Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 30 octobre 2021

Par : Benjamin AUBIGNAT
Titre : Construction d'un zonier à la maille ville en assurance santé collective à l'aide de méthodes de Data Science.

Confidentialité : Oui - (Durée: 2 ans)

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

Membre présent du jury de l'Institut des Actuaires :
Swan BROUTARD
Marine HABART
Léonard FONTAINE
Signatures :

Entreprise :
Optimind
Signature :

Membres présents du jury de l'EURIA :
Philippe LENCA

Directeur de mémoire en entreprise :
Thibault HOUSSAY
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

L'impact géographique sur la consommation médicale est un sujet complexe, tant sur le plan politique qu'actuariel. En assurance santé, la situation géographique constitue une des principales sources de risque. Négliger l'impact du risque spatial relatif à un portefeuille d'assurés peut conduire à un mauvais ajustement de la prime pure d'assurance. Cette prime, calculée actuariellement, vise à être la plus proche possible du risque intrinsèque des assurés. Une tarification trop peu précise sur ce point peut entraîner de l'antisélection, et à terme, la souscription du mauvais risque.

Le comportement d'un assuré quant à sa consommation médicale est en partie véhiculé par des facteurs liés à la zone, tels que l'offre de soins ou le niveau de vie. Pour segmenter les risques, les organismes complémentaires d'assurance maladie cherchent à isoler le plus justement possible l'influence de l'effet spatial afférent au comportement d'un assuré. Cet effet spatial peut être peu ou prou important d'une zone à l'autre. De ce fait, ces zones peuvent être regroupées dans des classes de risques homogènes. L'ensemble de ces classes est appelé *zonier*.

Ce mémoire a pour objectif de détailler la construction de zoniers sur un portefeuille santé collective à l'aide de méthodes de Data Science et de données médicales et socio-économiques en libre accès et exogènes au portefeuille initial, ainsi que l'interprétation des résultats qui en découlent.

Mots clefs: Assurance collective, santé, zonier, déserts médicaux, tarification, GLM, Apprentissage statistique, science des données, forêts aléatoires, gradient boosting, données ouvertes, lissage

Abstract

The impact of geographic effect on medical consumption is a complex issue on many levels, such as political, health, and actuarial levels. In health insurance, geographic information is one of the main sources of risk. Neglecting the impact of spatial risk on an insured's portfolio can lead to an under or over priced one. The health insurance premium, calculated based on actuarial formulas, aims to cover the insured's value of risk.

An insured's medical consumption behavior is partly driven by area-related factors, such as the health care supply, medical needs, standard of living, or other unexplained components. In order to split the risk, supplementary health insurance organizations try to isolate the effect of one variable at a time, in particular the influence of the spatial effect on an insured's behavior.The importance of this latter may differ from one area to another. As a result, these areas can be grouped into homogeneous risk classes. The set of classes is called *zoning*.

The objective of this thesis is to detail the ways of setting zonings' area on a group health insurance portfolio using Data Science methods and open source external medical and socio-economic data, and to analyze the results.

Keywords: Group insurance, health, zoning system, medical deserts data, GLM, machine learning, data science, random forests, gradient boosting, open data, smoothing

Remerciements

Je remercie toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire. Je remercie l'équipe Optimind pour la bienveillance, l'humanité et le sérieux qu'ils ont pu me partager.

J'adresse ma reconnaissance à mon tuteur d'entreprise, Thibault Houssay, qui m'a accompagné tout au long de mon stage et suivi dans la réalisation de ce mémoire.

Je tiens à remercier Coralie Bonnifait, ainsi que Antoine Pesneaud, pour m'avoir fait part de leurs connaissances, de leur expérience et de leur temps.

Également, je souhaite remercier ma tutrice académique, Anaëlle Le Berre, pour s'être tenue disponible dans la supervision de mon mémoire.

Un grand merci à Franck Vermet, directeur de l'EURIA, ainsi qu'à Paul Friedel, directeur de l'IMT Atlantique, et à l'ensemble du corps enseignant m'ayant offert la chance d'acquérir une double formation d'ingénieur-actuaire. Une attention particulière est donnée à Philippe Lenca, responsable du double diplôme IMT-EURIA.

Enfin, je remercie ma famille pour l'éducation et les valeurs qu'elle a pu me transmettre.

Note de synthèse

Contexte

Les contrats d'assurance complémentaire santé permettent aux assurés de se voir indemniser une partie de leurs dépenses de santé. Les organismes complémentaires d'assurance maladie (OCAM), qui couvrent une partie de ces dépenses via des prestations, cherchent à ajuster leurs primes pures de telle sorte que ces dernières soient afférentes au risque de versement d'une prestation. Pour calculer ces primes, les OCAM ont des besoins de segmentations tarifaires de plus en plus fines. L'information géographique constitue une variable importante pour les OCAM qui visent à différencier les assurés par des critères tarifaires discriminants.

Ainsi, tenir compte de l'information géographique dans les modèles de tarification permet de différencier les assurés résidents dans des zones où les facteurs incitent à la sur/sous-consommation. A titre d'exemple, un assuré résident dans une zone de désert médical serait moins enclin à consulter régulièrement un médecin spécialiste qu'un assuré résident à Paris où la densité de médecins spécialistes est importante. Toutefois, la disparité médicale n'est pas le seul facteur géographique lié à la consommation. En effet, le niveau de vie, l'activité socio-économique, ou d'autres facteurs font de la zone un critère discriminant.

La granularité de cette zone doit être choisie de façon adéquate afin que le lien entre l'appartenance à une zone et le comportement de l'assuré fasse sens.

Cependant, les modèles de tarification, tels que les modèles linéaires généralisés, ne sont pas toujours en mesure de supporter une information trop granulaire. Ainsi, pour pallier ce phénomène, des méthodes de Data Science permettent de classifier les zones de sorte que le risque intra-classe soit le plus homogène possible, avec un nombre de classe modéré. L'ensemble de ces classes est appelé *zonier*. Pour la construction d'un zonier, et d'un modèle de tarification robuste, il convient d'avoir à disposition des données conséquentes qualitativement et quantitativement.

Le portefeuille

La base de données utilisée pour cette étude est un portefeuille santé d'entreprise fourni par une entreprise cliente d'Optimind implantée sur plusieurs sites en France.

La base effectif a pour variables discriminantes l'âge, le sexe et le département de résidence. Le portefeuille étudié est composé de salariés actifs de l'entreprise ainsi que de leurs ayants droits.

Les assurés principaux de ce portefeuille ont une moyenne d'âge de 43 ans, tandis que les conjoints ont une moyenne d'âge de 49 ans et les enfants de 12 ans. La figure 1 présente la pyramide des âges du portefeuille des bénéficiaires.

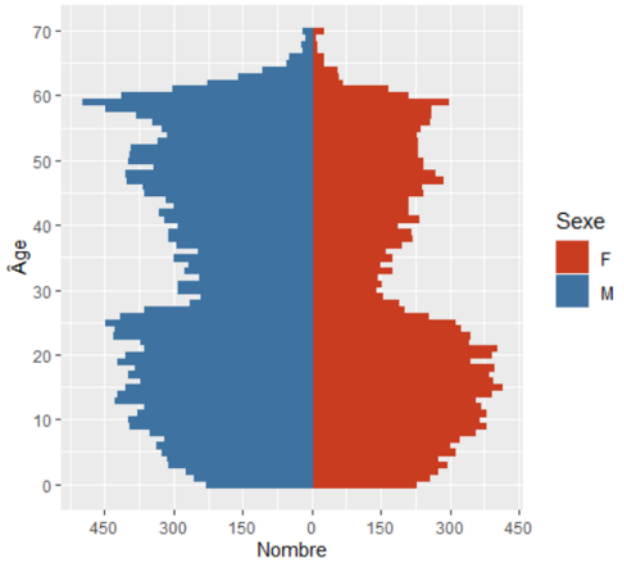

FIGURE 1 – Pyramide des âges de l'effectif des bénéficiaires

Dans ce portefeuille, 46% des bénéficiaires sont des femmes, et parmi les assurés principaux, 34% sont des femmes.

Les bénéficiaires sont répartis dans la majorité des départements de l'hexagone, mais pas de manière uniforme. La figure 2 présente cartographiquement la présence des assurés par département en indiquant la présence d'un site de l'entreprise.

Par la suite, les assurés de la base de données se verront attribuer un code postal de résidence, en fonction de leur département de résidence, qui sera généré de telle sorte que la distribution de ces codes postaux suive la distribution intra-départementale de la population française.

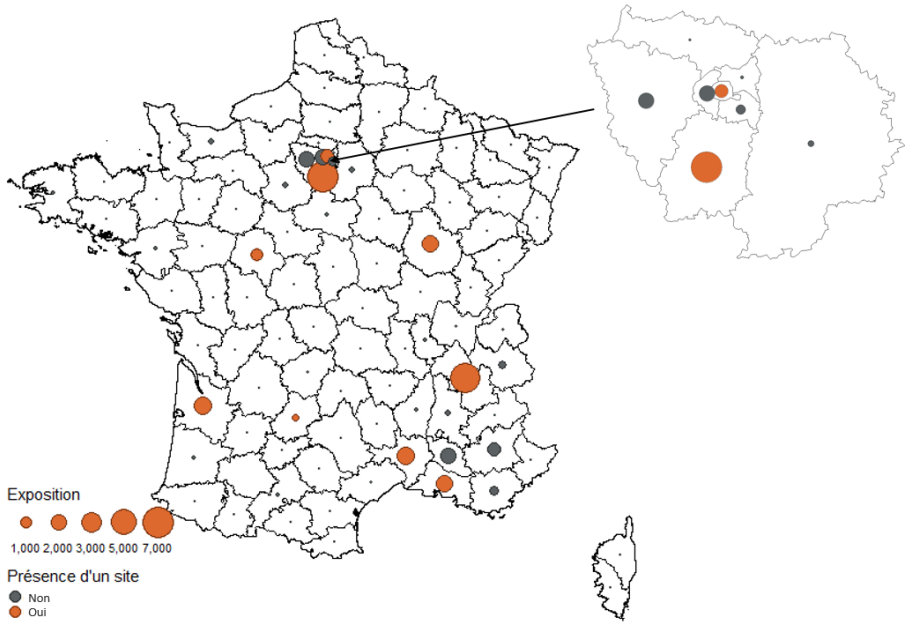

FIGURE 2 – Exposition par département

Étapes de construction d'un zonier

Dans le cadre de ce mémoire, le zonier se construit à partir de l'erreur résiduelle de la modélisation de la sinistralité, appelé résidu, agrégé à chaque zone. Si une zone possède un résidu positif, cela veut dire que la zone est plus sinistrée que l'assureur l'a prédit. Cette zone représente donc un risque de sous-tarification. A l'inverse, un résidu négatif indique que la sinistralité a été surestimée par le modèle, ce qui n'est pas nécessairement une bonne chose pour l'assureur dans le cadre d'un contrat santé collective. En effet, un contrat « sur-tarifé » inciterait l'entreprise à se tourner vers un assureur concurrent pour assurer ses salariés, et par extension conduirait à de l'antisélection et de la perte de chiffre d'affaire sur du bon risque. Le schéma 3 représente le spectre du risque spatial lié au résidu.

FIGURE 3 – Spectre du risque spatial

Les étapes de construction d'un zonier sont les suivantes :

1. Modéliser la sinistralité du portefeuille par un modèle linéaire généralisé (GLM) à l'aide de variables explicatives non-géographiques. Ce modèle est appelé *modèle contraint* ;
2. Récupérer les résidus issus du modèle contraint. Ces résidus contiennent l'information géographique, ou effet spatial ;
3. Agréger les résidus pour chaque zone en pondérant par l'exposition de celle-ci ;
4. Effectuer un lissage spatial des résidus de manière à homogénéiser géographiquement la répartition des résidus ;
5. Modéliser les résidus par Machine Learning à l'aide de variables géographiques afin de capter l'effet spatial contenu dans ceux-ci ;
6. Classifier l'effet spatial et réduire le nombre de classes.

Le schéma 4 illustre la décomposition de la sinistralité en ces éléments.

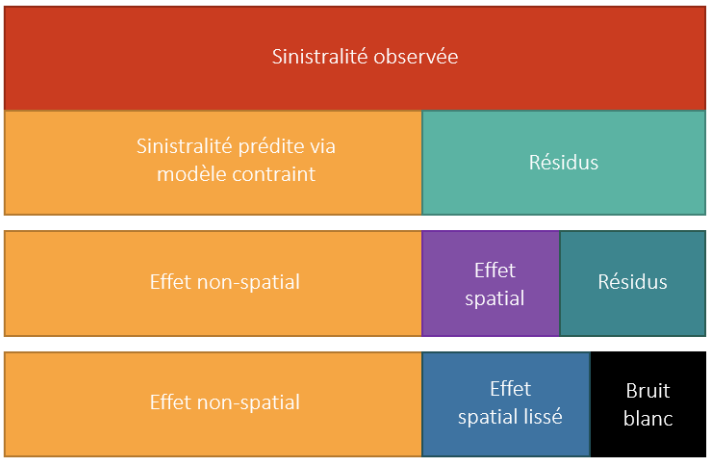

FIGURE 4 – Décomposition de la sinistralité selon les étapes du zonier

L'exemple ci-dessus de la sinistralité peut également s'appliquer à la modélisation de la fréquence et/ou du coût.

Mise en pratique : Construction du zonier

Le zonier a été construit sur deux types de sinistralité : la fréquence et le coût.

La fréquence a été modélisée par un GLM avec pour loi de distribution la loi Zéro-inflated Binomiale négative (ZIBN) et le coût a été modélisé par un GLM, avec la loi Log-normale. Après avoir construit des classes d'âges à l'aide d'un arbre de décision, les variables non-géographiques ont été sélectionnées par une méthode *Stepwise*. Les variables conservées sont : la classe d'âge et le sexe. Afin de vérifier la cohérence des coefficients linéaires, la sinistralité moyenne par modalité de variable a été observée. A titre d'exemple, les figures 5 et 6 confrontent la fréquence moyenne prédite pour chaque modalité de variable face à la fréquence moyenne observée, pour l'acte *Consultations spécialistes*.

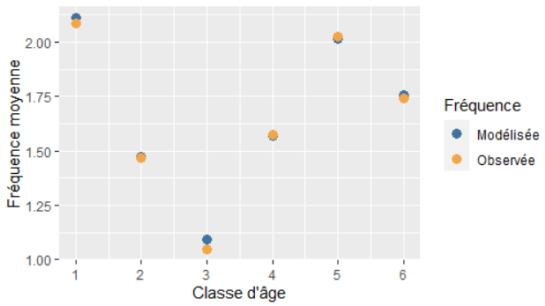

FIGURE 5 – Fréquence moyenne par classe d'âge

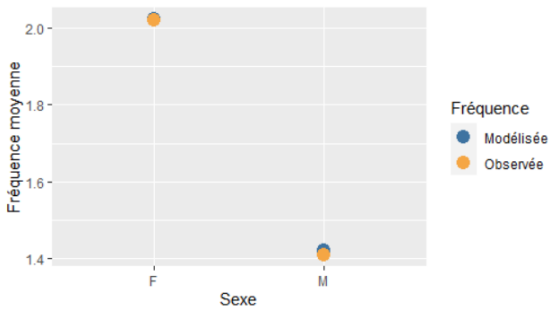

FIGURE 6 – Fréquence moyenne par sexe

L'erreur résiduelle découlant du modèle a été agrégée à chaque zone selon la maille géographique désirée, à savoir la maille départementale d'une part, et la maille code postal d'autre part. Certaines zones étant peu voire non couvertes par le portefeuille, un lissage spatial a été effectué sur les résidus en vue de couvrir les zones non observées, et atténuer les valeurs extrêmes.

Le lissage spatial des résidus se fait par la théorie de la crédibilité. L'idée est de donner à une zone une nouvelle valeur de résidu qui sera une pondération entre sa valeur effective et la valeur de l'ensemble du portefeuille. Pour une zone i ayant pour résidu r_i, son résidu lissé r_i^* s'exprime comme suit (voir équation 1).

$$r_i^* = Z_i \times r_i + (1 - Z_i) \times \frac{\sum_{j=1}^{n} r_j \times e_j \times f(d_{i,j})}{\sum_{j=1}^{n} e_j \times f(d_{i,j})} \tag{1}$$

Avec :
— Z_i le facteur de crédibilité, compris entre 0 et 1, associé à la zone i ;
— e_i l'exposition totale au sein de la zone i ;
— $f(d_{i,j})$ une fonction monotone de la distance entre la zone i et la zone j ;
— n le nombre de zones.

La figure 7 présente le résultat du lissage spatial des résidus issus du modèle GLM fréquence, agrégés à la maille code postal. Les nuances de couleurs sont ajustées par quantile à pas de 20%.

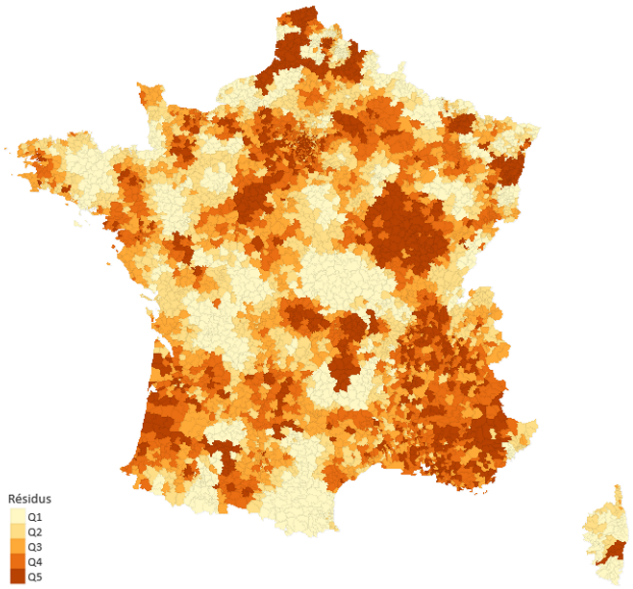

FIGURE 7 – Résidus fréquence à la maille code postal lissés

Par la suite, deux modèles de Machine Learning sont utilisés, afin de capter l'effet spatial inclus dans le résidu lissé, à l'aide de variables géographiques provenant de données médicales et socio-économiques en libre accès. Ces modèles sont le Random Forest et le GBM. Le choix des paramètres optimaux pour calibrer ces modèles s'est fait par validation croisée.

A titre indicatif, les figures 8 et 9 présentent cartographiquement l'effet spatial capté par chacun des modèles. Les nuances de couleurs sont ajustées par quantile à pas de 20%.

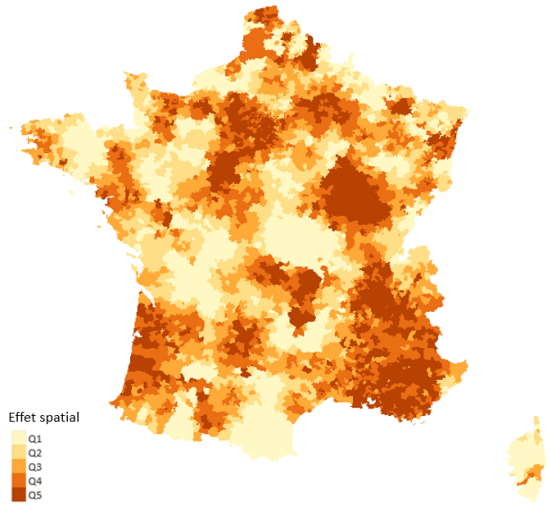

FIGURE 8 – Effet spatial fréquence modélisé par Random Forest

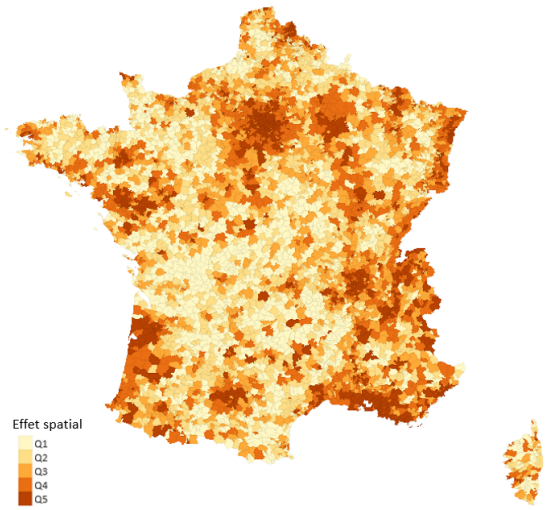

FIGURE 9 – Effet spatial fréquence modélisé par GBM

Enfin, l'effet spatial a été classifié par méthode de classification ascendante hiérarchique (CAH), qui a pour objectif de classer des éléments dans des groupes de sorte que l'inertie intra-groupe soit minimale, et l'inertie inter-groupe soit maximale. Le choix du nombre de classes k est basé sur le seuil de stabilité de l'inertie intra-groupe en fonction du nombre de classes. A titre d'exemple, la figure 10 représente l'évolution de l'inertie intra-classe de l'effet spatial fréquence par nombre de classes k, pour nos deux modèles.

