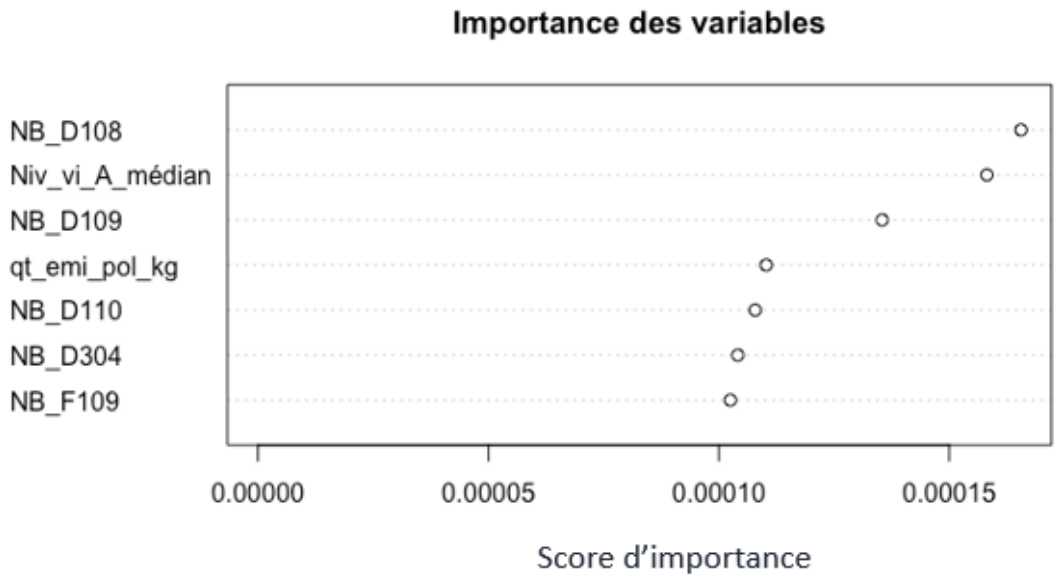


FIGURE 7.3 – Score d'importance des variables externes

Nous constatons que le nombre de centres de santé, le niveau de vie médian et le nombre de structures psychiatriques semblent influencer sur la fréquence des sinistres par département, en effet ils constituent les variables les plus importantes parmi les variables externes.
La pollution et le nombre de parcours santé semblent avoir un impact sur la fréquence des sinistres en santé : les assurés vivant dans les zones plus polluées ont tendance à avoir une santé plus fragile tandis que les assurés qui habitent dans des départements ayant plus de parcours santé ont une meilleure qualité de vie. Une interprétation possible serait : de tels équipements de loisirs jouent un rôle contre le stress et d'autres pathologies, ce qui pourrait diminuer la fréquence des sinistres.
Enfin, on retrouve le nombre de centres de médecine préventive et le nombre de transfusions sanguines par département qui sont des variables qui semblent impacter également la fréquence des sinistres par département. A partir de ces variables externes qui ont une grande importance sur notre modèle, nous recalculons les résidus.
Nous avons enlevé la part de résidus expliquée par ces variables externes puis nous avons classifié les résidus par zone de risque en utilisant les KNN et la méthode des quantiles.
Étant donné que le zonier est une variable tarifaire, il est nécessaire d'avoir des zones homogènes, en effet entre deux départements proches l'écart de tarif ne doit pas être élevé.

7.4 Classification zone

Une étape préalable au découpage s'agissait d'une agrégation des résidus par département pour ensuite les classer. A ce stade, nous devons choisir le nombre de classes optimal à retenir pour le regroupement des zones en fonction de deux métriques : la méthode du coude et la méthode silhouette pour déterminer le nombre de clusters optimal adapté pour nos données.

Méthode du coude (Elbow method) : La méthode consiste à faire une représentation graphique de la variance expliquée (dispersion) de nos données en fonction du nombre de clusters k qui est choisi là où la cassure est la plus flagrante.

La première figure ci dessous est le résultat de l'implémentation de la méthode du coude, qui s'appuie sur la notion d'inertie intra-classe (within-cluster inertia), définie comme la somme des distances euclidiennes entre chaque point et son centroïde associé. Elle mesure l'hétérogénéité au sein des classes, ce critère de sélection du nombre de clusters doit donc être minimisé. Or, nous pouvons constater sur le graphique que plus le nombre de clusters est élevé, plus l'inertie intra-classe diminue. Toutefois, la tendance paraît ralentir au-delà de 4 : la réduction liée à l'ajout d'un cluster supplémentaire n'est plus aussi significative qu'au début. 4 semblait donc être un bon choix pour le découpage.