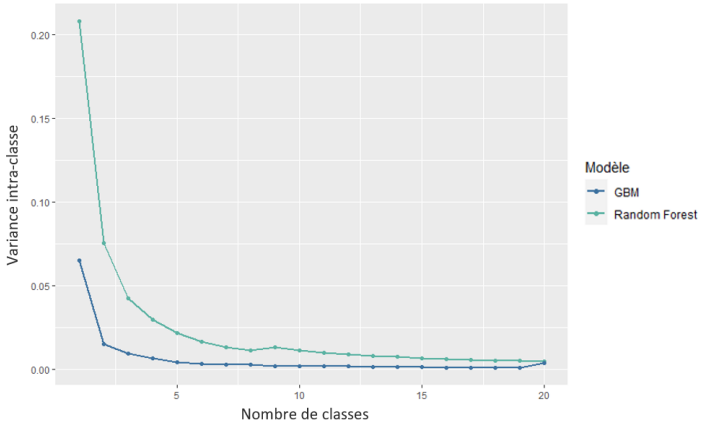

FIGURE 10 – Évolution de la variance intra-classe de l'effet spatial maille code postal

Ici, l'inertie intra-classe de l'effet spatial se stabilise à partir de $k = 10$. C'est le nombre de classes qui a été choisi pour la maille code postal, et le même nombre pour la maille départementale en procédant par la même démarche.

Résultats

L'effet spatial de chaque zone a été classifié, et, par extension, les zoniers ont été construits pour la fréquence et le coût. A titre indicatif, le zonier fréquence à la maille départementale se présente cartographiquement dans la figure 11.

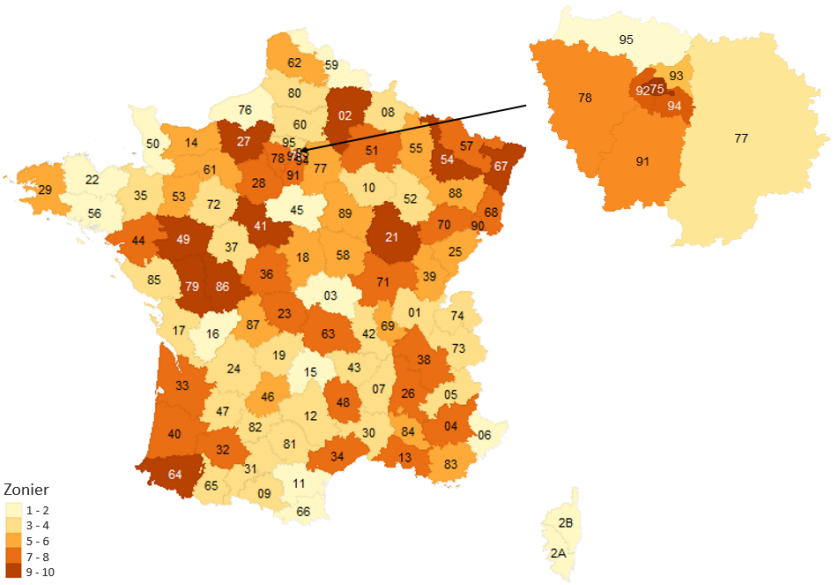

FIGURE 11 – Zonier fréquence à la maille départementale

Les zoniers fréquence à la maille ville, respectivement construits par Random Forest et GBM, se présentent cartographiquement dans les figures 12 et 13.

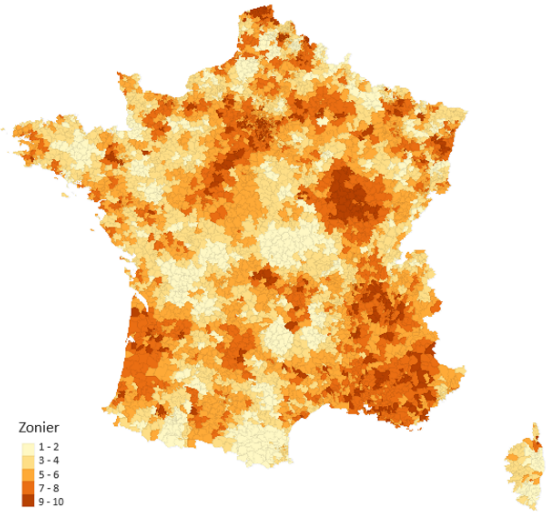

FIGURE 12 – Zonier fréquence maille code postal construit par Random Forest

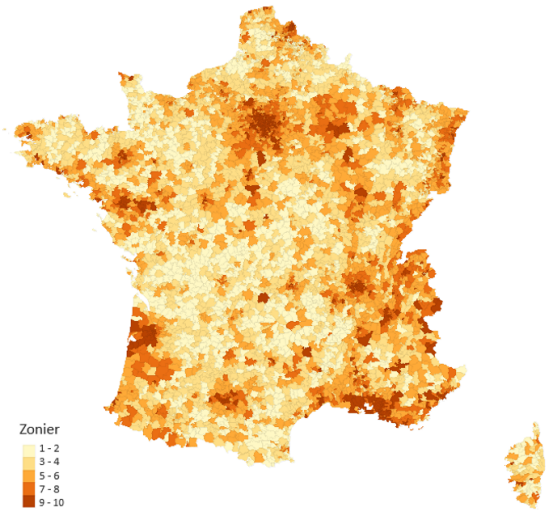

FIGURE 13 – Zonier fréquence maille code postal construit par GBM

Enfin, les zoniers étant construits, il est alors possible de créer une nouvelle variable tarifaire pour chaque zonier, et analyser leur impact respectif sur les performances des modèles de prédiction. Le modèle contraint enrichi de la variable *zonier* s'appelle *modèle complet*.

Le tableau 1 (resp. 2) compare l'impact de chaque zonier sur la modélisation de la fréquence (resp. du coût).

Modèle	AIC	MSE	MAE
Modèle contraint	132 244	8.7694	1.7282
Modèle complet maille départementale	131 875	8.7118	1.7184
Modèle complet maille ville Random Forest	130 865	8.5611	1.7001
Modèle complet maille ville GBM	132 101	8.7552	1.7242

TABLE 1 – Indicateurs de performance pour chaque modèle GLM fréquence

Modèle	AIC	MSE	MAE
Modèle contraint	17 608	287.5	12.69
Modèle complet maille départementale	16 507	272.8	12.26
Modèle complet maille ville Random Forest	16 152	268.6	12.14
Modèle complet maille ville GBM	17 114	279.9	12.52

TABLE 2 – Indicateurs de performance pour chaque modèle GLM coût

Il apparaît que chaque zonier diminue les indicateurs de performance par rapport au modèle contraint, ce qui est un gage de qualité. Par ailleurs, ces indicateurs vont unanimement en faveur du zonier construit par Random Forest.

L'ensemble des études réalisées a donc permis une amélioration de nos modèles prédictifs de fréquence et de coût pour notre portefeuille.

Summary

Context

Supplemental health insurance contracts allow policyholders to be reimbursed for part of their health expenses. Supplemental health insurance organizations, which cover part of these expenses through benefits, try to adjust their pure premiums such that they are related to the risk of payment of a benefit. In order to calculate these premiums, OCAMs need increasingly fine tariff segmentations. Geographical information is an important variable for OCAMs that aim to differentiate policyholders by discriminating tariff criteria.

Thus, taking geographic information into account in pricing models makes it possible to differentiate between policyholders living in areas where factors encourage over- or under-consumption. For example, a policyholder living in a medical desert zone would be less likely to consult a specialist on a regular basis than a policyholder living in Paris where the density of specialists is high. However, medical disparity is not the only geographical factor linked to consumption. Indeed, the standard of living, needs, pollution, or other unexplained factors make the zone a discriminating criterion.

The granularity of this zone must be chosen appropriately so that the link between zone membership and policyholder behavior makes sense.

However, pricing models, such as generalized linear models, are not always able to support too much granular information. Thus, to overcome this phenomenon, data science methods can be used to classify zones so that the intra-class risk is as homogeneous as possible, with a moderate number of classes. The set of these classes is called *zoning*. To build a zoning, and a robust pricing model, it is necessary to have qualitatively and quantitatively consistent data.

Presentation of the portfolio

The database used for this study is a corporate health portfolio provided by an Optimind client company located in several regions in France.

The discriminant variables in the database are age, gender and county of residence. The studied portfolio is composed of active employees of the company as well as their beneficiaries.

The average age of the main insureds in this portfolio is 43 years old, while the average age of their spouses is 49 years old and their children is 12 years old. The age pyramid for the portfolio of beneficiaries is shown in Figure 14.

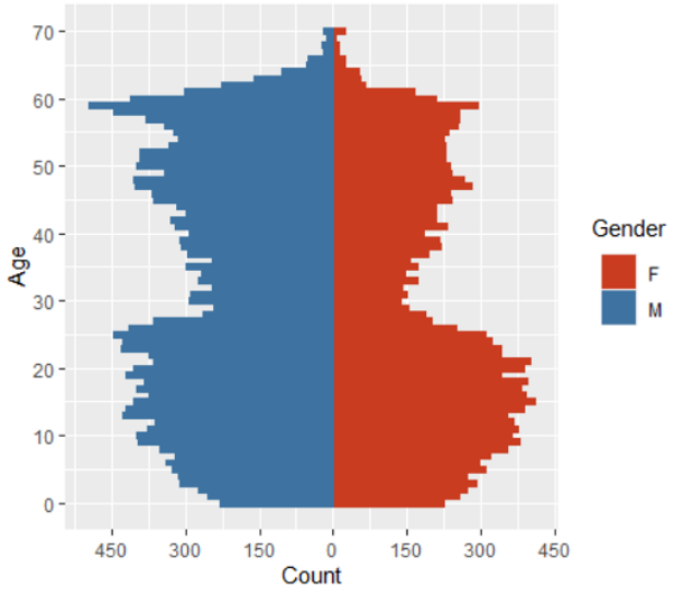

FIGURE 14 – Age pyramid of the insured workforce

In this portfolio, 46% of the beneficiaries are women, and among the principal insured, 34% are women.

The beneficiaries are distributed in the majority of the French counties, but not in a uniform manner. The figure 15 shows the presence of policyholders by county, indicating the presence of a company site.

Subsequently, the policyholders in the database will be assigned a postcode of residence, according to their county, which will be generated so that the distribution of these postcodes follows the within-county distribution of the French population.

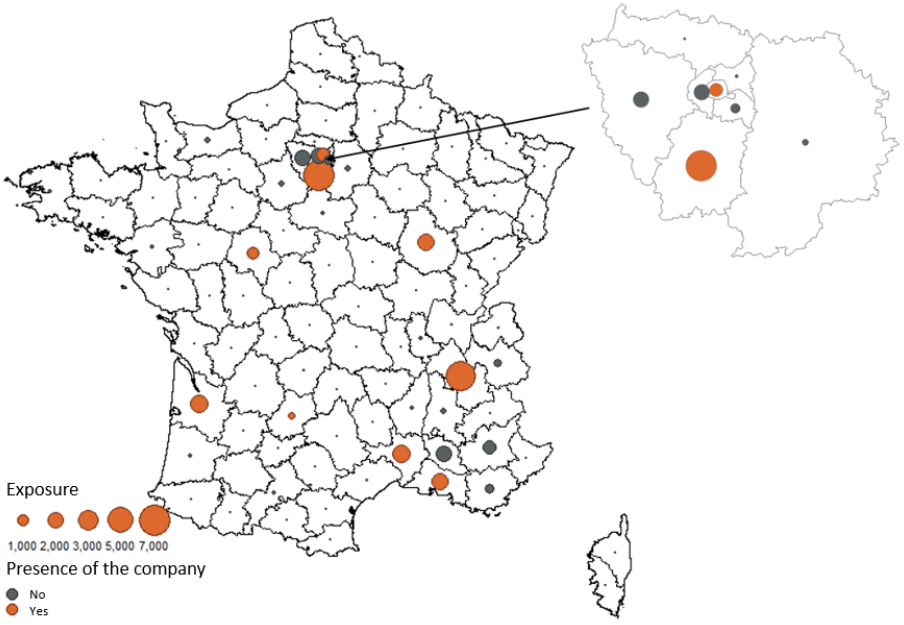

FIGURE 15 – Exposure by county

Steps to build a zoning

For the purposes of this thesis, the zoning is constructed from the residual error of the loss modeling, aggregated to each zone. If a zone has a positive residual, it means that the zone has a higher loss than the insurer predicted. This zone therefore represents an underwriting risk. Conversely, a negative residual indicates that the claims experience has been overestimated by the model, which is not necessarily a good thing for the insurer in the context of a group health contract. Indeed, an overpriced contract would encourage the company to turn to a competing insurer to insure its employees, and by extension would lead to anti-selection and loss of turnover on good risk. The diagram 16 represents the spectrum of spatial risk related to the residual.

FIGURE 16 – Spatial risk spectrum

The steps for building a zoning are as follows :

1. Model the portfolio loss experience by a generalized linear model (GLM) using non-geographic explanatory variables. This model is called *restrained model* ;
2. Retrieve the residuals from the restrained model. These residuals contain the geographic information, or spatial effect ;
3. Aggregate the residuals for each zone by weighting by the zone's exposure ;
4. Spatial smoothing of the residuals in order to homogenize the distribution of the residuals geographically ;
5. Model the residuals by Machine Learning using geographical variables in order to capture the spatial effect contained in them ;
6. Classify the spatial effect and reduce the number of classes.

The figure 17 illustrates the breakdown of the loss into these elements.

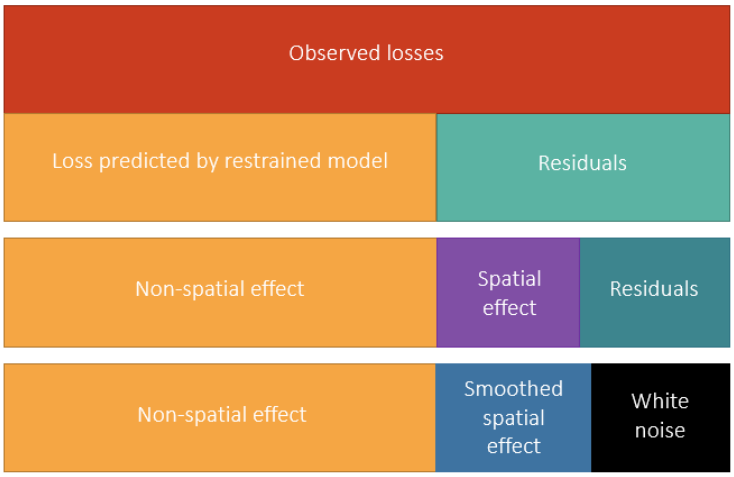

FIGURE 17 – Breakdown of losses following zoning steps

The above example of losses can also be applied to frequency and/or cost modeling.

Practice : Building the zoning

The zoning was built on two types of losses : frequency and cost.

The frequency was modeled by a GLM with the Zero-inflated Binomial negative distribution (ZINB) and the cost was modeled by a GLM, with the Log-normal distribution. After constructing age ranges using a decision tree, the non-geographic variables were selected by a *Stepwise* method. The retained variables are : age range and gender. In order to check the consistency of the linear coefficients, the average loss by variable modality was observed. As an example, the figures 18 and 19 compare the predicted average frequency for each variable modality with the observed average frequency, for the medical act *Specialist consultations*.

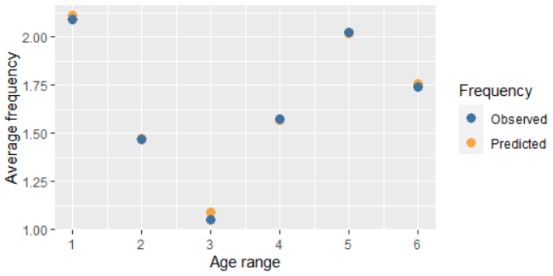

FIGURE 18 – Average frequency by age range

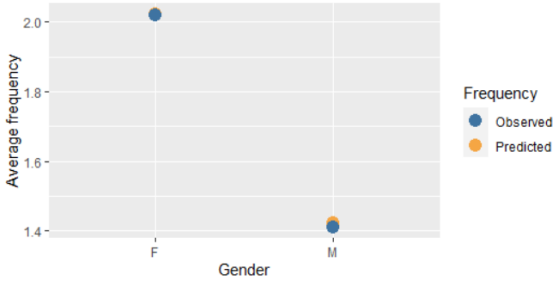

FIGURE 19 – Average frequency by gender

The residual error resulting from the model was aggregated for each zone according to the desired geographical level, i.e. the county level on the one hand, and the postcode level on the other. As some areas are not covered by the portfolio, a spatial smoothing was performed on the residuals in order to cover the unobserved areas and to attenuate the extreme values.

The spatial smoothing of the residuals is done by the credibility theory. The idea is to give a zone a new residual value which will be a weighting between its actual value and the value of the whole portfolio. For an area i with residual r_i, its smoothed residual r_i^* is expressed as follows (check equation 2).

$$r_i^* = Z_i \times r_i + (1 - Z_i) \times \frac{\sum_{j=1}^{n} r_j \times e_j \times f(d_{i,j})}{\sum_{j=1}^{n} e_j \times f(d_{i,j})} \tag{2}$$

Where :
— Z_i is the credibility factor, between 0 and 1, associated with the i^{th} zone ;
— e_i the total exposure in the i^{th} zone ;
— $f(d_{i,j})$ a monotonic function of the distance between the i^{th} zone and the j^{th} zone ;
— n the total number of zones.

Figure 20 shows the result of the spatial smoothing of the residuals from the GLM frequency model, aggregated to the zip code grid. The color shades are adjusted by quantile with a step of 20%.

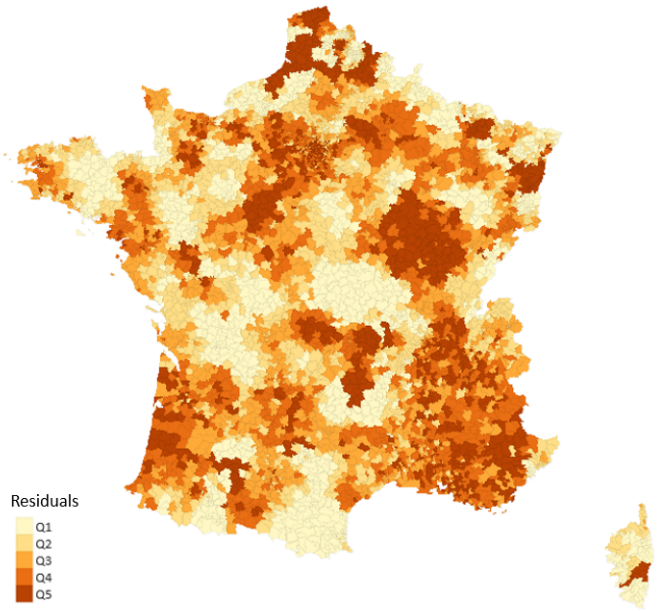

FIGURE 20 – Residuals frequency at the level postcode smoothed

For the next step, two Machine Learning models are used to capture the spatial effect included in the smoothed residual, using geographic variables from open source medical and socioeconomic data. These models are Random Forest and GBM. The choice of optimal parameters to calibrate these models was made by cross-validation.

As an indication, figures 21 and 22 present cartographically the spatial effect captured by each model. The color shades are adjusted by quantile with a step of 20%.

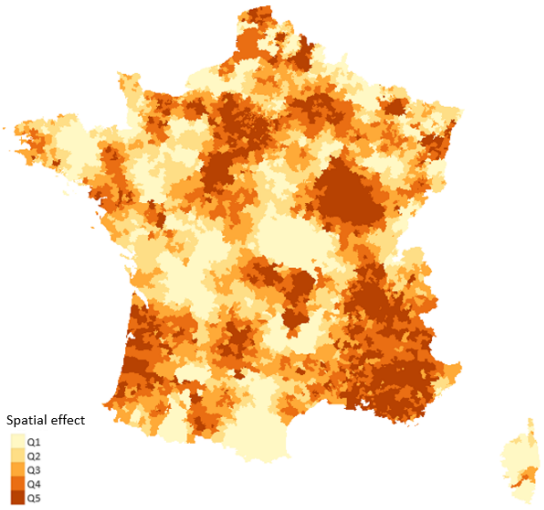

FIGURE 21 – Spatial frequency effect modeled by Random Forest

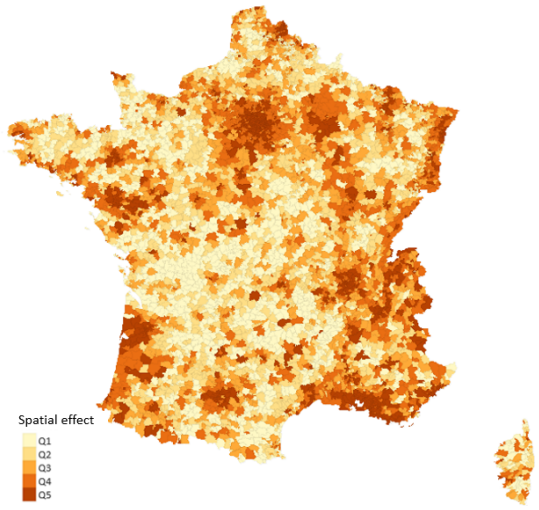

FIGURE 22 – Spatial frequency effect modeled by GBM

Finally, the spatial effect was clustered using the hierarchical clustering method, which aims to classify items into groups so that the within-group inertia is minimal and the between-group inertia is maximal. The choice of the number of clusters k is based on the stability threshold of

the intra-group inertia according to the number of clusters. As an example, figure 23 represents the evolution of the intra-group inertia of the spatial frequency effect by number of clusters k, for our two models.