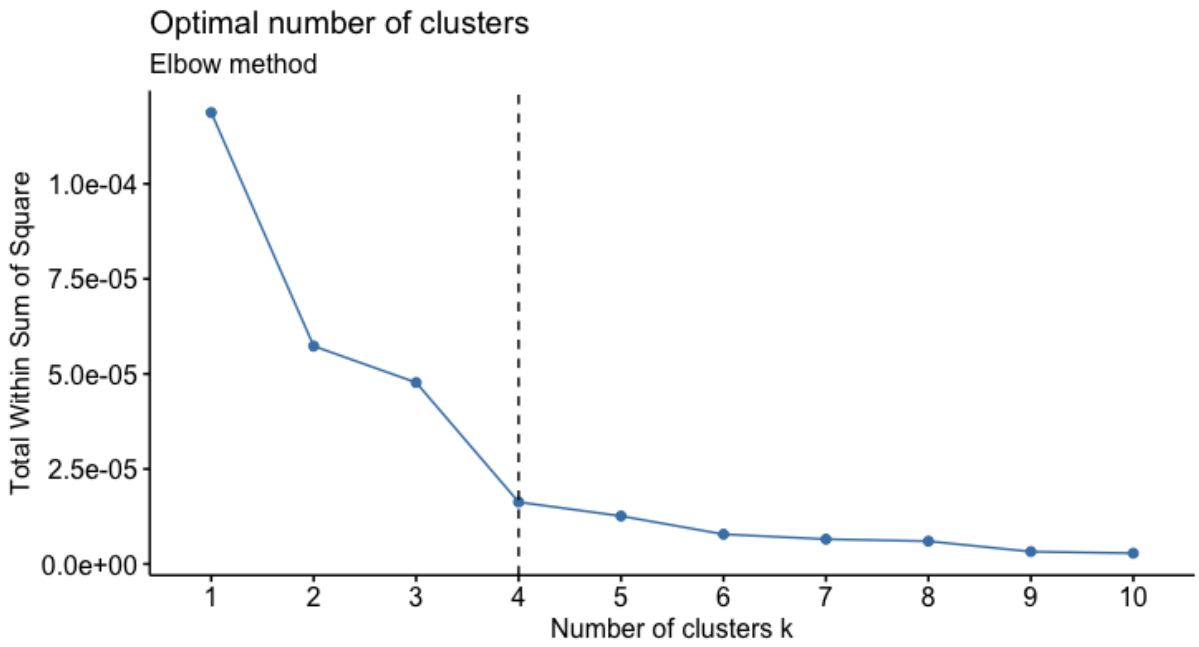


FIGURE 7.4 – Elbow method

Méthode de la silhouette : Afin de challenger les résultats obtenus avec l'inertie intra-classe et la méthode du coude, nous avons testé une approche alternative : l'indice silhouette. La plage de valeur de ce coefficient varie entre -1 et 1. Il rend compte de la différence entre la distance moyenne d'un point donné aux points du même groupe que lui (cohésion) et la distance moyenne de ce point aux points du groupe voisin (séparation). Une valeur proche de 1 signifie que la distance qui le sépare de la classe la plus proche est bien supérieure à celle qui le sépare de sa classe. Le point est donc

bien classé. A l'inverse, une valeur qui tend vers -1 indique que le point n'a pas été correctement classifié, puisqu'il est en moyenne plus proche de la grappe voisine que de son propre cluster. L'indice Silhouette de la partition est calculé à partir de la moyenne des indices de ses éléments.
La méthode de la silhouette appliquée à nos données nous donne le graphique suivant :

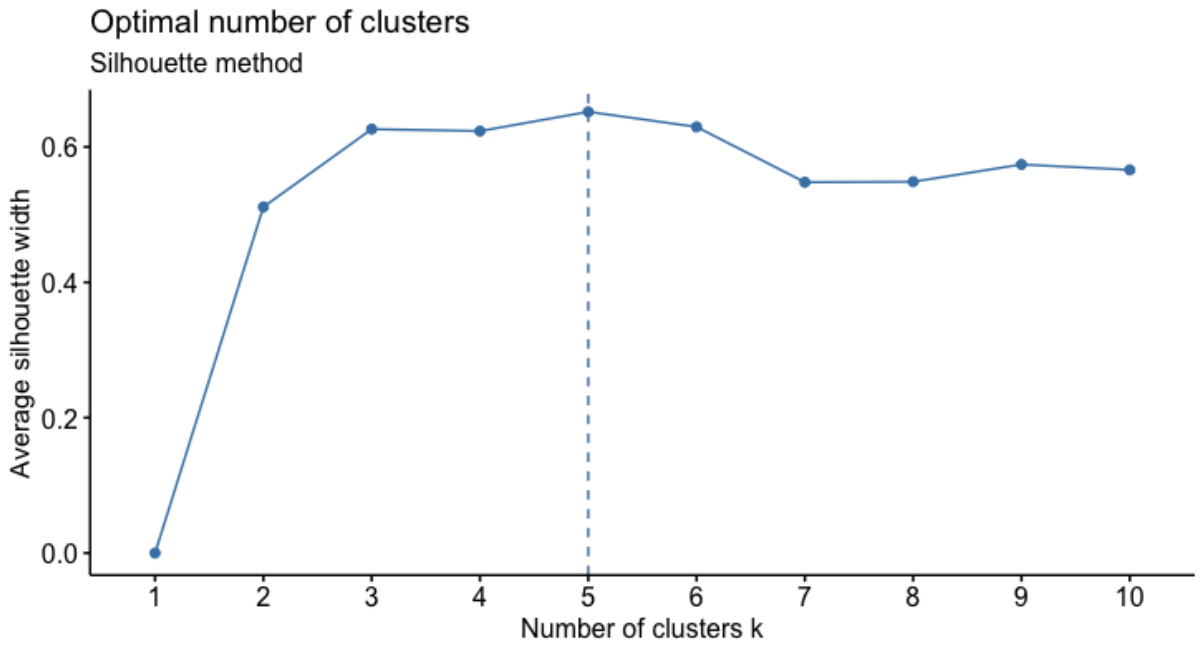


FIGURE 7.5 – Silhouette méthode

La méthode du KNN nous permet d'avoir le nombre de classes optimale pour nos données. Cependant une des limites de cette approche est le déséquilibre des effectifs par classe.
Pour pallier à ce problème, nous allons utiliser le nombre de classe optimale proposée par les KNN et classifier nos résidus avec la méthode des quantiles qui nous permet d'avoir une bonne répartition des effectifs par classe.
On les classe du risque le moins élevé au risque le plus élevé : la classe 1 étant la Zone la moins risquée et la classe 5 étant la zone la plus risquée.
Ces différentes approches nous permettent de voir que le nombre de clusters optimal est de 5.

7.5 Comparaison zonier

Dans cette partie je compare le zonier élaboré avec le zonier déjà existant et le zonier des Agents Généraux afin de voir le meilleur zonier.