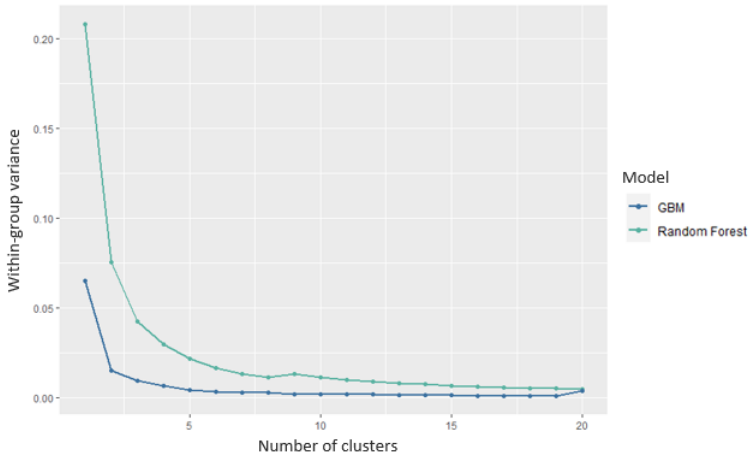

FIGURE 23 – Variation of the within-group variance of the postcode level spatial effect

Here, the intra-group inertia of the spatial effect stabilizes from $k = 10$. This is the number of clusters that was chosen for the postcode level, and the same number for the county level, using the same procedure.

Results

The spatial effect of each zone was clustered, and then, zonings were built for frequency and cost. As an example, the frequency zoning at the county level is shown in the figure 24 below.

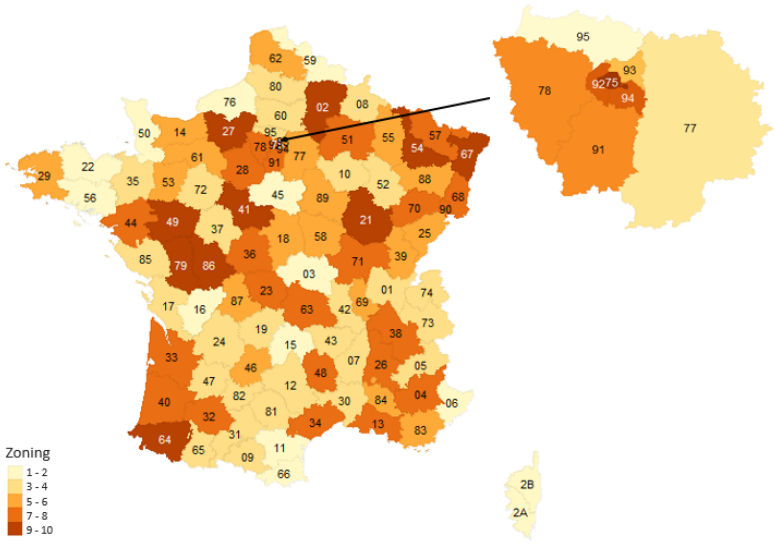

FIGURE 24 – Frequency zoning at the county level

The frequency zonings at the postcode level, respectively built by Random Forest and GBM, are shown cartographically in figures 25 and 26.

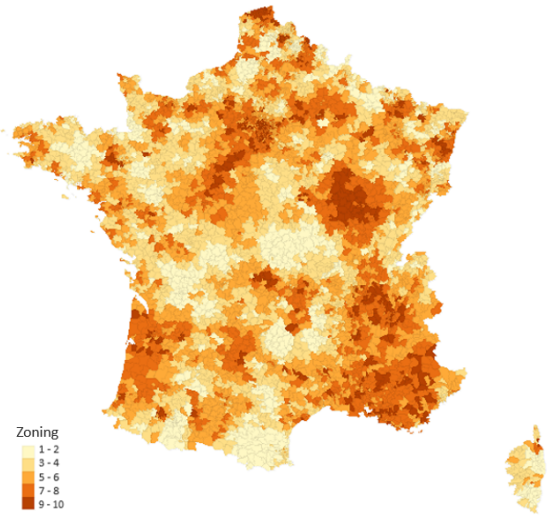

FIGURE 25 – Frequency zoning at the postcode level built by Random Forest

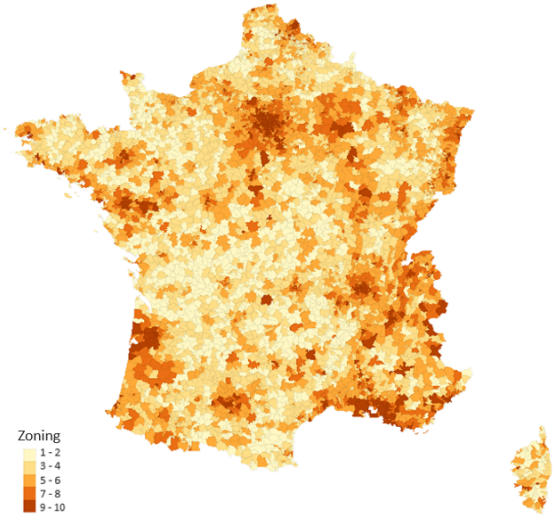

FIGURE 26 – Frequency zoning at the postcode level built by GBM

Finally, once the zonings have been built, it is then possible to create a new variable for each zoning, and analyze their respective impact on the performance of the models. The restrained model to with we have added the variable *zoning* is called *complete model*.

The table 3 (resp. 4) compares the impact of each zoning on the frequency (resp. cost) modeling.

Model	AIC	MSE	MAE
Restrained model	132 244	8.7694	1.7282
Complete model county level	131 875	8.7118	1.7184
Complete model postcode level Random Forest	130 865	8.5611	1.7001
Complete model postcode level GBM	132 101	8.7552	1.7242

TABLE 3 – Performance indicators for each GLM frequency model

Model	AIC	MSE	MAE
Restrained model	17 608	287.5	12.69
Complete model county level	16 507	272.8	12.26
Complete model postcode level Random Forest	16 152	268.6	12.14
Complete model postcode level GBM	17 114	279.9	12.52

TABLE 4 – Performance indicators for each GLM cost model

It can be noted that each zoning decreases the statistics compared to the constrained model, which is a guarantee of quality. Furthermore, these indicators are unanimously in favor of the zoning constructed by Random Forest.

All the studies carried out have therefore led to an improvement in our predictive models of frequency and cost for our portfolio.

Table des matières

Table des figures

Liste des tableaux

Introduction

Le système de santé en France est particulièrement performant en termes de couverture. En plus de la sécurité sociale, les organismes complémentaires d'assurance maladie (OCAM) permettent de réduire le reste à charge des assurés. Les primes pures, calculées actuariellement par les organismes complémentaires, visent à être les plus proches possibles des dépenses potentielles des assurés. Pour ce faire, la tarification doit être segmentée par des informations relatives à la consommation des assurés. La France recense des zones à plus ou moins forte consommation médicale, qui peut être dues à une offre de soins amoindrie, ou autres facteurs sociaux ou médicaux directement ou indirectement liés à la zone. Ainsi, tenir compte du lieu de résidence dans le calcul des tarifs permettrait de différencier les assurés résidents dans des zones dont la consommation est peu ou prou conséquente. Cependant, cette information se doit d'être à la fois suffisamment granulaire pour identifier de manière précise l'impact de l'effet géographique, et suffisamment lisse pour ne pas complexifier les modèles de tarification. La mise en place d'un zonier vise à construire une variable géographique combinant ces deux critères.

Ce mémoire propose une approche de classification des zones en fonction du risque qu'elles représentent d'un point de vue assurantiel, sur un portefeuille santé collective. L'ensemble de ces classes est appelé *zonier*.

La structure de ce mémoire comprend quatre grandes parties.

La première partie présente le cadre et le contexte de l'étude, à savoir l'assurance santé en France, ses acteurs, sa structure, son fonctionnement, et les grands principes de la tarification santé.

La deuxième partie introduit les données sur lesquelles les études seront effectuées. Elle détaille les traitements effectués sur ces dernières, et comprend une présentation statistique du portefeuille santé collective à disposition, ainsi qu'une présentation de données open data qui permettront de consolider notre étude.

La troisième partie porte sur l'aspect théorique, mathématique et actuariel des modèles et méthodes utilisés. Elle rappelle notamment les grands principes des modèles linéaires généralisés, des modèles d'apprentissage statistique, et les étapes et méthodes visant à construire une variable tarifaire géographique : le zonier.

La quatrième partie met en application les méthodes présentées dans la partie précédente, et présente les résultats qui en découlent. Elle vise dans un premier temps à modéliser la sinistralité via un modèle GLM contraint n'incluant aucune variable géographique avec de récupérer de façon optimale l'effet non-spatial relatif à la consommation médicale. Dans un second temps, cette partie vise à étudier et traiter les résidus issus de ce modèle contraint. Le zonier sera mis en place à partir de ces résidus, à l'aide du Machine Learning et des données en open source. Par ailleurs,

deux zoniers seront construits : un zonier à la maille départementale, et un zonier à la maille code postal. Cette double étude permettra de conforter la nécessité d'une maille géographique granulaire, ou bien la rejeter. Enfin, cette partie s'achève par une analyse de l'apport de ce zonier, ses limites, et les perspectives qu'il entraîne. Elle vise entre autres à analyser l'impact de l'effet spatial dans le cadre de notre portefeuille santé en comparant les modèles non-géographiques aux modèles enrichis de la variable construite, le zonier.

Première partie

Cadre de l'étude

Chapitre 1

L'assurance santé en France

La France est caractérisée par un système de santé universel particulièrement performant en termes de prise en charge des frais de santé, notamment grâce aux organismes de santé tels que l'Assurance Maladie ou les organismes complémentaires d'assurance maladie (OCAM). Pour maintenir cette performance, le contexte réglementaire de l'assurance santé connaît des réformes régulières.

1.1 L'Assurance Maladie de la Sécurité sociale

Créée en 1945, la Sécurité sociale est un service public de l'Etat. Elle a pour objectif de couvrir l'ensemble de la population contre les conséquences financières des risques sociaux.

La Sécurité sociale regroupe trois types de régimes légaux, couvrant chacun une ou plusieurs catégories socioprofessionnelle :

- Le **régime général** pour les travailleurs salariés et les indépendants ;
- Le **régime agricole** pour les exploitants et les salariés agricoles ;
- Les **régimes spéciaux** (SNCF, militaire, marin...).

La Sécurité sociale est composée de cinq branches, couvrant chacune un ou plusieurs types de risques sociaux :

- la branche **Maladie** (CPAM), qui prend en charge une partie des frais de santé des assurés et favorise l'accès aux soins ;
- la branche **Famille** (CAF), qui prend en charge les prestations familiales. Ces prestations comprennent : l'accompagnement des familles dans leur vie quotidienne, l'accueil d'un jeune enfant, l'accès au logement, la lutte contre la précarité et le handicap ;
- la branche **Accidents du travail et Maladies professionnelles** (CPAM), qui gère les accidents professionnels auxquels sont confrontés les travailleurs, et les maladies professionnelles ;
- la branche **Retraite** (CARSAT, CNAV, CGSS, CSS), qui prend en charge l'ensemble des prestations liées à la vieillesse : la retraite, la pension minimum ;
- la branche **Cotisation et Recouvrement** (URSSAF), qui collecte l'ensemble des cotisations et contributions sociales. Ces fonds sont redistribués dans les caisses de la Sécurité sociale pour financer les branches précédentes.

L'Assurance Maladie est incluse dans la branche Maladie de la Sécurité sociale. Souvent appelée « Sécurité sociale » ou « Sécu » par abus de langage, l'Assurance Maladie prend en charge une partie des frais des actes et soins médicaux liés aux postes suivants :

— Les honoraires de médecins et auxiliaires médicaux ;
— Les analyses et examens de laboratoire ;
— Les médicaments ;
— Les frais d'optique, appareillages et prothèses ;
— Hospitalisation (hôpital ou à domicile) ;
— Transports (transport lié à une hospitalisation, transport ambulancier, etc.) ;
— Cure thermale (avec ou sans hospitalisation).

La Sécurité sociale définit pour chaque acte sur lesquels elle intervient une base de remboursement et un taux de remboursement. Ainsi, le remboursement de la Sécurité sociale peut être partiel ou intégral. L'articulation du remboursement d'un acte médical sera financé plus tard dans ce mémoire.

Pour financer ces aides, la Sécurité sociale dispose de trois sources de financement principales :
— Les **cotisations sociales**, qui représentent plus de la moitié du budget de la Sécurité sociale. Ce sont des versements obligatoires effectués par les salariés, les employeurs, et les non-salariés ;
— la **Contribution sociale Généralisée (CSG)** et la **Contribution pour le Remboursement des Dettes sociales (CRDS)**, prélevées à la source et qui concernent les personnes résident fiscalement en France ;
— les **impôts et autres taxes** sur les produits (tabac et alcool principalement).

En 2019, près de 90% de la population française est affiliée au régime général, et les cotisations et contributions représentent 90% du budget de la branche Maladie. Les soins de ville (consultation d'un professionnel de santé, médicaments, actes de biologie, indemnités journalières) représentent plus de 45% des dépenses de santé financées par l'Assurance Maladie.

Dans l'ensemble, l'Assurance Maladie, couplée à une assurance complémentaire santé, réduit de manière conséquente le reste à charge des assurés.

1.2 L'assurance complémentaire santé

La complémentaire santé, souvent appelée mutuelle par abus de langage, a pour objectif de compléter le remboursement des frais médicaux de la Sécurité sociale. Le contrat d'une complémentaire santé peut être individuel ou collectif. Si le système de remboursement est le même pour les deux types de contrats, les modalités diffèrent selon que le contrat soit collectif ou individuel.

En effet, le contrat individuel ne comprend que deux signataires : un assureur et un assuré. Le lien entre les deux parties est direct : l'assuré, qui est également le souscripteur du contrat, paye une prime régulière contre la promesse de versement d'une prestation par l'assureur en cas de réalisation d'un acte médical couvert.

Le contrat collectif est quant à lui passé entre l'assureur et un groupe de personnes, généralement les membres d'une entreprise ou d'une association. En conséquence, l'assuré et le souscripteur sont deux entités distinctes. Les clauses du contrat concernent les membres de l'entreprise et sont négociées entre l'employeur et l'assureur. A l'inverse du contrat individuel, les termes du contrat collectif peuvent être modifiés sans l'accord de l'assuré.

La complémentaire santé comprend trois types d'acteurs :
— Les **sociétés d'Assurance**, régies par le Code des Assurances. Il existe deux types de sociétés d'Assurance : les Sociétés Anonymes (SA) qui sont des sociétés à but lucratif commercialisant des produits d'assurance et reversant ses profits à ses actionnaires, et les Sociétés Mutuelles d'Assurance (SMA) qui n'ont pas d'actionnaires à rémunérer, et sont donc à but non-lucratif.

— Les **sociétés Mutuelles**, régies par le Code de la Mutualité. Les mutuelles sont des sociétés à but non-lucratif spécialisées dans l'assurance Santé individuelle. Elles reposent sur le principe de mutualité financière entre les assurés. Les fonds proviennent des cotisations des assurés.
— Les **Institutions de Prévoyance**, régies par le code de la Sécurité sociale, sont spécialisées dans les contrats collectifs et couvrent les risques de maladie, incapacité, invalidité, dépendance et décès.

En Santé individuelle, les mutuelles détiennent la majeure partie du marché, suivi des sociétés d'assurance. Les institutions de prévoyance sont davantage présentes en Santé collective, comme le montre la figure 1.1.

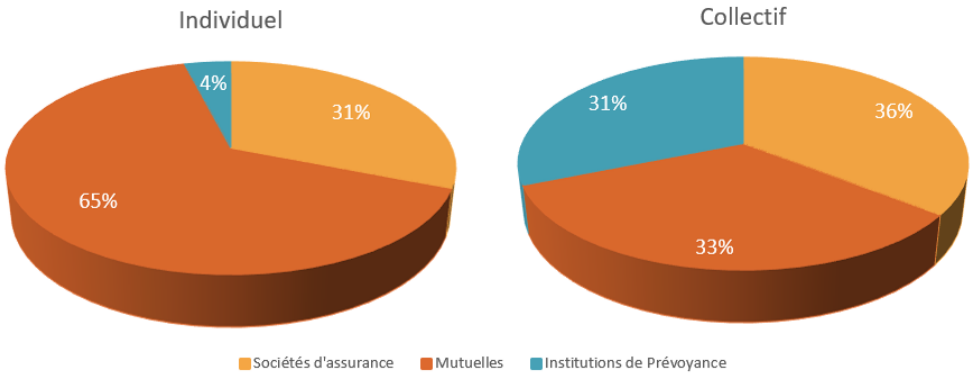

FIGURE 1.1 – Parts de marché des organismes de complémentaire santé

Les sources de financement des complémentaires santé viennent du paiement des primes par les assurés. En santé individuelle, les primes varient en fonction de plusieurs facteurs comme le type de garantie, l'âge et le lieu de résidence de l'assuré. En santé collective, chaque assuré paye la même prime fixée selon une grille de garantie, et le montant de celle-ci dépend des caractéristiques globales de l'entreprise, comme l'âge moyen ou la localisation de l'entreprise. Le détail théorique de ces calculs sera donné dans la partie 1.5.

1.3 Le remboursement d'un acte médical

Le remboursement d'un acte médical par la Sécurité sociale et la complémentaire santé comprend plusieurs éléments :

— Les **Frais Réels (FR)** représentent le coût total de l'acte médical exigé par le professionnel de santé ;
— la **Base de Remboursement (BR)** est le montant fixé par la Sécurité sociale inférieur ou égal au coût de l'acte ;
— le **Taux de Remboursement (TR)** est le taux de BR remboursé par l'assurance maladie de la Sécurité sociale ;
— le **Remboursement de la Sécurité sociale (RSS)** correspond au montant pris en charge par l'Assurance Maladie de la Sécurité sociale. C'est le produit $BR \times TR$;
— le **Ticket Modérateur (TM)** est la différence entre la base de remboursement et le remboursement de la Sécurité sociale.
— le **Remboursement complémentaire (RC)** est le remboursement effectué par la complémentaire santé ;

— le **dépassement d'honoraires** représente le montant dépassant la base de remboursement. Certains professionnels de santé peuvent en effet choisir de tarifer leurs prestations à un prix supérieur à la base de remboursement, le dépassement d'honoraires est la différence entre les frais réels et la base de remboursement ;
— le **Reste à Charge (RAC)** représente le montant que l'assuré aura à payer une fois les frais réels déduits des différents remboursements. C'est donc la différence : $FR - RSS - RC$. Pour certains actes, le reste à charge comprend une **participation forfaitaire** de 1 € [1].

Ces éléments varient en fonction de l'acte. La figure 1.2 représente la décomposition des frais d'une consultation généraliste en ces différents éléments.

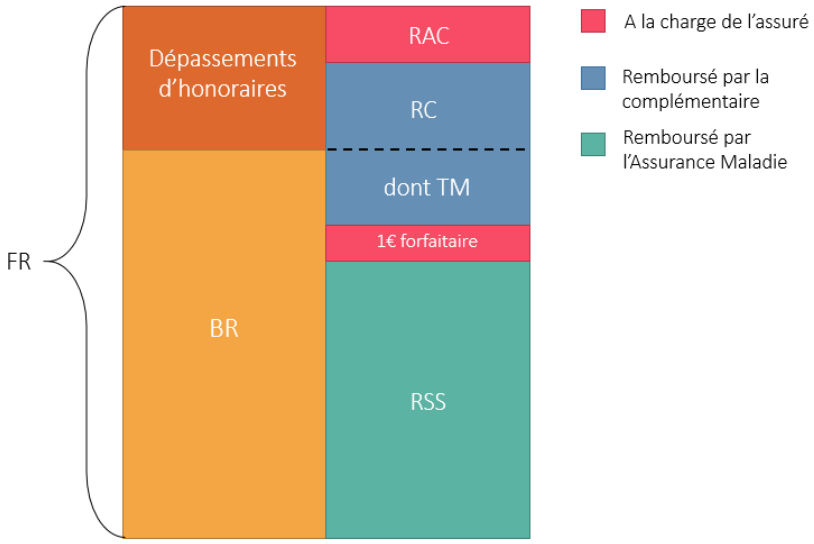

FIGURE 1.2 – Décomposition du remboursement d'une consultation généraliste [2]

Certains actes sont majoritairement pris en charge par la complémentaire santé, et inversement. Le tableau 1.1 présente des exemples du système de remboursement de la Sécurité sociale pour différents actes.