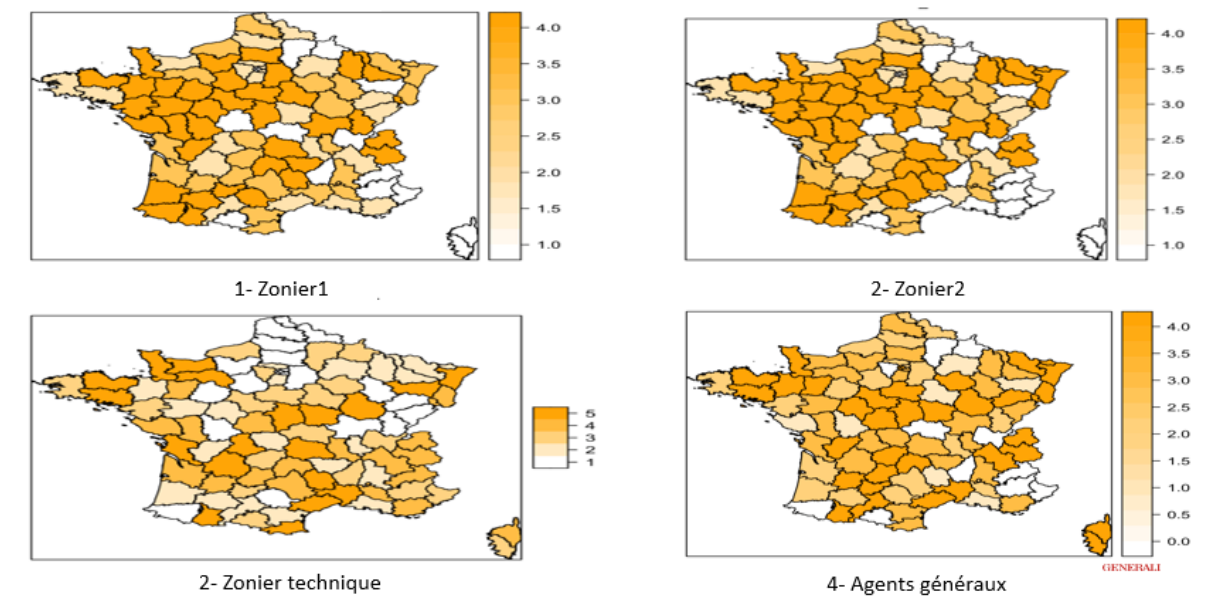


FIGURE 7.6 – Comparaison zonier

- La cartographie (figure 2) de notre zonier nous permet de voir que nous avons une meilleure répartition du facteur de risque géographique, nous avons une répartition plus homogène par zone de risque. Notre zonier prend en compte le risque géographique tandis que le zonier du partenaire est un zonier commercial et il n'est pas discriminant. En effet, certains départements comme le sud et de la France ont été classifiés dans des zones risquées contrairement au zonier du partenaire qui les classe dans des zones non risquées pour des raisons commerciales.
  Ainsi nous pouvons dire que notre zonier technique prend bien en compte le facteur de risque géographique.
- Comme nous l'avons vu précédemment, le distributeur cible majoritairement les seniors du régime général. Nous constatons que dans le sud de la France la population des seniors est bien représentée, mais aussi cette zone géographique est caractérisée par la présence des offres concurrentielles importantes. Pour être à la fois attractif et compétitif, le distributeur diminue son tarif et propose des tarifs commerciaux.
- Néanmoins, nous notons que notre partenaire dispose d'un portefeuille globalement rentable. Il se caractérise par une mutualisation du risque géographique entre les zones sous-tarifées et les zones sur-tarifées. Il comble une sous-tarification dans les zones où il veut être présent par la sur-tarification dans les autres zones.
- Le zonier des Agents Généraux (hors partenariats)(figure 4) est également un zonier commercial. Cependant, nous constatons les mêmes classifications du risque

géographique dans certains départements. On note par exemple certains départements du sud-ouest et quelques départements du sud-est notamment Marseille, communs au zonier technique

Comparaison entre le Glm avec ou sans zonier

Nous allons pousser plus loin notre étude, en comparant le modèle initial de fréquence au nouveau modèle linéaire généralisé qui prend en compte la répartition géographique, l'objectif c'est de comparer le GLM sans zonier et le GLM avec le nouveau zonier en utilisant le critère AIC qui est un critère purement statistique.

Dans le tableau suivant, nous retrouvons les résultats des performances selon le critère de l'AIC de la modélisation par poste de garantie après avoir intégré le zonier comme variable tarifaire.

Poste de garanties	**Sans zonier**	**Avec zonier**
HOSPITALISATION	150 633	141 168
OPTIQUE	143 287	102 847
PHARMACIE	157 977	103 543
SOINS COURANT	460 052	134 755
DENTAIRE	510 649	137 869

TABLE 7.1 – Comparaison AIC avec ou sans zonier

Nous constatons que les modèles binomiaux dans le cadre du nouveau zonier minimisent le critère AIC. Nous constatons également que l'écart est très significatif sur les soins courants et le dentaire, le zonier est efficace sur ces deux postes. On conclut que le GLM avec le nouveau zonier est meilleur que le GLM sans zonier, d'où l'intérêt de rajouter le zonier lors d'une tarification.