Exemple d'acte	Exemple de frais réels	BR	TR	RSS	Euro forfaitaire
Consultation médecin généraliste conventionné secteur 1	25 €	25 €	70%	16,50 €	1 €
Monture optique tarif libre	135 €	0,05 €	60%	0,03 €	0 €
Verre blanc simple tarif libre	143 €	0,05 €	60%	0,03 €	0 €
Prothèses auditives adulte tarif libre	1500 €	400 €	60%	240 €	0 €

TABLE 1.1 – Exemples de remboursement de la Sécurité sociale

1. La participation forfaitaire doit être versée : pour toute consultation de médecin généraliste ou spécialiste ; lors d'examens de radiologie ; lors d'analyses de biologie médicale.

2. Le schéma 1.2 comprend une participation forfaitaire dans le cas d'une consultation généraliste. Celle-ci n'apparaît pas dans la décomposition du remboursement d'un acte autre que : Consultation généraliste, consultation spécialiste, examen de radiologie, analyse de biologie médicale.

Dans cet exemple, la consultation d'un médecin généraliste tarifée à la BR aura donc un prix réel de 25 €. Ce montant sera pris en charge à 70% par l'assurance maladie, hors participation forfaitaire de 1 €. Soit un RSS de 16,50 €. Les 30% restants, soit 7,50 €, correspondent au ticket modérateur et sont, dans la majorité des cas, pris en charge par la complémentaire santé. Toutefois, certains médecins peuvent pratiquer des dépassements d'honoraires. Le RSS et le TM resteront inchangés, et la complémentaire couvrira ou non ce dépassement. L'assuré pourrait dans cette situation faire face à un RAC.

Le tableau 1.1 montre que les remboursements de la Sécurité sociale sont nettement plus faibles pour les soins optiques, qui sont pourtant parmi les plus coûteux et les plus consommés. C'est pour cela qu'Emmanuel Macron a lancé en 2019 la réforme 100 % Santé, qui vise à rendre le reste à charge nul pour une certaine gamme de soins optiques, auditifs et dentaires.

1.4 Contexte réglementaire en assurance santé

L'assurance santé est un secteur particulièrement réglementé et régulièrement réformé ce qui rend le marché de la santé très évolutif. Notamment en France où une importance considérable est donnée à l'accès aux soins, étant le pays de l'OCDE [3] où le reste à charge moyen est le plus faible.

1.4.1 La loi Evin

L'article 4 de la Loi Evin a été mis en place le 31 Décembre 1989 et concerne les contrat de complémentaire santé collective. Cet article impose aux organismes assureurs de maintenir, selon un tarif encadré, la couverture frais de soins de santé à :

— Les anciens salariés bénéficiaires d'une rente d'incapacité ou d'invalidité, d'une pension de retraite ou, s'ils sont privés d'emploi, d'un revenu de remplacement, sans condition de durée ;
— Les bénéficiaires dont le chef de famille est décédé, pendant une durée minimale de douze mois à compter du décès.

Les tarifs applicables aux personnes visées peuvent être supérieurs — mais encadrés selon l'ancienneté du départ — aux tarifs globaux applicables aux salariés actifs. Depuis le 1er Juillet 2017, les tarifs sont plafonnés selon les modalités suivantes :

— La première année, les tarifs ne peuvent être supérieurs aux tarifs globaux applicables aux salariés actifs de l'entreprise ;
— La deuxième année, les tarifs ne peuvent être supérieurs de plus de 25% aux tarifs des salariés actifs de l'entreprise ;
— La troisième année, les tarifs ne peuvent être supérieurs de plus de 50% aux tarifs des salariés actifs de l'entreprise.

Au delà de la troisième année, les tarifs ne sont plus plafonnés.

1.4.2 L'Accord National Interprofessionnel

Depuis 2008, l'Accord National Interprofessionnel (ANI) permet à un salarié quittant son entreprise de bénéficier, selon certaines conditions, du maintien des garanties de la couverture complémentaire santé à titre gratuit pendant une durée limitée. C'est le principe de **portabilité**.

Depuis 2013, l'ANI permet aux salariés de bénéficier du principe de portabilité pour une durée égale à la durée passée dans l'entreprise dans une limite de 12 mois. De plus, depuis 2014,

3. OCDE : Organisation de coopération et de développement économiques

l'accord impose un financement par mutualisation. C'est à dire que le coût de la couverture d'un assuré à titre gratuit va être reporté sur l'ensemble des salariés encore en activité.

Les conditions pour bénéficier de cette portabilité sont les suivantes :

— Être victime d'une rupture de contrat de travail pour un motif autre que la faute lourde ;
— Être éligible à l'assurance chômage ;
— Avoir adhéré à un contrat complémentaire santé collective proposé par l'entreprise ;
— Avoir travaillé au moins un mois au sein de l'entreprise.

A titre d'exemple : si un salarié respectant ces conditions et ayant travaillé 10 mois dans une entreprise de 12 employés payant une couverture Santé collective à 100 €, fait l'objet d'une rupture de contrat, il conservera ses garanties pendant 10 mois à titre gratuit pendant que les 11 autres employés continueront de payer 100 €.

La figure 1.3 illustre le la durée de portabilité des droits en fonction de la durée de présence dans l'entreprise.

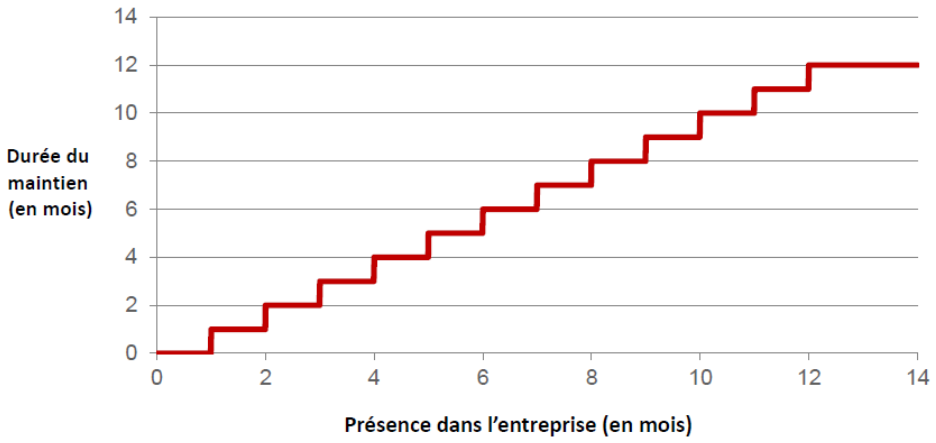

FIGURE 1.3 – Durée de maintien de portabilité des droits [Caritat, 2-3 Nov. 2020] [4]

Cet accord a augmenté le nombre d'assurés couverts par un contrat d'assurance Santé collective de 5% entre 2014 et 2017, et de 3% le nombre de chômeurs bénéficiant d'une couverture complémentaire santé [5]. Ces hausses correspondent surtout au passage d'un contrat individuel à un contrat collectif. Cette réglementation pousse donc les assurances complémentaires santé à affiner leurs modèles de tarification.

1.4.3 Les contrats responsables

Le **parcours de soins coordonnés** est entré en vigueur suite à la Loi du 13 Août 2004 [6] avec pour objectif de réduire l'aléa moral, c'est-à-dire inciter l'assuré à ne pas abuser de ses garanties Santé. Cette loi impose aux assurés le respect d'un parcours de soins coordonnés (consulter son médecin traitant avant de prendre rendez-vous avec un autre généraliste ou un spécialiste).

Le **contrat responsable** est entré en vigueur en 2006, et consiste à inciter les assurés à respecter le parcours de soins coordonnés s'ils veulent être mieux remboursés.

Depuis 2014 [7], le contrat responsable doit respecter un cahier des charges comprenant des

4. A. Stephan, S. Lazic, Tarification et suivi d'un régime frais de soins, Caritat, 2-3 Nov. 2020, s. 28
5. Source : DREES
6. Source : UGIPS, « Les contrats responsables et le parcours de soins, les évolutions... »
7. Décret n°2014-1374 du 18 novembre 2014

garanties minimales et maximales applicables aux postes suivants : Optique, auditif, dépassement d'honoraires pour des médecins non adhérents à l'option pratique tarifaire maîtrisée (OPTAM) [8].

Le cahier des charges des contrats responsables est régulièrement révisé. Depuis 2019, un contrat de complémentaire santé responsable doit respecter les conditions suivantes :

- Remboursement intégral du ticket modérateur pour tout acte ou soin à l'exception des cures thermales, de l'homéopathie et de certains médicaments [9] ;
- Remboursement intégral du forfait hospitalier journalier, quelle que soit la durée d'hospitalisation ;
- Prise en charge de l'intégralité du reste à charge pour les équipements « 100% Santé » ;
- Prise en charge des dépassements d'honoraires à hauteur de 100% de la BR pour les médecins non adhérents à l'OPTAM à condition que ce remboursement soit inférieur d'au moins 20% au remboursement susceptible d'être perçu par un médecin adhérent à l'OPTAM. Et prise en charge des dépassements d'honoraires sans limite pour les médecins adhérents à l'OPTAM.

Un contrat Santé responsable offre plusieurs avantages fiscaux et sociaux à l'employé et l'employeur :

- Exonération des charges sociales de la contribution employeur ;
- Déductibilité de la cotisation patronale des impôts sur les sociétés ;
- Réduction du taux de la Taxe de Solidarité Additionnelle [10] (TSA) passant à 13,27% du montant des cotisations, contre 20,27% pour les contrats non-responsables ;
- Cotisation salariale déductible des impôts sur le revenu pour les salariés.

L'arrivée des contrats responsables a été un facteur pondérant dans la révision de la tarification des contrats complémentaires santé collective puisqu'ils ont instauré la distinction des médecins adhérants à l'OPTAM, qui ont signé un contrat limitant les dépassements d'honoraires, des médecins non adhérents à l'OPTAM.

1.4.4 La réforme 100% Santé

Lancée par Emmanuel Macron en 2019, la réforme 100% Santé a pour objectif de réduire le renoncement aux soins en proposant des paniers de soins dentaires, optiques et auditifs pris en charge à 100% par l'Assurance Maladie et la complémentaire santé. Cette réforme s'est déployée progressivement depuis sa lancée pour être finalisée le 1er Janvier 2021.

Les domaines du dentaire, de l'optique, et de l'auditif, bien qu'étant parmi les plus consommés, sont particulièrement coûteux, et souvent trop faiblement pris en charge par la Sécurité sociale. Ce qui rendait le reste à charge trop élevé et donc le soin difficile d'accès pour les ménages les plus modestes. En effet, en 2017, le renoncement aux soins est évalué à : 10,1% en dentaire, 16,8% en optique, et 65% en audiologie. Ces renoncements sont souvent dûs à des raisons financières.

Les prix étant fixés librement, et parfois bien supérieurs à la base de remboursement, les complémentaires santé sont à présent dans l'obligation de proposer des paniers de soins et équipements de qualité sans reste à charge pour l'assuré.

8. Un médecin adhérant à l'OPTAM est un médecin ayant signé une convention avec l'Assurance Maladie limitant les dépassements d'honoraires.

9. Les contrats responsables ne concernent pas les médicaments dont le service médical rendu a été classé faible ou modéré. Il s'agit des médicaments remboursés à 15% ou 30% par l'Assurance Maladie.

10. La TSA est applicable à tout contrat de complémentaire santé. Elle finance d'une part la couverture maladie universelle (CMU) et l'URSSAF de l'autre. Elle est exprimée en pourcentage des cotisations nettes émises par les assurés.

Les garanties optique

Pour le poste optique, qui concerne les lunettes de vue (monture et verres), deux paniers d'équipements sont mis en place depuis le 1er Janvier 2020 :

— Le panier 100% Santé, ou **Classe A**, qui propose une gamme de 34 montures adultes et 20 montures enfants, d'un montant inférieur ou égal à 30€, et des verres répondant aux besoins de l'assuré à des prix plafonnés. De plus, les BR ont été revalorisées pour les produits de ce panier ;

— le Panier Libre, ou **Classe B**, dont les tarifs sont librement choisis par l'opticien. La prise en charge des montures par la complémentaire santé est plafonnée à 100€ dans le cadre du contrat responsable, et la BR est de 0,05€ pour les équipements de ce panier. Dans le cadre d'un contrat non-responsable, le remboursement n'est pas plafonné et peut être supérieur à 100 €.

Ainsi, si l'assuré choisit un équipement du panier 100% Santé, il aura un reste à charge nul mais une gamme limitée de produits à disposition. S'il souhaite bénéficier d'un choix plus large ou de meilleure qualité, il prendra un équipement du Panier Libre mais s'expose à un reste à charge potentiellement élevé. Par ailleurs, l'assuré peut aussi choisir de prendre des verres de la Classe A, et une monture de la Classe B ou l'inverse.

Les garanties auditives

Le poste audiologie, qui concerne les prothèses et autres équipements auditifs, propose depuis le 1er Janvier 2021 deux paniers de soins auditifs :

— Le Panier 100% Santé, ou **Classe 1**, qui propose une gamme d'appareils auditifs dont les prix de vente sont limités à : 1400 € avec une base de remboursement au même montant pour les moins de 20 ans, et 950 € avec une base de remboursement de 400 € pour les plus de 20 ans. Ainsi l'assuré n'a aucun reste à charge pour ce panier ;

— Le Panier Libre, ou **Classe 2**, qui propose des produits aux tarifs libres avec un dispositif de contrôle des prix, et un plafond de remboursement de 1700 € par oreille pour un contrat responsable.

Pour les deux paniers, les équipements sont renouvelables tous les 4 ans minimum. Jusqu'à présent le reste à charge sur les appareils était généralement extrêmement élevé, ce qui obligeait beaucoup d'assurés à renoncer à s'équiper. Avec l'instauration d'un panier proposant désormais des équipements de qualité intégralement remboursés, le renoncement aux soins devrait significativement reculer sur ce poste, l'argument financier n'étant alors plus un sujet.

Les garanties dentaires

Pour le poste dentaire, qui concerne les prothèses dentaires (couronnes, bridges, et dentiers) trois paniers sont en place depuis le 1er Janvier 2021 :

— Le panier « 100% Santé » qui propose une gamme de prothèses répondant aux besoins buccaux-dentaires essentiels sans reste à charge pour l'assuré ;

— Le panier « Tarifs Maîtrisés » pour lequel le reste à charge n'est pas nul, mais sera maîtrisé notamment par une limitation des prix de vente ;

— Le panier « Tarifs Libres » qui propose une gamme de prothèses aux matériaux plus qualitatifs et/ou esthétiques avec prix fixés librement et un reste à charge variable.

Les soins buccaux-dentaires, tels que le traitement des caries ou détartrage par exemple, n'entrent pas dans le cadre de la réforme du 100% Santé, n'étant généralement pas sujets à un reste à charge trop élevé.

La réforme a eu pour effet de modifier le cahier des charges des contrats responsables. Depuis le 1er Janvier 2020, tous les contrats responsables doivent proposer les paniers « 100% Santé » pour les postes optique, auditif et dentaire.

La tarification Santé nécessitant une distinction de chaque garantie au vu de la différence de risque, la réforme 100% Santé pourrait influer de manière conséquente les modèles de tarification.

1.5 La tarification Santé

Pour qu'un assureur accepte de garantir la couverture des conséquences financières d'un aléa, il doit être capable de tarifer le risque associé. Le marché de l'assurance santé étant particulièrement compétitif, les acteurs cherchent à proposer des contrats aux garanties de plus en plus ciblées. La tarification santé consiste à ajuster le tarif d'un contrat d'assurance complémentaire santé de sorte qu'il représente au mieux le risque santé de l'assuré.

1.5.1 Le risque santé

En assurance, le risque santé représente le risque pour un assureur qu'un assuré effectue des dépenses pour un acte médical couvert par la complémentaire santé.

A la différence d'autres secteurs de l'assurance, comme l'assurance automobile où le risque de sinistre est généralement très proche du risque d'accident (puisqu'un assuré ayant eu un accident de voiture fait presque systématiquement appel à son assurance), le risque santé ne représente pas nécessairement le risque pour un assuré de tomber malade, mais de bénéficier d'une prestation de sa complémentaire santé.

En assurance santé, on parle généralement de risque court. C'est-à-dire qu'il y a un délai relativement faible entre l'acte médical et le versement de la prestation par la complémentaire (généralement moins de trois mois), contrairement à l'assurance automobile où un sinistre peut être réglé par l'assureur plusieurs années après la déclaration de celui-ci.

Le risque santé doit tenir compte d'un facteur conséquent qui est l'**aléa moral**[11]. En effet, une personne ayant souscrit à un contrat d'assurance complémentaire santé peut, une fois assurée, augmenter considérablement ses dépenses de santé. En outre, l'aléa moral est particulièrement présent en assurance santé. En effet, les personnes bénéficiant d'une couverture santé ont des dépenses de santé bien plus élevées que les personnes non-assurées.

De plus, l'assurance complémentaire santé, et notamment en santé collective, est confrontée au risque d'**antisélection**[12]. C'est-à-dire le fait d'inciter à la souscription des personnes qui effectuent plus fréquemment des dépenses de santé, qu'on appelle des « mauvais risques ». Le coût de ces mauvais risques n'est alors pas en adéquation avec ce qu'ils versent via leur prime d'assurance.
De manière générale, une bonne tarification des contrats santé permet à l'assureur de se prémunir contre les risques associés.

1.5.2 Principe de la tarification en assurance

Comme dit plus tôt, le principe de la tarification vise à calculer la prime qu'un assuré devra verser à l'assureur pour pour se protéger financièrement contre un risque précis.

11. Aléa moral : Absence d'incitation à se prémunir contre un risque lorsque l'on est protégé contre celui-ci.
12. En assurance, l'antisélection désigne les situations dans lesquelles une compagnie accorde une couverture d'assurance à un candidat dont le risque réel est sensiblement plus élevé que le risque connu par la compagnie d'assurance.

La tarification comprend deux parties : le calcul de la **prime pure**, et le calcul de la **prime commerciale**.

La prime pure est l'estimation de la charge totale de prestations (ou sinistres) que l'assureur devra verser à l'assuré. Le montant de ces prestations, alors inconnu, est considéré comme aléatoire.

Pour illustrer cela mathématiquement, considérons S la charge totale de prestations à régler par l'assureur. La prime pure P s'exprimera alors :

$$P = \mathbb{E}[S] \tag{1.1}$$

La démonstration de l'équation 1.1 ci-dessus se trouve en Annexe B.1.

Par ailleurs, la charge totale de prestations s'exprimera comme suit (voir équation 1.2) :

$$S = \sum_{i=1}^{N} X_i \tag{1.2}$$

Avec : N représentant le nombre de prestations versées, et X_i le montant de la i^{e} prestation.

Plusieurs modèles permettent d'estimer cette charge de prestations, notamment le modèle « Coût × Fréquence », sur lequel nous reviendrons plus tard.

La prime commerciale est alors la prime pure à laquelle sont ajoutés les chargements commerciaux puis fiscaux (ou taxes).

Les chargements commerciaux sont généralement exprimés en pourcentage de la prime pure, et comprennent :

- Les frais de gestion, qui couvrent les frais d'encaissement des primes ;
- Les frais d'administration, qui permettent de financer la gestion du contrat (charges salariales, matériels et autres frais fixes) ;
- Les frais d'acquisition, qui financent la communication marketing pour l'acquisition de nouveaux clients.

A ces chargements peut s'ajouter une marge de sécurité, qui permet de faire face à la volatilité naturelle des sinistres. Cette marge permet aussi de pallier le phénomène d'antisélection.

Les chargements fiscaux sont les taxes calculées sur la prime commerciale nette, qui varient selon la branche d'assurance. En santé, on parle de la Taxe de solidarité additionnelle (TSA), qui est de :

- 13,27% pour un contrat responsable ;
- 20,27% pour un contrat non-responsable ;
- 6,27% pour un contrat du régime Agricole responsable ;
- 20,27% pour un contrat du régime Agricole non-responsable ;

La figure 1.4 illustre la décomposition de la prime commerciale en ces différents chargements.

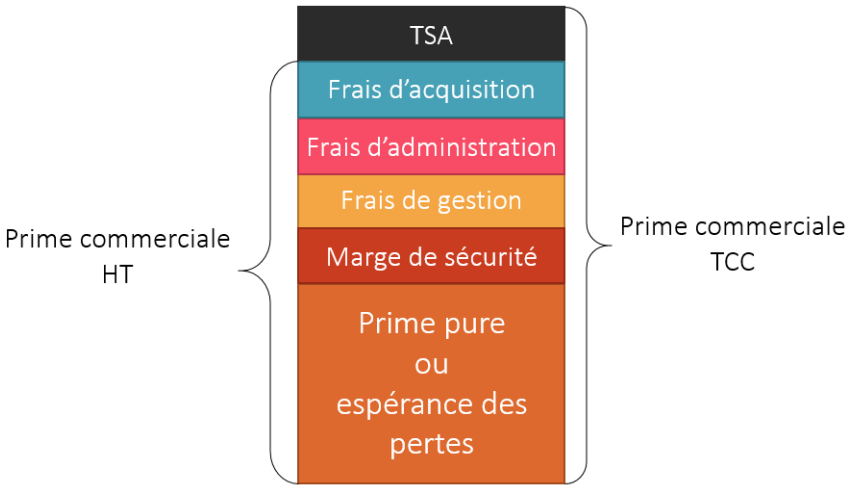

FIGURE 1.4 – Décomposition de la prime commerciale TTC

Ainsi, la prime commerciale nette peut s'écrire comme suit (voir équation 1.3) :

$$\text{Prime Commerciale HT} = \text{Prime Pure} \times \frac{1 + \text{Marge de Sécurité}}{1 - \sum \text{Frais}} \tag{1.3}$$

Et la prime commerciale brute s'écrit de la façon suivante (voir équation 1.4) :

$$\text{Prime Commerciale TTC} = \text{Prime Commerciale HT} \times (1 + \text{Taxes}) \tag{1.4}$$

L'assuré payera donc une prime commerciale brute à l'assureur pour bénéficier des garanties de son assurance.

1.5.3 La tarification en santé collective

En santé collective, l'actuaire tarifie selon le risque santé que représente un groupe, non une personne. Le principe de la tarification d'un contrat de complémentaire santé collective est de déterminer, par des modèles statistiques, les dépendances entre la consommation médicale et les caractéristiques de l'entreprise et des assurés, que l'on appelle **variables discriminantes**.

Les principales variables discriminantes sont les suivantes :

— **L'âge moyen** : L'âge est le facteur le plus discriminant en assurance santé. En effet, la fréquence de consommation augmente considérablement avec l'âge dans pratiquement tous les postes médicaux [13]. Dans le cadre d'un contrat collectif, l'âge moyen d'un groupe est donc généralement représentatif du risque de celui-ci.

— **La proportion femmes / hommes** : Si le Code des assurances interdit de proposer un tarif différencié en fonction du sexe de l'assuré [14], un assureur a tout de même le droit de calculer dans un premier temps un « tarif hommes » et un « tarif femmes », avant de proposer un tarif pondéré entre ces derniers en fonction de la proportion de femmes et hommes du groupe.

13. Le dentaire connaît un léger pic de consommation dans l'adolescence avant de reprendre un rythme croissant. Ce pic est principalement dû aux soins d'orthodontie.

14. Code des assurances, Article L111-7 : « Toute discrimination directe ou indirecte fondée sur la prise en compte du sexe comme facteur dans le calcul des primes et des prestations ayant pour effet des différences en matière de primes et de prestations est interdite. »

— **La catégorie socioprofessionnelle (CSP)** : D'une CSP à une autre, le niveau de vie est souvent significativement différent, et par extension, la propension à effectuer des dépenses de santé aussi. C'est donc une variable discriminante dans le cadre d'un contrat santé.
— **La zone géographique** : La présence médicale est très disparate selon les villes. Or, un assuré n'ayant pas de médecin a proximité de son domicile ou de son entreprise aura moins tendance à effectuer de dépenses médicales. Ce qui fait de la zone géographique une variable particulièrement corrélée à la fréquence de consommation.

Pour tarifer un contrat santé, il est nécessaire de distinguer chaque acte, les risques n'étant pas les mêmes. A titre d'exemple, une consultation de médecin généraliste n'a ni la même fréquence, ni le même ordre de coût moyen qu'une demande de remboursement de prothèse auditive. Il faut donc calculer une prime pure pour l'acte consultations généralistes, une prime pure pour l'acte prothèses auditives, etc. La prime pure individuelle sera la somme de ces primes calculées.

Les contrats d'assurance santé collective destinés aux salariés d'une entreprise sont généralement distingués selon la **structure de cotisation**, choisie par l'employeur. Les principales structures de cotisation sont les suivantes :

— **La cotisation familiale unique** : Lorsque l'employeur choisit ce type de tarification, la prime est unique quelle que soit la situation familiale du salarié (célibataire, marié, avec ou sans enfant). Cette structure de cotisation incite les salariés à affilier leur famille puisque le coût de la cotisation reste le même quel que soit le nombre de bénéficiaires.
— **La cotisation isolé/famille** : Dans cette structure, il y a deux types de cotisations possibles en fonction de la situation familiale du salarié :
  ◦ Si le salarié est célibataire sans enfant, il cotise pour un tarif isolé ;
  ◦ Si le salarié a des ayants droits (conjoint et/ou enfants), il a la possibilité de choisir un tarif famille qui le couvrira lui et ses ayants droits.

  Les salariés choisissant un tarif famille payeront le même tarif qu'importe le nombre d'ayants droits.
— **La cotisation adulte/enfant** : Cette structure permet de fournir un tarif par adulte et par enfant affilié. Il peut donc y avoir une multitude de tarifications différentes selon que le salarié soit célibataire sans enfant, marié sans enfant, marié avec 1 enfant, etc. Les tarifs augmentent en fonction du nombre de bénéficiaires.
— **La cotisation 1,2,3 personnes et +** : Cette structure contient trois tarifs :
  ◦ Un tarif isolé, pour les salariés célibataires sans enfant
  ◦ Un tarif 2 personnes, pour les salariés mariés sans enfant, ou célibataires avec 1 enfant.
  ◦ Un tarif 3 personnes et plus, pour les salariés souhaitant affilier plus de deux ayants droits.

  Les tarifs dépendent donc du nombre de personnes à affilier. A noter que cette structure ne distingue pas l'adulte de l'enfant.

La mise en place d'une structure de cotisation segmentée permet de réduire l'antisélection. En effet, la structure de cotisation familiale unique favorise les salariés ayant une famille à assurer puisqu'ils payeront le même tarif que les salariés isolés, bien que le niveau de risque soit considérablement plus élevé.

Dans le cadre d'un contrat collectif, les salariés appartenant à la même structure de cotisation payent la même prime. Une bonne estimation de la charge des sinistres permet une estimation plus juste de cette prime.

Conclusion

Les organismes d'assurance santé et les réformes réglementaires de ce marché permettent à une majorité des français de bénéficier d'un accès aux soins sans être freinés par le reste à charge. Toutefois, le renoncement aux soins n'est pas expliqué que par la contrainte financière. En effet, nous recensons en France aujourd'hui des zones dites « déserts médicaux » dans lesquelles les professionnels sont rares ou inexistants. L'absence de ces derniers à proximité empêche de nombreux assurés à consulter. Inversement, une zone ayant une importante offre médicale aura une consommation plus élevée. Ce qui fait de la zone un indicateur de consommation dans l'assurance santé. Pour s'en conforter, il est nécessaire d'avoir à disposition des données importantes et conséquentes.

Deuxième partie

Présentation et traitement des données

Chapitre 2

Présentation des données

Les données utilisées pour cette étude ont été fournies par un client d'Optimind proposant à ses salariés un contrat collectif de complémentaire santé. Pour des raisons de confidentialité, nous l'appellerons Entreprise E. Cette entreprise possède un grand nombre de salariés en France, auxquels elle propose une couverture santé d'entreprise garantie par un assureur.

2.1 La base de données bénéficiaires

L'effectif de la base de données bénéficiaires est composé des salariés de l'entreprise E ayant souscrit à un contrat d'assurance de leur entreprise, les retraités, les anciens salariés bénéficiant de la portabilité, ainsi que leurs ayants droits. Ces bénéficiaires sont couverts par un assureur. Cet assureur sera appelé Assureur A, pour des raisons de confidentialité.

2.1.1 Présentation des variables

La base de données bénéficiaires comprend au total 12 variables donnant des informations sur le bénéficiaire ou sur le contrat auquel il est rattaché. Les variables sont indiquées et décrites dans le tableau 2.1 ci-dessous.

Variable	Description
Reference.contrat	Code du contrat de travail
Contrat.Juridique	Contrat de travail
Libellé.Rg	Libellé de population (actif, retraité, etc.)
No.Assuré	Numéro attaché à la personne adhérant au contrat et payant la cotisation
Affiliation	Date d'affiliation au contrat
Sortie	Date de fin du contrat
Naissance	Date de naissance du bénéficiaire
Sexe	Sexe du bénéficiaire
Position	Position du bénéfiniaire (Assuré principal ou ayant droit)
Lien	Si ayant droit, lien avec l'assuré principal
Debut.Lien	Date à laquelle le bénéficiaire est devenu ayant droit
Dpt	Département de résidence

TABLE 2.1 – Variables de la base bénéficiaires

Lorsqu'un assuré principal affilie ses bénéficiaires à son contrat, ces derniers se voient attribuer le même numéro d'assuré. En conséquence, il existe un unique numéro d'assuré pour un assuré et ses ayants droits.

A noter que l'entreprise E regroupe ses salariés dans trois catégories :

— La catégorie A réunit tous les salariés actifs, leurs enfants, et leur conjoint si le revenu imposable mensuel de ces derniers est inférieur au seuil, qui est de 8 fois le SMIC brut mensuel. Elle regroupe également les enfants d'actifs décédés, qui deviennent alors cotisants. La structure de cotisation est uniforme à adhésion obligatoire [1].

— La catégorie B concerne les retraités et invalides de première catégorie affiliés au contrat, et leurs ayants droits si le revenu imposable de ces derniers est inférieur au seuil. La structure de cotisation est uniforme à adhésion facultative [2].

— La catégorie C représente donc les autres : les conjoint dont le revenu imposable est supérieur au seuil (le cas échéant, ces derniers sont considérés comme assurés principaux et sont cotisants) que le salarié soit actif, retraité, ou décédé ; les enfants cotisants dont le parent anciennement actif est décédé ; les chômeurs ; les agents détachés. La structure de cotisation est Adulte/Enfant à adhésion facultative [3].

La variable *Libellé.Rg* contient 20 modalités, et permet de retrouver la catégorie à laquelle appartient l'assuré. Les modalités et correspondances de la variable *Libellé.Rg* sont indiquées dans le tableau 2.2 ci-dessous.

1. Cotisation indépendante du nombre d'ayants droits. Ces derniers doivent être affiliés sauf dispense. Plus d'informations en partie 1.5.3.

2. Cotisation indépendance du nombre d'ayants droits. L'affiliation de ces derniers n'est pas obligatoire.

3. La cotisation diffère en fonction du nombre d'adultes et du nombre d'enfants de la famille. L'affiliation de ces derniers n'est pas obligatoire. Plus d'informations en partie 1.5.3.

Libellé	Description	Catégorie
E ACTIF+RET ANTIC	Salarié actif	A
E ENFANT/PETIT ENF	Enfant ou petit enfant au revenu supérieur au seuil	A
E INVALIDE GID	Invalide de première catégorie actif	A
E SALAR EN CGE PAR	Actif en congés parentales	A
CJ SURV+55 D'AC<SL	Conjoint d'actif décédé, de plus de 55 ans	A
CJ SURV-55 ACTIF<SL	Conjoint d'actif décédé, de moins de 55 ans	A
ENF.ORPHEL GRATUIT	Enfant d'actif décédé	A
E INVALIDES 1ERE CATEGORIE	Invalide de première catégorie inactif	B
E RETR CATEGORIE B	Retraité	B
CJT SURV/RETR<SL	Conjoint de retraité décédé	B
E AGENT DETACHE	Agent détaché	C
E AGENTS ASSEDIC	Chômeur	C
E CJ SURV ACTF COT	Conjoint d'actif décédé et cotisant	C
E CJT COT D'ACTF	Conjoint d'actif au revenu supérieur au seuil	C
E CJT COT DE RET	Conjoint de retraité au revenu supérieur au seuil	C
E CJT DIV COTIST	Conjoint divorcé au revenu supérieur au seuil	C
E SAL CSS	Actif en congé sans solde	C
CJ SURV DE RET COT	Conjoint de retraité décédé au revenu supérieur au seuil	C
ENF.ORPHEL COTISANT	Enfant d'actif décédé au revenu supérieur au seuil	C
ENF.ORPHEL COTISANT DE RETRAITE	Enfant de retraité décédé au revenu supérieur au seuil	C

TABLE 2.2 – Modalités de la variable *Libellé.Rg*

Cela permet ainsi de créer la variable *Catégorie* qui indique la catégorie à laquelle appartient l'assuré principal.

La première date de souscription disponible est le 1er Janvier 1995, et la dernière est le 31 Décembre 2019. Quant aux dates de sortie du contrat, la première est le 1er Janvier 2019, et la dernière le 31 Décembre 9999 lorsque celle-ci est indéfinie. C'est donc au 31 Décembre 2019, fin de l'année comptable, qu'il faut se placer pour connaître l'effectif total sur l'année 2019.

2.1.2 Retraitement de la base bénéficiaire

La base de bénéficiaires contient au total 62 680 lignes et ne présente pas d'anomalie apparente : aucune valeur aberrante sur les dates et aucune valeur manquante. Cependant, plusieurs points doivent être pris en compte pour obtenir l'effectif total.

Un certain nombre de bénéficiaires ont mis fin à leur contrat avant la fin de l'année comptable. Il ne font donc plus partie de l'effectif. Parmi ces bénéficiaires, certains ont mis fin à leur contrat pour souscrire à nouveau par la suite dans la même année. La raison principale est le change-

ment de contrat juridique qui oblige le bénéficiaire à mettre fin à son contrat pour souscrire à un autre. Dans d'autres cas, il peut simplement s'agir de décisions arbitraires des bénéficiaires. Ces bénéficiaires sont donc en double dans la base. Pour ne pas les compter plusieurs fois, il suffit de filtrer la base pour ne garder que les lignes dont la date de sortie du contrat est ultérieure au 31 Décembre 2019.

Certains bénéficiaires se sont affiliés en cours d'année, d'autres ont rompu leur contrat en cours d'année. Précisément, 731 bénéficiaires sont concernés. Ces derniers ont donc une exposition inférieure à 1 sur l'année. En d'autres termes, leur durée d'affiliation entre le 1er Janvier 2019 et le 31 Décembre 2019 est inférieur à 1 an. Grâce aux variables *Affiliation* et *Sortie*, on parvient à calculer l'exposition de chaque bénéficiaire.

La base contient 174 doublons. Ces doublons sont tous des bénéficiaires enfants de l'assuré principal. N'ayant pas de numéro individuel, il serait plausible de conclure qu'il s'agisse de jumeaux. En effet, ces doublons représentent 0,3% de la base, ce qui est cohérent au vu de la démographie gémellaire en France [4]. Pour cette raison, ces lignes sont donc conservées.

Enfin, 18 personnes n'ont pas renseigné leur genre. Ces personnes sont toutes des bénéficiaires enfants de l'assuré principal. Ces lignes seront retirées, elles n'interviendront pas dans les études démographiques à suivre.

Après retraitement, la base de données présente 58 549 lignes, chaque ligne correspondant cette fois à un bénéficiaire.

2.1.3 Analyse démographique

Pour mieux connaître le portefeuille que l'on étudie, il est intéressant d'effectuer des études statistiques. Comme dit plus tôt, la base contient au total 58 549 bénéficiaires au 31 Décembre 2019. Parmi eux, on retrouve :

— 36 551 assurés principaux
— 17 259 enfants
— 4 734 conjoints
— 5 autres (parents, ascendants...)

La plupart des assurés principaux sont des hommes (64%), et les conjoints sont majoritairement des femmes (97%). Le tableau 2.3 présente la répartition des sexes par position dans le portefeuille.

	Femmes	Hommes
Assurés principaux	13 301	23 250
Conjoints	4 596	138
Enfants	8 525	8733

Table 2.3 – Répartition des sexes

La composition familiale

Près de deux tiers des assurés principaux sont célibataires sans enfant à charge. Mais beaucoup de familles nombreuses sont également présentes dans le portefeuille. Cela s'explique par les

4. Ined (Institut national d'études démographiques) : « Evolution du nombre d'accouchements multiples ».

clauses du contrat qui obligent les assurés à affilier leurs ayants droits, à condition que ceux-ci ne soient pas déjà couverts par une autre complémentaire santé. Le tableau 2.4 présente la répartition des assurés principaux par composition familiale.

Composition familiale	Nombre	Pourcentage
Seul sans enfant	23 473	64%
Seul 1 enfant	3 093	8,5%
Seul 2 enfants	3 994	11%
Seul 3 enfants ou plus	1 256	3,4%
En couple sans enfant	3 566	10%
En couple 1 enfant	449	1,2%
En couple 2 enfants	460	1,2%
En couple 3 enfants ou plus	259	0,7%

TABLE 2.4 – Composition familiale

La famille la plus nombreuse est composée de 8 enfants et un conjoint. Au vu des dates de naissance, rien ne pousse à considérer cette observation comme une valeur aberrante.

La catégorie de salarié

Comme vu en partie 2.1.1, il existe trois catégories de salariés. Parmi les assurés principaux, on retrouve :

— 20 663 assurés de la catégorie A ;
— 11 951 assurés de la catégorie B ;
— 3 937 assurés de la catégorie C.

La figure 2.1 présente la répartition de cette catégorie parmi les assurés principaux, et parmi les bénéficiaires.

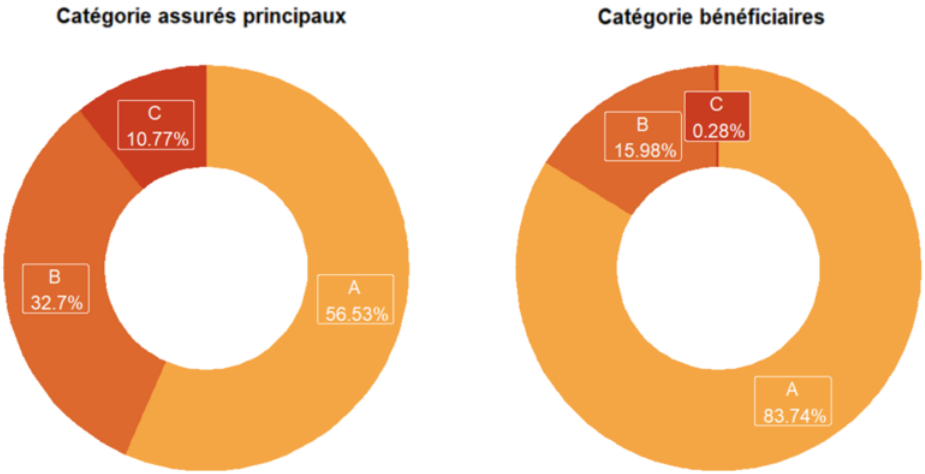

FIGURE 2.1 – Répartition des catégories pour les assurés principaux et pour les bénéficiaires

Une grande majorité des bénéficiaires est de la catégorie A. La plupart des bénéficiaires étant des enfants, ceux-ci se font plus rares parmi les ayants droits des retraités ou autres.

Les bénéficiaires de la catégorie C sont presque négligeables. Ils représentent les anciens salariés bénéficiant de la portabilité et les agents détachés, minoritaires dans un portefeuille santé d'entreprise.

L'âge

En se plaçant au 31 Décembre 2019, il est possible de créer une variable *Age* donnant l'âge du bénéficiaire grâce à sa date de naissance. Ainsi, l'âge moyen du portefeuille est de 45 ans au 31 Décembre 2019, mais la moyenne d'âge des assurés principaux est de 58 ans. Celle des conjoints est de 67 ans, et celle des enfants est de 12 ans. Mais cette moyenne varie selon les catégories, comme l'indique le tableau 2.5 ci-dessous.

Catégorie	A	B	C
Assurés principaux	43 ans	77 ans	77 ans
Conjoints	49 ans	75 ans	77 ans
Enfants	12 ans	28 ans	28 ans

TABLE 2.5 – Âge moyen par catégorie

La catégorie B est la plus âgée puisqu'elle contient en grande majorité des retraités. La catégorie C contient entre autres des assurés aux revenus importants donc sensiblement en fin de carrière, et des conjoints veufs, d'où l'âge moyen plus élevé qu'en catégorie A.

Il est aussi constaté que les conjoints de la catégorie A sont en moyenne plus âgés que les assurés principaux de la catégorie A. Cette statistique s'explique par le fait que la plupart des assurés de cette catégorie sont célibataires, comme vu au tableau 2.4. De plus, les assurés mariés sont globalement plus âgés.

La figure 2.2 ci-dessous présente la pyramide des âges de l'effectif global.