Conclusion

Dans ce présent mémoire, nous avons créé un nouveau zonier en utilisant les variables tarifaires et des variables géographiques externes. Ainsi, ce nouveau zonier prend en compte le facteur de risque géographique.
Premièrement, nous avons donné un intérêt particulier à nos données d'études, nous sommes partis des flux mensuels reçus de nos partenaires pour construire une base technique d'études. Nous avons donné une attention particulière à la qualité des données qui est une étape importante, car elle nous permet de mieux connaître nos données.
Ensuite, nous allons modéliser la fréquence des sinistres par poste de garanties. En effet, pour chaque poste de garanties, nous avons dû créer un modèle GLM, ce qui nous a permis d'avoir des résidus, que nous avons par la suite lissés en utilisant les méthodes de lissage notamment le krigeage. Ce mémoire nous a permis d'étudier l'effet des variables géographiques externes en utilisant les techniques de machines Learning telles que le Random Forest.
Enfin, nous avons pu élaborer le nouveau zonier ce qui nous a permis de le comparer avec le zonier du partenaire et le zonier des Agents Généraux (hors partenariats). Il en ressort que les zoniers du partenaire ainsi que des Agents Généraux (hors partenariats) sont peu discriminants car ils ne prennent pas en comptent le facteur de risque tandis que notre nouveau zonier est un zonier technique et prend bien en compte le risque géographique.
La comparaison avec ces zoniers fait partie des limites du mémoire. En effet, nous avons comparé notre zonier fréquence avec un zonier qui prend en compte les charges.
La suite de cette étude sera de faire un zonier coût moyen. Même si nous avons comparé notre zonier avec un zonier commercial, ce nouveau zonier que nous avons élaboré nous permettra de faire des analyses complémentaires des études de rentabilité de notre portefeuille : le leakage.
Le leakage nous permet de faire la comparaison entre la prime commerciale payée par le client et la prime technique qui ressortira de notre zonier technique ainsi que d'un GLM technique, pour comparer la prime technique et la prime commerciale. Cela nous permettra de voir dans la globalité comment est distribué le portefeuille et de définir la revalorisation de notre portefeuille chaque année.
La suite donnée à ces travaux consistera à faire un zonier complet avec la charge puis appliquer cette donnée technique pure à tout notre portefeuille pour étudier la rentabilité, la distribution de rentabilité par rapport à la prime commerciale.
Enfin, un autre axe d'amélioration a été noté : lors de l'élaboration d'un nouveau zonier, nous tenterons de le faire à la maille commune, cela nous permettra d'avoir des zones de risques plus fines.

Annexe

Loi de poisson
La loi de Poisson est une loi de probabilité discrète qui décrit le comportement d'un phénomène se produisant dans un intervalle de temps donné.
Loi de Gamma
La loi Gamma est un type de loi de probabilité de variables aléatoires réelles positives. La famille des distributions Gamma inclut, entre autres, la loi du χ^2 et les distributions exponentielles.
Loi de binomiale
La loi binomiale modélise la fréquence du nombre de succès obtenus lors de la répétition de plusieurs expériences aléatoires identiques et indépendantes.
Loi normale, les lois normales sont parmi les lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires.

Mesures d'importance variable dans les forêts aléatoires
La fonction varImpPlot() de R appliqué sur le modèle nous pour vérifier visuellement l'importance de la variable
La mesure d'importance des variables est basée sur Mean Decrease Gini (IncNodePurity).
Il s'agit d'une mesure d'importance variable basée sur l'indice d'impuretés de Gini utilisé pour le calcul des fractionnements dans les arbres.
Plus la valeur de la précision de la diminution moyenne ou du score de Gini de la diminution moyenne est élevée, plus l'importance de la variable pour notre modèle est élevée.
Gini Impureté Gini Impurityest une mesure de la probabilité d'une classification incorrecte d'une nouvelle instance d'une variable aléatoire, si cette nouvelle instance était classée au hasard en fonction de la distribution des étiquettes de classe à partir de l'ensemble de données.
L'impureté Gini est limitée par 0 , 0 se produisant si l'ensemble de données ne contient qu'une seule classe.

Table des figures

Liste des tableaux

Bibliographie

1. https ://statsandr.com/blog/correlogram-in-r-how-to-highlight-the-most-correlated-variables-in-a-dataset/>
2. https ://fr.wikipedia.org/wiki/Methodedeskplusprochesvoisins>
3. https ://eric.univlyon2.fr/ ricco/cours/slides/baggingboosting.pdf
4. https ://www.kongakura.fr/article/Random-Forest-explication-et-impl
5. https ://medium.com/analytics-vidhya/mathematics-behind-random-forest-and-xgboost-ea8596657275
6. http ://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-inù-r/
7. www.drees.sante.gouv.fr.
8. Insee. www.insee.fr.
9. Irdes. www.irdes.fr.
10. L'assurance maladie en ligne. www.ameli.fr.
11. https ://www.georisques.gouv.fr
12. data.gouv.fr
13. Impact de la réglementation 100
14. Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial : Catalina SEPULVEDA, Mémoire d'actuariat.
15. Maud THOMAS. Econométrie de l'assurance non-vie.
16. DI BERNARDINO Elena. Théorie des copules.
17. Construction d'un zonier en assurance MRH, Issam MEZRAG Mémoire d'actuariat.
18. Construction d'un zonier en assurance auto, Jean de Dieu Mémoire d'actuariat.