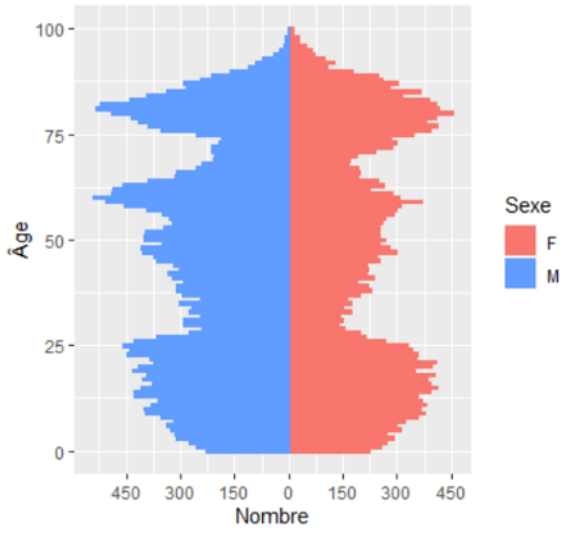

FIGURE 2.2 – Pyramide des âges de l'effectif global

Dans l'effectif, 12 313 bénéficiaires ont 75 ans ou plus, soit 21% de l'effectif total. Ce ratio est très supérieur à celui de la démographie française qui compte 9% de personnes ayant 75 ans ou plus en 2021 [5]. Dans l'ensemble, la population est assez âgée dans la mesure où les bénéficiaires de la catégorie B augmentent considérablement la moyenne d'âge.

Si on se restreint à la catégorie A, on constate un frein de l'exposition sur les âges avancés. La figure 2.3 présente la pyramide des âges de l'effectif de catégorie A.

5. Source : Population par sexe et groupe d'âges, INSEE, 2021

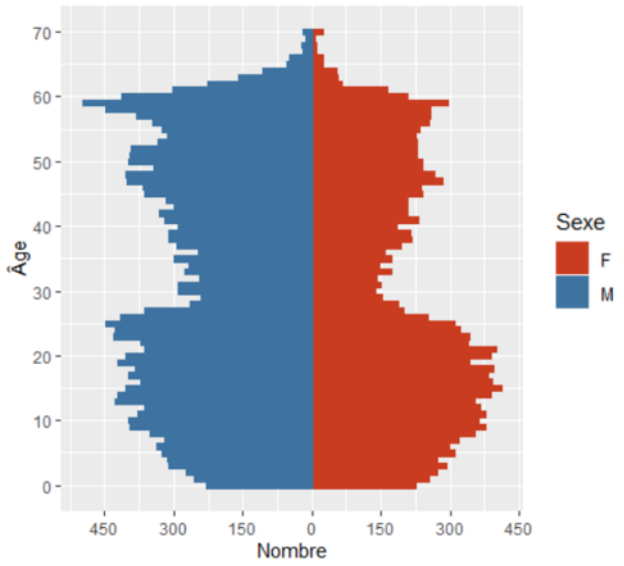

FIGURE 2.3 – Pyramide des âges de l'effectif des actifs

Dans cette catégorie, moins de 1% de la population a 75 ans ou plus, ce qui est cette fois très inférieur à la démographie française. On constate une nette chute de l'effectif à partir de 62 ans, âge légal de départ en retraite en 2019 [6].

Par ailleurs, comme dans la figure 2.2, on constate une chute de l'exposition autour de 25 ans. Il s'agit en réalité des enfants qui ne sont plus à charge des parents.

Répartition spatiale

L'entreprise E est implantée sur plusieurs sites en France :
— en Ile-de-France ;
— en Occitanie ;
— au Centre-Val de Loire ;
— en Bourgogne-Franche-Comté ;
— en Provence-Alpes-Côte d'Azur ;
— en Auvergne-Rhône-Alpes ;
— en Gironde ;
— en Nouvelle-Aquitaine.

Les bénéficiaires sont toutefois répartis dans toute la France, bien que la majorité se trouve en Essonne. La figure 2.4 représente la répartition des assurés et de leurs ayants droits à la maille départementale, en indiquant la présence ou non d'un site de l'entreprise.

6. Loi n°2010-1330 du 9 novembre 2010 portant réforme des retraites

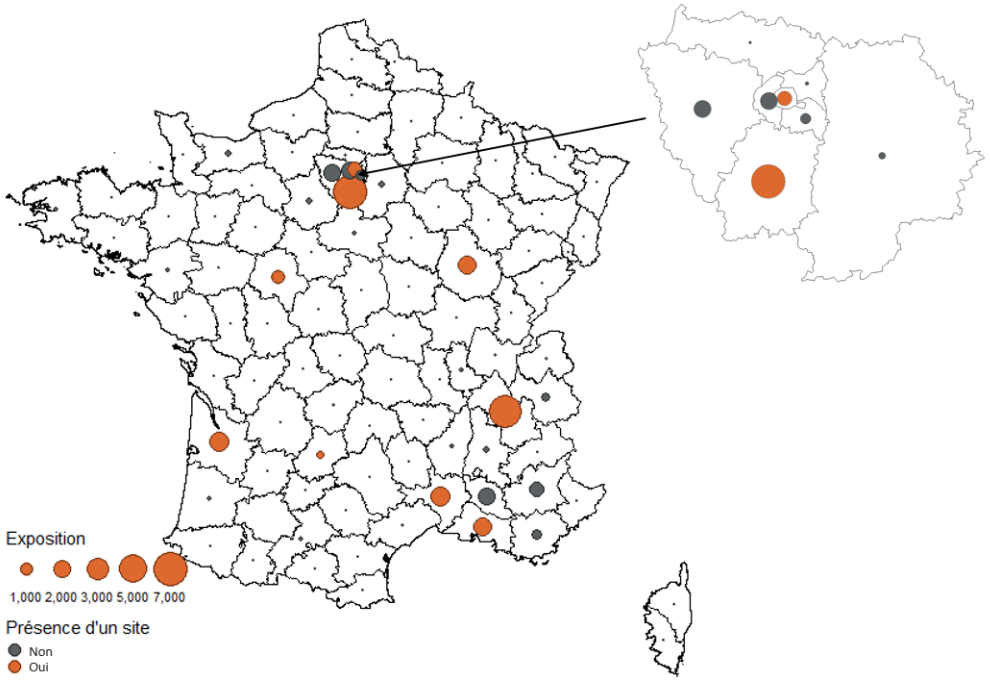

FIGURE 2.4 – Exposition par département

Si certains départements sont très faiblement représentés, 95 des 96 départements de l'hexagone sont occupés par au moins une personne.

Comme on peut s'y attendre, les bénéficiaires résident majoritairement dans les départements disposant d'un site, mais également dans les départements voisins. Le département le plus exposé étant l'Essone avec près de 7 000 assurés actifs et ayants droits, suivi de l'Isère avec 6 600 bénéficiaires. Le département le moins exposé de l'hexagone est le Gers, qui ne contient aucun bénéficiaire, suivi de la Haute-Garonne, qui n'en contient que 3.

Le portefeuille à disposition est donc assez disparate, avec des assurés aux caractéristiques plutôt discriminantes dans la mesure où les assurés principaux sont principalement des hommes, et les conjoints sont principalement des femmes, bien que cette caractéristique ne puisse pas être un indicateur tarifaire, conformément à l'article L.111-7 du Code des Assurances. D'autre part, le grand nombre de personnes âgées rend le groupe sensiblement risqué. Pour s'en conforter, il est nécessaire d'observer le comportement des assurés via les analyses statistiques de la base de prestations.

2.2 Données prestations

La base de données de prestation présente l'historique de tous les actes médicaux effectués en 2019 par les salariés de l'entreprise ayant adhéré à la complémentaire santé collective, ainsi que par leurs bénéficiaires.

2.2.1 Présentation des variables

La base prestations contient une vingtaine de variables. La description de ces variables est dans le tableau 2.6.

Variable	Description
Code.Acte	Code de l'acte effectué
Libellé.Acte	Libellé de l'acte effectué
Référence.contrat	Code du contrat de travail
Société	Site de l'entreprise
Population	Type de salarié
No.Pg	Numéro d'assuré
Sexe	Sexe de l'assuré
Age	Âge de l'assuré au moment de l'acte
Lien	Si bénéficiaire, lien avec l'assuré principal
Dpt	Numéro de département de résidence de l'assuré principal
Numero.finess	Numéro FINESS, RPPS ou ADELI
Acquittement	Date de versement de la prestation
Date.Acte	Date de l'acte
Regroupement	Famille d'acte
Quantite	Nombre d'actes effectués
Frais.réels	Frais réels de l'acte
Remb.Ss	Remboursement de la sécurité sociale
Mt.Rc	Remboursement de la complémentaire santé
Autre.Mut	Remboursement d'autre complémentaire
Reste.a.Charge	Reste à charge de l'assuré
Autres.Ct	Éventuels autres coûts
Taux.Ro	Taux de remboursement de la sécurité sociale

TABLE 2.6 – Variables de la base prestations

Cette base contient donc plusieurs variables en commun avec la base bénéficiaires :

— La variable *No.Pg* est équivalente à la variable *No.Assuré* de la base effectif ;
— La variable *Référence.contrat* est équivalente à la variable *Contrat.Juridique* de la base effectif ;
— Les variables *Sexe* et *Lien* sont les mêmes que dans la base effectif.

Les garanties du contrat couvrent 8 postes médicaux :

— Les consultations et visites, qui concernent entre autres les consultations généralistes et spécialistes ;
— L'hospitalisation, qui concerne les frais de séjours, les forfaits journaliers, etc.
— Les frais médicaux de ville, qui concernent les analyses de laboratoire, les sages femmes, la kinésithérapie, etc.
— La pharmacie ;
— Le dentaire, qui concerne les soins dentaires, l'orthodontie, etc.
— L'optique, qui concerne les montures, verres, lentilles, etc.
— L'appareillage, qui concerne les prothèses auditives ;
— Les forfaits cures thermales.

Le détail des actes couverts par le contrat de complémentaire santé se trouve en Annexe A.1.

Si le code postal du lieu de soin n'est pas donné, le numéro de département de résidence, ainsi que les numéros FINESS, RPPS et ADELI donnent une information géographique :

— Le numéro de fichier national des établissements sanitaires et sociaux (FINESS) permet d'identifier un établissement de santé (hopitaux, cliniques, cabinets médicaux, laboratoires et maisons de soins).
— Le numéro de répertoire partagé des professionnels de santé (RPPS) permet d'identifier certains professionnels de santé (médecins, dentistes, sages-femmes, pharmaciens, kinésithérapeutes, podologues).
— Le numéro d'automatisation des listes (ADELI) permet d'identifier les professionnels de santé qui ne sont pas dans le répertoire RPPS (infirmiers, assistants dentaire, psychologues, etc.).

Ces numéros sont composés de 9 chiffres dont les deux premiers représentent le département de l'acte effectué.

2.2.2 Restructuration et traitement de la base prestations

La base prestations contient au total 3 401 034 lignes, une ligne par prestation effectuée. La date comptable de ces prestations s'étend de 2019 à 2021. La majorité étant survenues en 2019, c'est sur cette année de survenance que vont s'effectuer les études à suivre. Ainsi, en se restreignant à l'année de survenance 2019 [7], la base fait à présent 3 356 120 lignes.

Par la suite, plusieurs points ont dû être pris en compte pour la préparation de la base. Certains ont été traités, d'autres omis.

Montants négatifs

Un certain nombre de montants de frais réels sont négatifs, précisément 1 216 lignes. La quantité, le remboursement de la complémentaire, le remboursement de la sécurité sociale, et le reste à charge le sont aussi dans ces cas précis. Il ne s'agit pas d'erreur mais d'annulations de versements. Ces derniers sont systématiquement compensés par le même montant en valeur absolue. Il n'est donc pas nécessaire de les retirer.

Consultations majorées

Le tableau de l'Annexe A.1 présente les actes sous différentes granularités. La maille la plus fine étant le libellé de l'acte. Certains libellés indiquent une majoration pour diverses raisons (jour férié, travail de nuit, etc.). Dans notre base, 63 361 lignes sont concernées par ces majorations. Les montants de ces majorations sont déduits des frais réels de l'acte effectué. A titre d'exemple, si un assuré a effectué une visite de nuit chez un médecin généraliste conventionné de secteur 1, majorée de 5 €, il aura deux lignes associées à cet acte : une ligne avec pour libellé d'acte « Consultation généraliste », et pour montant de frais réels 25 €, et une seconde avec pour libellé d'acte « Majoration nuit généraliste », et pour montant de frais réels 5 €. Cependant, la variable *Quantité*, qui dénombre les actes pour chaque ligne, comptera un acte pour chacune de ces deux lignes. Il convient donc de corriger cela en considérant que ces deux lignes totalisent un acte.

Variable Famille d'acte

La base ne contient pas de variable indiquant la famille d'acte (ou poste) à laquelle appartient l'acte effectué. Il a donc été nécessaire de la créer afin d'avoir une variable médicale moins granulaire pour les études statistiques. Pour cela, a été fourni un fichier *Transco actes* contenant

7. De Janvier 2019 à Janvier 2020.

les correspondances entre familles d'actes et actes. Il suffit de relier ce fichier à la base de données avec pour clé primaire la variable *Libellé.Acte* qui est la maille la plus fine de description d'un acte.

Information géographique manquante

Concernant la variable *Numero.finess*, 86 634 numéros n'ont pas été renseignés. Soit un peu moins de 3% de la base, ce qui n'est pas négligeable. Toutefois, seulement 2 675 départements de résidence ne sont pas renseignés, et dans les lignes concernées, 251 numéros FINESS n'apparaissent pas. En d'autres termes, 0.007% des lignes de la base ne contiennent pas d'information géographique. Ces dernières se verront attribuer un numéro de département simulé de telle sorte que la probabilité d'attribution d'un département suive la distribution des départements du portefeuille.

Cependant, considérer que le département du lieu de soin est également le département de résidence de l'assuré reste une hypothèse relativement forte. Typiquement, une cure thermale peut avoir lieu dans une zone très éloignée de la résidence. Dans notre base, 84% des actes effectués ont eu lieu dans le département de résidence de l'assuré.

Après retraitement de la base, celle-ci peut être analysée afin de connaître le comportement des bénéficiaires en fonction des postes, et de leurs caractéristiques.

2.2.3 Statistiques descriptives de la base prestations

Consommation par poste

Le poste *Pharmacie* est le plus consommé en termes de quantité d'actes en 2019, tandis que le poste *Hospitalisation* est le plus consommé en termes de frais réels et de remboursement complémentaire. Ceci est cohérent puisque les actes pharmaceutiques sont généralement à faibles coûts, et une ordonnance médicale peut prescrire plusieurs dizaines de médicaments, alors que l'hospitalisation, qui comprend par exemple les honoraires chirurgicales, peut se montrer bien plus coûteuse. Dans notre portefeuille, les frais moyens (Frais réels / Nombre d'actes) d'un acte pharmaceutique est de 3,45€, et celui d'un acte hospitalier est de 97€. Le poste *Cures termales* est quant à lui le moins consommé. Le tableau 2.7 présente le montant total des prestations par grand poste.

Poste	FR	RC	RSS	RAC	Quantité
Hospitalisation	16 228 060 €	7 370 975 €	7 931 504 €	917 075 €	166 458
Frais médicaux de ville	12 248 636 €	3 978 174 €	7 146 644 €	1 117 272 €	638 529
Pharmacie	8 951 110 €	3 928 547 €	5 011 193 €	11 365 €	2 593 892
Consultations et visites	8 919 384 €	2 475 146 €	4 648 073 €	1 783 946 €	288 582
Cures Thermales	420 352 €	417 781 €	0 €	2 570 €	2 437
Dentaire	12 678 251 €	6 721 485 €	3 041 062 €	2 809 676 €	99 823
Optique	8 164 681 €	5 554 565 €	194 387 €	2 378 860 €	44 710
Appareillage	5 697 008 €	2 800 147 €	1 646 207 €	1 241 292 €	57 138

TABLE 2.7 – Consommation médicale par poste

Comme on peut le voir, les restes à charge des postes *Dentaire, Optique* et *Appareillage* sont particulièrement élevés. En effet, ces données datant de 2019, la réforme 100% Santé est alors

récente (ou inexistante pour certains actes), et les soins de ces postes sont encore assez coûteux.

Si l'on regarde la répartition des frais médicaux, on voit que les actes pharmaceutiques sont bien davantage consommés mais finalement très peu coûteux, et vice versa pour l'hospitalisation. Plus de 60% des actes médicaux effectués sont des actes pharmaceutiques, mais ces actes occupent seulement 12% des dépenses médicales totales.

Remboursements des frais médicaux par poste

Comme vu plus tôt, les postes optique, dentaire et auditif ont un reste à charge relativement élevé au regard de la faible consommation en nombre d'acte. Pourtant, ces trois postes ont un ratio *Remboursement complémentaire/Frais réels* particulièrement conséquent. La figure 2.5 reporte la décomposition des frais réels en remboursements pour chaque poste.

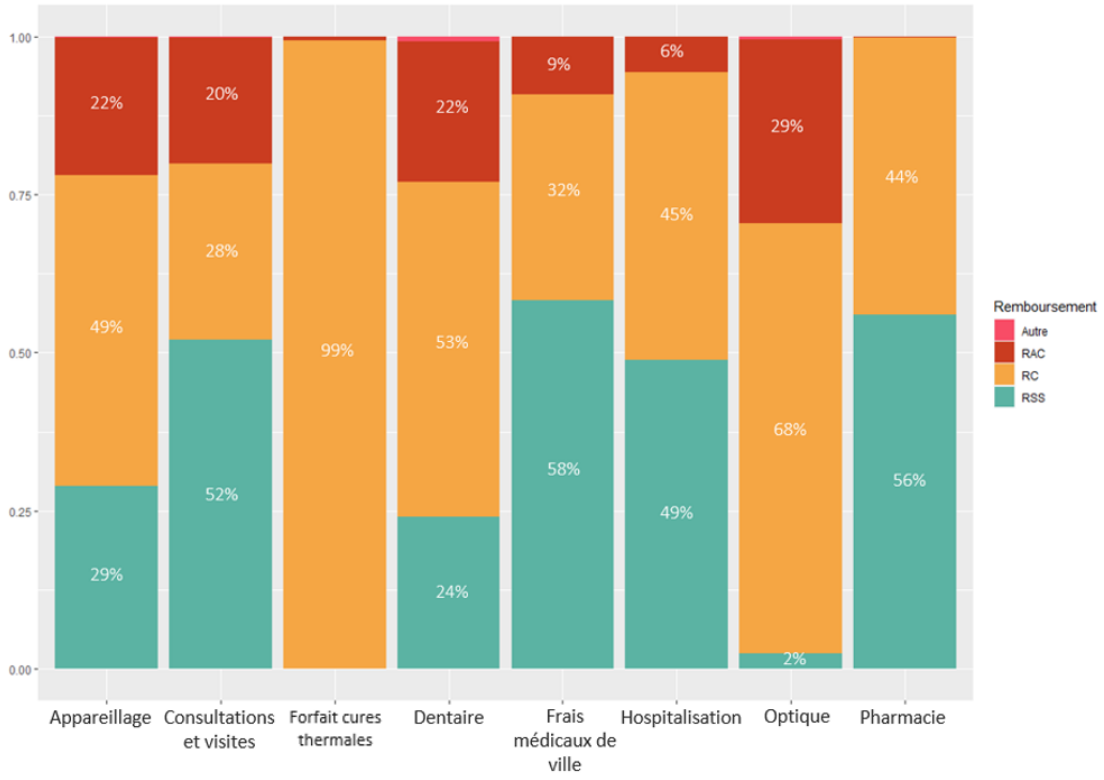

FIGURE 2.5 – Décomposition des frais réels par poste

Hors cure thermale, les postes optique, dentaire et appareillage ont les ratios *Remboursement complémentaire/Frais réels* les plus élevés. Cependant, ils ont également les ratios *Remboursement de Sécurité sociale/Frais réels* les plus faibles, ce qui veut dire que la complémentaire compense conséquemment le faible remboursement de la sécurité sociale. On constate aussi que le poste Pharmacie a un reste à charge presque négligeable (inférieur à 1%).

Consommation par catégorie

Au global, la catégorie B, qui ne représente qu'un tiers de l'effectif, est la plus dépensière avec 44% des frais réels totaux, contre 42% pour la catégorie A, et 14% pour la catégorie C.

Cela s'explique par un volume important d'actes hospitaliers, essentiellement consommés par les retraités. La figure 2.6 présente la répartition des frais réels par catégorie pour chaque poste.

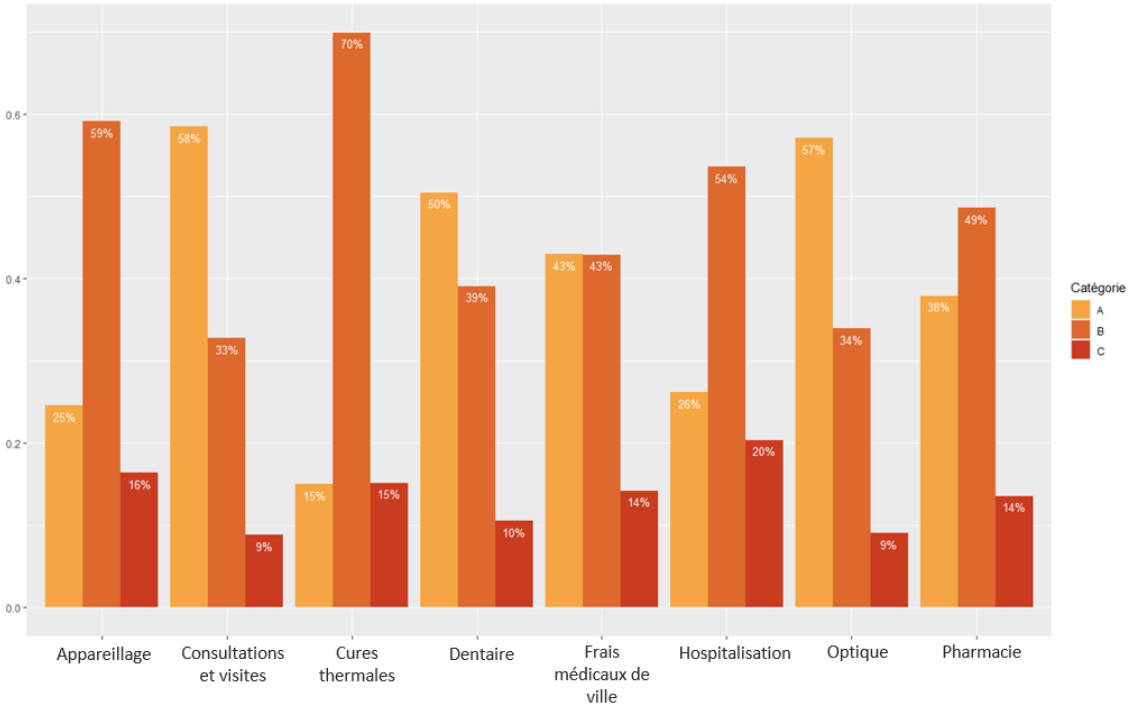

FIGURE 2.6 – Répartition des frais réels par catégorie pour chaque poste

On constate que les postes Appareillage, Cures Thermales, et Hospitalisation sont très largement consommés par les bénéficiaires de catégorie B. Ce sont des postes souvent destinés aux personnes âgées, nous le verrons dans la sous-partie suivante. L'optique et les consultations sont bien davantage consommés par les bénéficiaires de catégorie A, ce qui s'explique par la forte présence d'enfants dans la catégorie A, qui, comme vu dans la figure 2.7, consomment fortement ces deux actes.

Étant donné la disparité des besoins par catégorie, il convient donc de les distinguer une à une pour les études à venir dans ce mémoire. L'assurance santé collective vise en premier lieu à proposer un contrat de complémentaire santé aux salariés actifs d'une entreprise. Bien qu'étant moins consommatrice, la catégorie A étant la plus volumineuse et la plus homogène, les études se focaliseront sur cette dernière.

Ainsi, **dans la suite de cette partie, nous observerons les statistiques de consommation au sein de la catégorie A.**

Consommation par lien familial

On distingue trois modalités de liens : Les assurés principaux, les conjoints, et les enfants.

En raisonnant en termes de prime pure (*Coût moyen par personne* × *Fréquence de consommation par personne*), les conjoints consomment en moyenne davantage que les assurés principaux, environ 20% de plus. Le graphique 2.7 présente la prime pure par lien familial pour chaque poste.

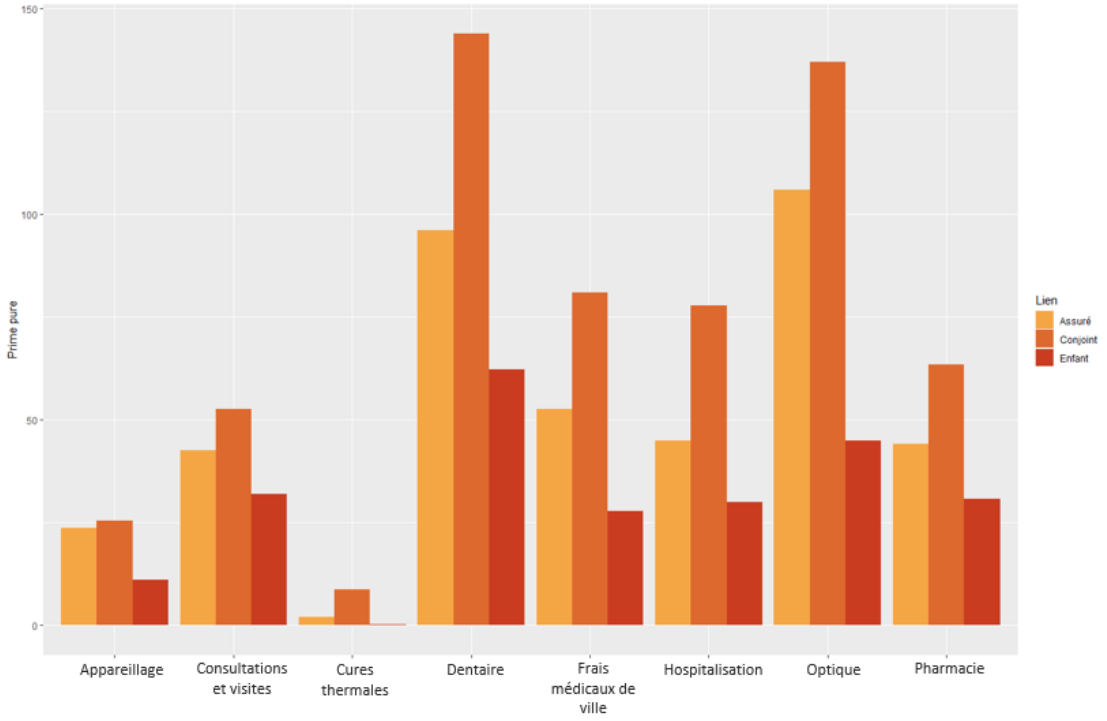

FIGURE 2.7 – Prime pure par lien familial pour chaque poste

Il apparaît clairement que les conjoints sont les plus consommateurs. Côté enfants, on observe une bien plus faible consommation. En moyenne, les enfants consomment près de 60% de moins que les assurés principaux. La santé étant généralement inversement proportionnel à l'âge, on peut considérer que les enfants, plus jeunes, sont en bonne santé, ce qui explique leur faible consommation d'actes médicaux.

Consommation par âge

Pour se faire une idée de la population étudiée, il est intéressant d'observer la fréquence de consommation par âge pour chaque poste. Dans un premier temps, en observant la fréquence de consommation globale, on observe une croissance assez claire en fonction de l'âge. Voir figure 2.8.

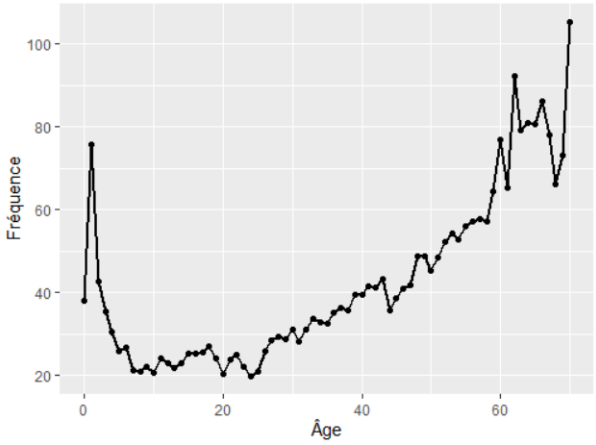

FIGURE 2.8 – Fréquence de consommation globale par âge

On voit que les assurés plus âgés consomment globalement davantage. Pourtant, si l'on distingue les postes, la fréquence n'est pas toujours croissante avec l'âge.

A titre d'exemple, les postes optique, appareillage, et dentaire sont assez volatiles en termes de fréquence de consommation par âge. La figure 2.9 illustre la fréquence de consommation par âge pour ces trois postes.

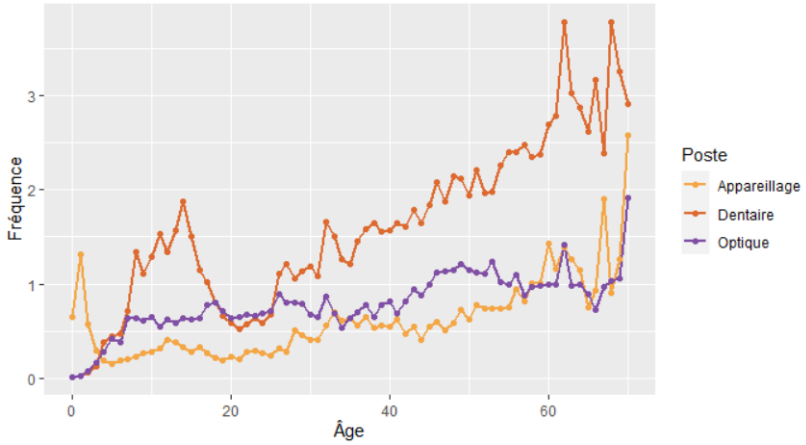

FIGURE 2.9 – Fréquence de consommation par âge pour les postes Optique, Appareillage et Dentaire

Le poste appareillage est bien croissant avec l'âge, ce qui peut s'expliquer par la presbyacousie (perte auditive) qui touche essentiellement les personnes âgées. La déficience auditive peut également concerner les nouveaux nés, d'où le pic entre 0 et 1 an pour ce poste.

Le poste dentaire présente une certaine tendance entre 13 et 16 ans, probablement dû à la consommation de traitements orthodontiques, qui ont souvent lieu dans l'adolescence.

Le poste optique quant à lui présente une importante consommation chez les quinquagénaires, période d'âge pendant laquelle arrive la presbytie.

Les postes Frais médicaux de ville et Pharmacie sont bien davantage consommés par les ménages. Ceux-ci sont les plus proches de la tendance globale, à savoir croissants avec l'âge. La figure 2.10 présente la fréquence de consommation par âge pour ces postes.

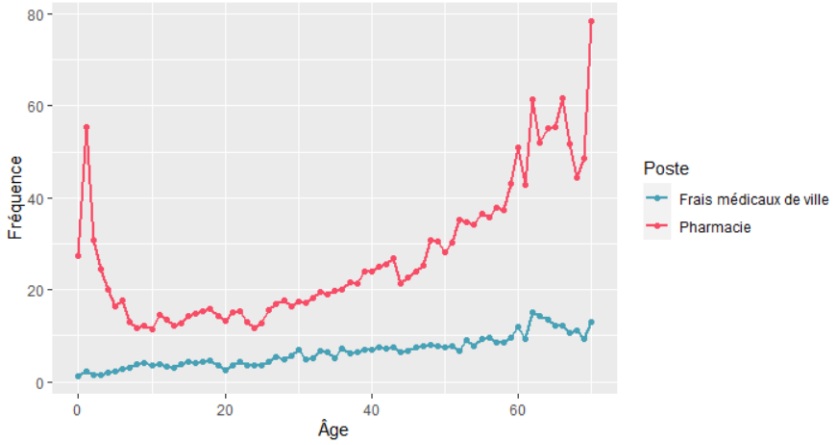

FIGURE 2.10 – Fréquence de consommation par âge pour les postes Frais médicaux de ville et Pharmacie

Le poste Pharmacie présente un pic dans les premiers âges qui peut s'expliquer par l'achat de produits de maternité. Par la suite, les besoins pharmaceutiques étant généralement fonction de l'âge, la courbe de consommation augmente avec l'âge pour ce poste.

Le poste Frais médicaux de ville reste relativement constant avec l'âge puisqu'il s'agit ici d'assurés actifs, peu âgés. Ils ne sont donc pas forcément sujets aux soins infirmiers et aux auxiliaires de vie, qui sont inclus dans ce poste.

Le poste Hospitalisation est également très fréquent, mais le risque associé à ce poste est davantage un risque de coût que de fréquence. La figure 2.11 présente la fréquence de consommation par âge pour les postes Hospitalisation et Consultations et visites.

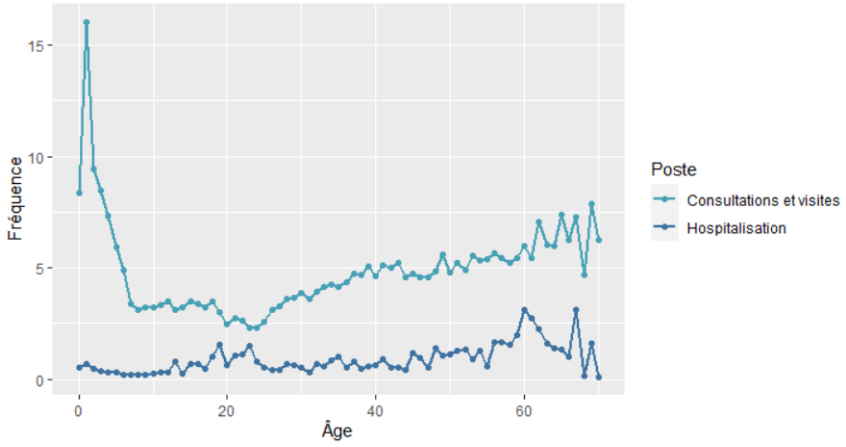

FIGURE 2.11 – Fréquence de consommation par âge pour les postes Consultations et Hospitalisation

Un pic de consommation important est observable dans le poste Consultations et visites chez les nourrissons, qui s'explique par le fait que ce poste inclut les actes de pédiatrie. Quant au poste Hospitalisation, il ne croît pas aussi fortement avec l'âge qu'on pouvait l'imaginer. Cela est dû d'une part à la jeunesse de l'effectif des actifs, d'autre part à leur milieu de travail. Il s'agit en grande majorité de chercheurs, ou en tout état de cause, de métiers aux risques physiques modérés.

Consommation par sexe

La consommation des assurés et ayants droits se distingue par leurs besoins. Ces besoins peuvent varier d'un sexe à l'autre. Notamment en assurance santé puisque certains soins peuvent être réservés à un genre (typiquement, la gynécologie ou l'obstétrique pour les femmes).

La figure 2.12 illustre le coût moyen (sous forme d'histogramme), ainsi que la fréquence de consommation (sous forme de courbe) par sexe pour chaque poste.

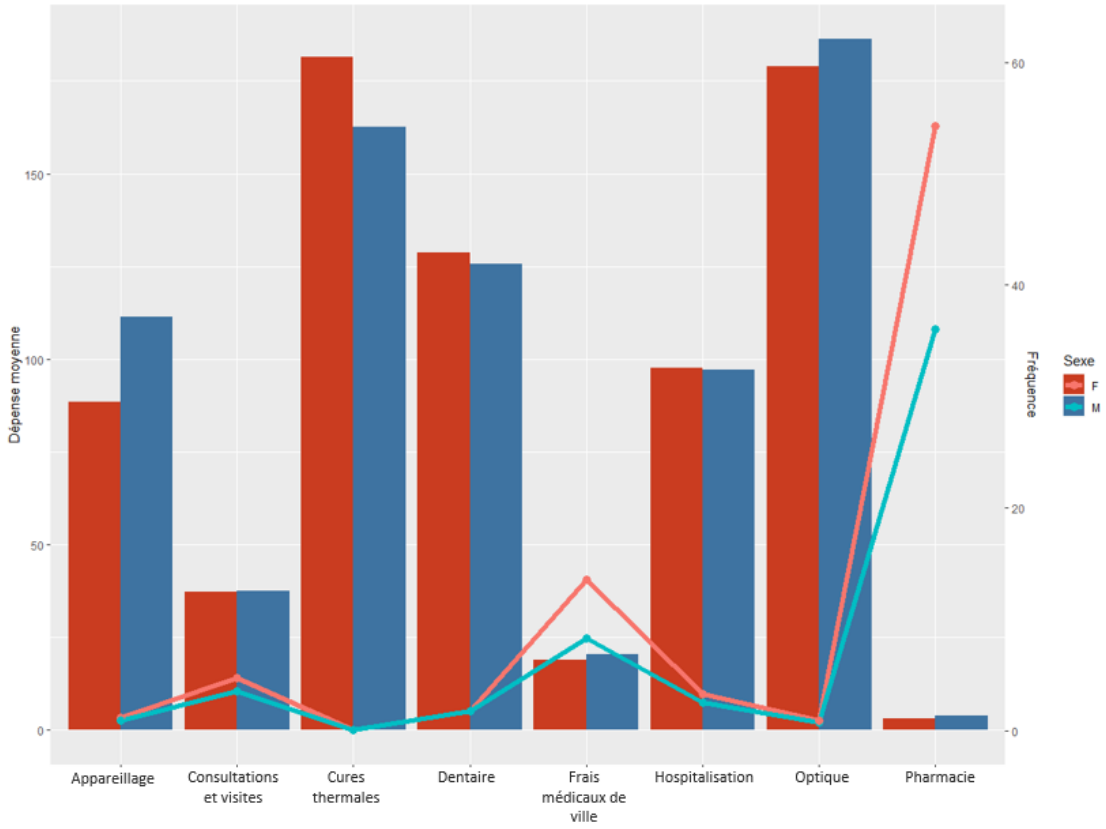

FIGURE 2.12 – Consommation par sexe pour chaque poste

Concernant les dépenses moyennes, peu de différences sont observables entre les femmes et les hommes. On constate par ailleurs que l'optique est le poste le plus cher en moyenne, et l'un des postes le moins consommé. En 2019, le 100% Santé n'étant pas encore finalisé pour les contrats responsables, on pourrait s'attendre à voir cette statistique évoluer dans les années futures.

En termes de fréquence, les femmes consomment en moyenne plus que les hommes.

Une différence est notable pour le poste Frais médicaux de ville, qui inclut les soins de sages femmes, réservés aux femmes.

Le poste Pharmacie est également nettement plus consommé par les femmes. Cette différence est difficilement explicable puisque les détails des médicaments prescrits ne sont pas donnés (cf : Annexe A.1). Toutefois, on peut supposer que les femmes ont plus de besoins pharmaceutiques spécifiques que les hommes (la contraception, l'infectiologie, les prescriptions post-grossesses, etc.), ce qui tend à expliquer cette différence.

Consommation par zone

La consommation par département de résidence peut donner une indication du risque spatial relatif à ce portefeuille et à cette catégorie d'assurés. Cependant, il convient de distinguer les actes, car la notion de risque peut différer. Pour certains actes d'un même poste, on peut parler de fréquence quand d'autres concernent le coût : typiquement, la médecine généraliste contre la médecine spécialiste.

Les figures 2.13 et 2.14 représentent les frais réels moyens [8] (en euros) et la fréquence moyenne [9] (en quantiles) par département pour l'acte *Consultations généralistes*.

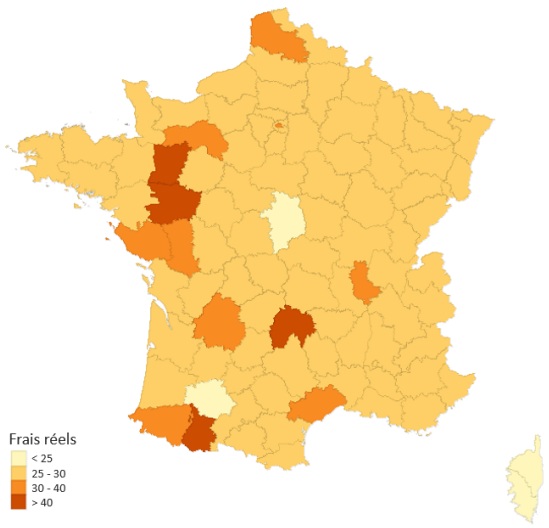

FIGURE 2.13 – Frais réels moyens par département pour *Consultations généralistes*

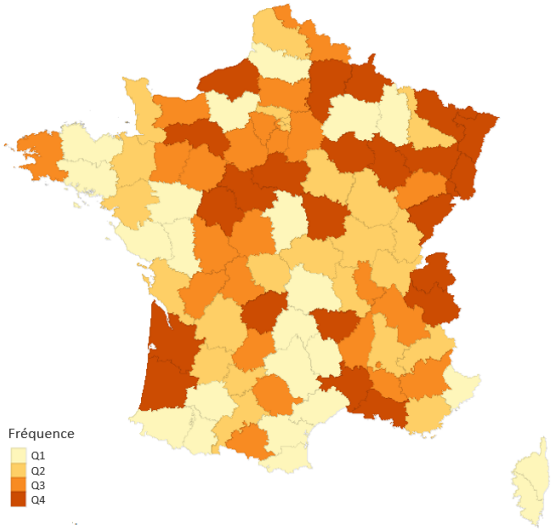

FIGURE 2.14 – Fréquence moyenne par département pour *Consultations généralistes*

Il s'avère que la plupart des départements ont pour frais réels moyens 25€, prix d'une consultation de médecin généraliste conventionné secteur 1. Les autres sont essentiellement des départements peu exposés dont un résident aurait consommé plus que la moyenne. Typiquement, le Cantal (Centre) ne dispose que de 5 assurés, dont un a effectué un acte de consultation généraliste conventionnée. Cette consultation a été majorée de 44€, ce qui explique que ce département ressorte davantage en frais réels moyens.

La fréquence, quant à elle, est assez disparate. Mais elle peut également être biaisée par des cas isolés qui consommeraient plus que la moyenne. A titre d'exemple, le Doubs (Centre-Est) contient 9 assurés dont un a effectué 15 consultations dans l'année, ce qui explique que ce département ressorte davantage en fréquence moyenne.

En comparaison, l'acte *Consultations spécialistes*, pourtant inclus dans le même poste que la consultation généraliste, représente un risque géographique relativement différent. Les figures 2.15 et 2.16 présentent les frais réels moyens et la fréquence moyenne par département pour l'acte *Consultations spécialistes*.

8. Pour un département i : $\textit{Frais réel moyen}_i = \frac{\sum \textit{Frais réels}_i}{\sum \textit{Nombre d'actes effectués dans le département } i}$
9. Pour un département i : $\textit{Fréquence}_i = \frac{\sum \textit{Nombre d'actes effectués dans le département } i}{\sum \textit{Exposition au sein du département } i}$

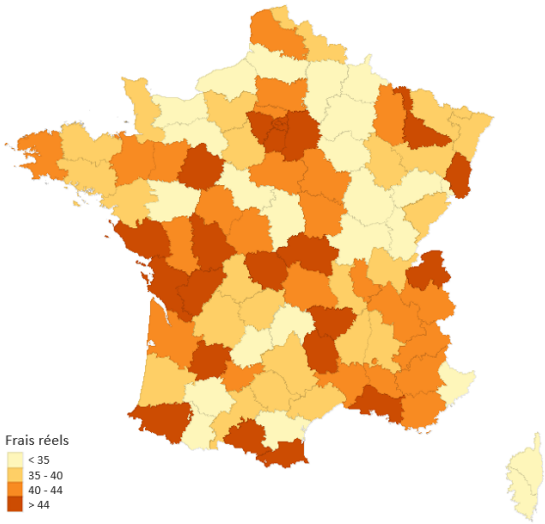

FIGURE 2.15 – Frais réels moyens par département pour *Consultations spécialistes*

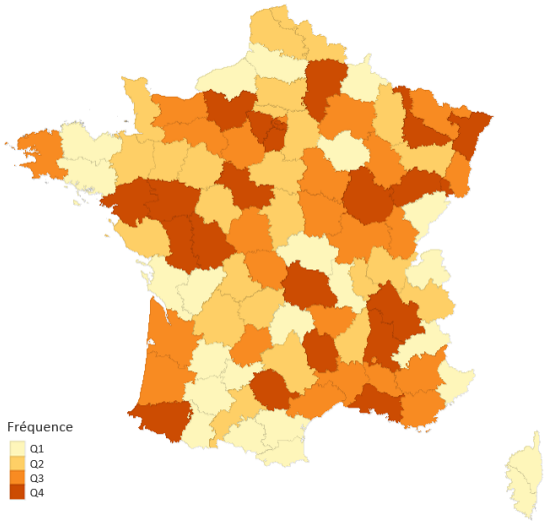

FIGURE 2.16 – Fréquence moyenne par département pour *Consultations spécialistes*

Pour cet acte, des départements urbains comme Paris ou les Bouches-du-Rhône ressortent fortement en termes de coût réel, mais aussi de fréquence. En comparaison avec l'acte *Consultations généralistes*, on voit que l'impact du risque spatial diffère d'un acte à l'autre, y compris lorsque ces actes appartiennent au même poste.

2.2.4 Création d'une variable *Code postal*

Dans notre base prestation, un grand nombre de numéros FINESS, RPPS et ADELI renseignés ne sont pas présents dans l'annuaire santé ni dans l'annuaire FINESS, plus précisément 90% de la base. Une des raisons pourrait être l'obsolescence de ces numéros. En effet, une réforme actuellement en place vise à faire basculer les professionnels disposant d'un numéro ADELI dans la base RPPS [10], les numéros ADELI deviendraient alors obsolètes. Une autre raison pourrait être la confidentialité des informations personnelles. En conséquence, la maille géographique la plus fine à disposition est le département de résidence de l'assuré principal, et dans le cas contraire le département dans lequel a été effectué l'acte.

Nous verrons plus tard que cette maille, bien que discriminante, peut être discutable pour discriminer une zone.

Ainsi, les codes postaux ont dû être simulés pour pouvoir aller au bout de notre étude. Ces derniers sont simulés pour chaque ligne en fonction du département de résidence, présent dans la variable *Dpt*, lorsque celui-ci est donné, ou dans la variable *Numero.finess* dans le cas contraire.

La simulation est faite en récupérant les codes postaux d'une base INSEE regroupant l'ensemble de la population par commune [11], de telle manière que la distribution des codes INSEE attribués soit en adéquation avec la population communale pour chaque département. A titre d'exemple, si un assuré a effectué un acte dans les Bouches-du-Rhône (13), département dont 7% de la population réside à Aix-en-Provence, cet assuré aura une probabilité de 0.07 de se voir

10. Projets ONI / EPARS
11. INSEE : Populations légales 2018

attribuer le code postal de Aix-en-Provence, à savoir 13080.

2.3 Présentation des données en Open Data

Pour être en mesure d'expliquer le comportement des assurés par un effet spatial, il convient d'avoir à disposition des données géographiques médicales ou sociales. L'INSEE propose un grand nombre de données à disposition à la maille communale.

2.3.1 Données médicales

Le site de l'INSEE (Institut national de la statistique et des études économiques) propose plusieurs bases de données liées aux services médicaux. Notamment une base permanente des équipements (BPE) [12] et fonctions médicales et paramédicales. Cette base indique pour chaque commune française et pour chaque spécialité, le nombre de professionnels de santé exerçant leur activité en libéral, en 2019. Elle comprend :

— Les médecins généralistes ;
— Les médecins spécialistes (toutes les spécialités ne sont pas données, spécialités les plus importantes en effectif de libéraux y sont indiquées) ;
— Les professions paramédicales (auxiliaires, sage-femmes, infirmiers, etc.) ;
— Les audio-prothésistes (libéraux ou non).

Ces données peuvent être des indicateurs du risque puisque la présence médicale pourrait influer sur la consommation d'actes médicaux couverts par la complémentaire santé.

Présentation des variables

La base BPE de l'INSEE contient une vingtaine de variables pour 34 968 lignes, soit une ligne par commune. Le détail des variables est donné dans le tableau 2.8 ci-dessous.

Variable	Description
CODGEO	Code INSEE
Médecin.généraliste	Nombre de médecins généralistes
Spécialiste.en.cardiologie	Nombre de cardiologues
Spécialiste.en.gynécologie	Nombre de gynécologues
Spécialiste.en.gastro.entérologie	Nombre de gastro-entérologues
Spécialiste.en.radiodiagnostic	Nombre de radiologues
Chirurgien.dentiste	Nombre de dentistes
Sage.femme	Nombre de sage-femmes
Infirmier	Nombre d'infirmiers
Masseur.kinésithérapeute	Nombre de kinésithérapeutes
Audio.prothésiste	Nombre d'audioprothésistes

TABLE 2.8 – Variables de la base BPE

Dans cette partie, toutes les professions ne seront pas étudiées. En revanche, il est intéressant de visualiser la présence médicale à la maille communale pour les professions les plus représentées et qui sont concernées par les postes de notre portefeuille. De plus, les études effectuées dans ce mémoire se restreignent géographiquement à la France métropolitaine. Il n'est donc pas nécessaire de conserver les communes des départements outre-mer et territoires outre-mer.

12. Source : INSEE, Base permanente des équipements 2019

Les professions de santé les plus représentées

Au total, la base BPE répertorie en France métropolitaine 321 692 professionnels de santé. Mais tous ne seront pas retenus par manque d'informations.

Dans l'ordre, les cinq professions les plus représentées de cette base [13] sont :

— Infirmier, avec 95 238 praticiens ;
— Kinésithérapeutes, avec 72 350 praticiens ;
— Médecin généraliste, avec 59 781 praticiens ;
— Chirurgien-dentiste, avec 37 518 praticiens ;
— Radiologue, avec 9 564 praticiens.

Finalement, ces 5 professions composent 85% des professionnels de santé totaux de la base. Ce sont les professions qui seront étudiées par la suite.

Zones de déserts médicaux

Si la base à disposition ne répertorie pas le nombre exacte de professionnels de santé en France (puisque, dans les cas des généralistes et spécialistes, elle n'inclut pas les médecins salariés exerçant en établissement sanitaires), celle-ci peut donner une vision des zones à forte présence médicale, et par extension permettre d'expliquer le risque par zone.

Ainsi, si l'on trace sur une carte thermique la présence médicale pour chaque commune, on peut voir les zones les plus à risque. La figure 2.17 présente le nombre de professionnels de santés exerçant une des cinq professions citées ci-dessus, sur une carte de la France.

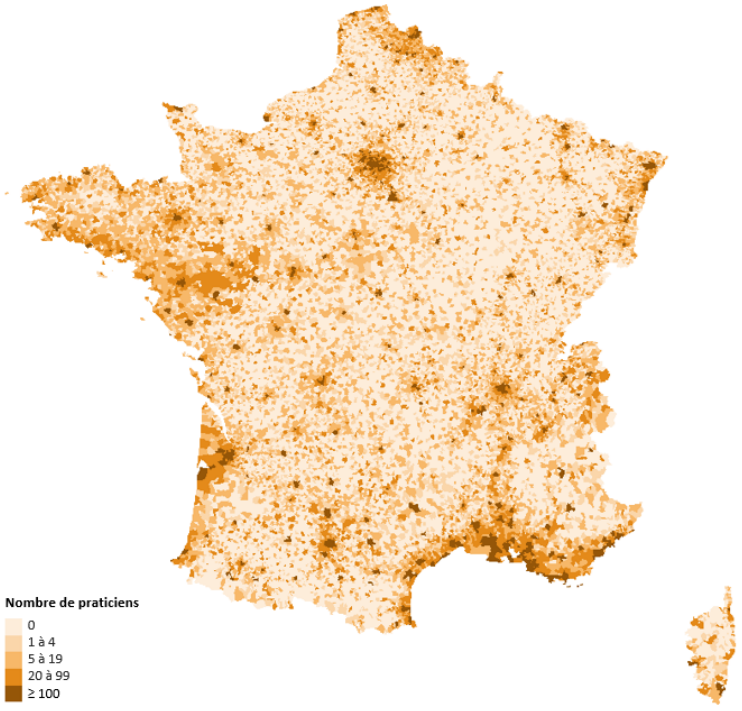

FIGURE 2.17 – Cartographie de la concentration médicale pour les professions Infirmier, Médedecin généraliste, Kynésithérapeute, Chirurgien-dentiste et Radiologue

Les grandes agglomérations sont particulièrement concentrées, et leurs villes voisines le sont

13. Dont la profession concerne les garanties de notre portefeuille

aussi. Typiquement, les villes ayant la plus forte démographie médicale sont : Paris, Marseille, Lyon, Nice et Toulouse, qui sont des villes très denses en habitants.

Il apparaît aussi un grand nombre de déserts médicaux, c'est-à-dire de villes n'ayant aucun professionnel de santé. Dans notre base, on recense 21 487 communes n'ayant aucun de ces cinq pratiquants. Ce phénomène est d'autant plus considérable qu'un individu résidant dans ces zones ayant des besoins médicaux est contraint d'effectuer des trajets de moyennes-longues distances pour se faire soigner.

Les concentrations médicales étant proportionnellement très proches pour les cinq professions, elles ne seront pas étudiées individuellement.

2.3.2 Données démographiques et socio-économiques

La base des dispositif sur les revenus localisés sociaux et fiscaux de l'INSEE [14] propose les indicateurs de revenus financiers déclarés par commune.

Les variables de cette base sont décrites dans le tableau 2.9 ci-dessous.

Variable	Description
CODGEO	Code INSEE
SUPERF	Superficie (km^2)
NAIS1318	Naissances entre 2013 et 2018
DECE1318	Décès entre 2013 et 2018
P18_MEN	Ménages en 2018
P18_LOG	Logements en 2018
NBMENFISC18	Nombre de ménages fiscaux en 2018
MED18	Médiane du niveau vie en 2018
P18_EMPLT	Emplois au lieu de travail en 2018
P18_CHOM1564	Chômeurs 15-64 ans en 2018
P18_ACT1564	Actifs 15-64 ans en 2018

TABLE 2.9 – Variables de la base des revenus fiscaux

L'INSEE propose aussi une base des principaux indicateurs sur les revenus et la pauvreté au niveau national [15]. Cette base est à disposition depuis 2020, mais les informations afférentes à cette base datent de 2017. Les variables retenues issues de cette base sont les suivantes (voir tableau 2.10) :

Variable	Description
NBMENFISC17	Nombre de ménages fiscaux
NBPERSMENFISC17	Nombre de personnes dans les ménages fiscaux
MED17	Médiane du niveau de vie (€)
PIMP17	Part des ménages fiscaux imposés (%)
TP6017	Taux de pauvreté-Ensemble (%)

TABLE 2.10 – Variables de la base des indicateurs de revenus et de pauvreté

En plus des données médicales, ces informations socio-économiques peuvent tendre à expliquer

14. INSEE : Dispositif sur les revenus localisés sociaux et fiscaux
15. INSEE : Principaux indicateurs sur les revenus et la pauvreté aux niveaux national et local

le comportement des assurés au sein d'une ville. Par la suite, ces données à la maille code INSEE seront agrégées à la maille code postal en vue d'être en adéquation avec notre domaine d'étude.

Conclusion

Les données à disposition une fois retraitées et analysées peuvent être considérées comme suffisamment robustes pour faire l'objet d'études poussées et de construction de modèles. L'ajout de variables externes permet notamment d'effectuer un parallèle entre le comportement des assurés et la situation médicale en France, donc par extension d'expliquer la consommation d'actes médicaux dans une zone, et optimiser les modèles de tarification en conséquence.

Troisième partie

Aspects théoriques

Chapitre 3

Modèles de tarification et GLM

L'assurance est caractérisée par le principe d'**inversement du cycle de production**. C'est-à-dire que, contrairement à un commerçant classique qui connaît le coût de production de ses biens et fixe les prix en conséquence, un assureur fixe un prix pour couvrir un aléa dont il ignore le montant et la date de réalisation. Le prix fixé s'appelle la prime d'assurance et est déterminé par les actuaires via des modèles de tarification.

3.1 La modélisation de la prime pure

La définition de la prime payée par l'assuré à l'assureur peut se faire à travers la modélisation de sa consommation en frais de santé.

3.1.1 Le modèle « Coût × Fréquence »

Pour estimer le coût moyen des prestations à verser pour un acte donné, les actuaires modélisent généralement indépendemment la fréquence de consommation, c'est-à-dire le nombre moyen de prestations versées par l'assureur à un assuré, et le coût moyen de ces versements. Le produit des deux donne alors la prime pure. C'est le modèle « Coût × Fréquence ».

L'utilisation de cette méthode nécessite deux hypothèses fortes :

1. Les coûts des prestations X_i sont indépendants et identiquement distribués ;
2. Les variables aléatoires X_i, représentant le montant des prestations, sont indépendantes du nombre de prestations N.

La seconde hypothèse n'est pas toujours vérifiée. En effet, il est avéré qu'en santé, les « petits » sinistres sont bien plus fréquents que les « gros », ce qui pose une certaine dépendance entre le nombre et les montants.

Ainsi, les hypothèses citées ci-dessus combinées aux équations 1.1 et 1.2 permettent d'estimer la charge totale de prestations S à travers l'équation suivante (voir équation 3.1) :

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[X_1] \times \mathbb{E}[N] \tag{3.1}$$

On lit mathématiquement que la charge moyenne pour un acte est égale au coût moyen de cet acte fois le nombre moyen de consommation de cet acte.

En pratique, cette estimation se base sur l'historique des prestations présentes dans le portefeuille.

Cette charge de prestations étant particulièrement volatile, il convient de calculer sa variance. La formule de décomposition de la variance s'écrit comme suit (voir équation 3.2) :

$$Var(S) = Var(\mathbb{E}[S|N]) + \mathbb{E}[Var(S|N)] \tag{3.2}$$

Le terme $\mathbb{E}[S|N]$, c'est-à-dire l'espérance de la charge totale sachant le nombre de prestations survenues, se décompose aisément grâce aux hypothèses d'indépendance entre les coûts et le nombre de prestations et d'identique distribution de celles-ci (voir équation 3.3) :

$$\mathbb{E}[S|N] = \mathbb{E}\left[\sum_{i=1}^{N} X_i|N\right] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{E}[X_i] = N \times \mathbb{E}[X_1] \tag{3.3}$$

Toujours en conservant les hypothèses, le terme $Var(S|N)$, c'est-à-dire la variance de la charge totale sachant le nombre de prestations survenues, se décompose de la façon suivante (voir équation 3.4) :

$$Var(S|N) \;=\; Var\left(\sum_{i=1}^{N} X_i|N\right) \;=\; Var\left(\sum_{i=1}^{N} X_i\right) \;=\; \sum_{i=1}^{N} Var(X_i) \;=\; N \times Var(X_1) \tag{3.4}$$

En conséquence, les équations 3.2, 3.3 et 3.4 permettent de calculer la variance de la charge totale (voir équation 3.5) :

$$\begin{aligned} Var(S) &= Var(N \times \mathbb{E}[X_1]) + \mathbb{E}[N \times Var(X_1)] \\ &= Var(N) \times (\mathbb{E}[X_1])^2 + \mathbb{E}[N] \times Var(X_1) \end{aligned} \tag{3.5}$$

Le premier terme de l'égalité 3.5 représente la part de volatilité liée au nombre de prestations versées, le second met en avant la part de volatilité du montant des prestations.

3.1.2 Choix des lois paramétriques

La modélisation de la prime pure nécessite d'être capable d'approcher les éléments de celle-ci par des lois de distributions connues. Dans le cadre de la méthode » Coût × Fréquence », ces éléments sont la fréquence de survenance des sinistres et le coût moyen de ceux-ci. Dans le cadre de la méthode « Occurrence × Charge totale », ces éléments sont la probabilité d'occurrence d'un sinistre et la charge de sinistralité de celui-ci.

Lois discrètes

Modéliser la fréquence revient à modéliser le nombre de sinistres susceptibles de survenir dans la période de temps. Il est donc nécessaire de l'approcher par une loi dite discrète.

La fréquence d'un sinistre représente le nombre de sinistres par personne, divisé par le prorata temporis de présence dans le contrat de celle-ci sur une période de temps donnée, appelé l'exposition. Dans un portefeuille à T assurés, la fréquence de sinistre du $i^{\text{ème}}$ assuré sera :

$$\textit{Fréquence}_i = \frac{\textit{Nombre de sinistres de l'assuré } i}{\textit{Exposition de l'assuré } i}$$

A titre d'exemple : on considère un contrat d'assurance dont la date de renouvellement est le 31 Décembre, auquel a souscrit un assuré le 31 Juin, soit en milieu d'année. Son exposition est alors d'environ 0.5. Si ce même assuré a un sinistre au cours de cette période, sa fréquence de sinistres sera alors de : $\frac{1}{0.5} = 2$.

Et la fréquence de sinistre du portefeuille sera :

$$\textit{Fréquence} = \frac{\sum_{i=1}^{T} \textit{Nombre de sinistres de l'assuré } i}{\sum_{i=1}^{T} \textit{Exposition de l'assuré } i}$$

Comme dit plus tôt, la fréquence d'un sinistre se modélise par une loi de distribution discrète. En pratique, on utilise soit une loi Poisson, soit une loi Binomiale négative, mais il en existe bien d'autres.

— **La loi Poisson** est une loi décrivant la probabilité qu'un évènement se produise un certain nombre de fois sur une période donnée. Soit X un évènement et $N(X)$ le nombre de survenance de cet évènement suivant une loi Poisson de paramètre $\lambda > 0$. La probabilité que l'évènement X survienne k fois s'écrit alors :

$$\mathbb{P}\left[N(X) = k\right] = \frac{\lambda^k}{k!} \times e^{-\lambda}$$

La loi Poisson présente une caractéristique spécifique : sa variance est égale à son espérance, elle-même égale au paramètre λ (voir équation 3.6) :

$$\mathbb{E}[N(X)] = Var(N(X)) = \lambda \tag{3.6}$$

La démonstration de l'équation 3.6 se trouve en Annexe B.3.

— La **loi Binomiale négative** est une loi visant à probabiliser le nombre d'épreuves indépendantes nécessaires pour obtenir un certain nombre de succès. Elle a pour paramètres une probabilité p que l'évènement se produise, et un entier n représentant le nombre de succès souhaité. Soit X un évènement de probabilité p et $N(X)$ le nombre d'essais à réaliser pour que X se réalise n fois. La probabilité d'avoir besoin de k ($\geq n$) essais s'écrit donc :

$$\mathbb{P}[N(X) = k] = C_{k-1}^{n-1} \times p^n \times (1-p)^{k-n}$$

Son espérance s'écrit $\mathbb{E}[N(X)] = \frac{n}{p}$ et sa variance $Var(N(X)) = n \times \frac{1-p}{p^2}$.

— La **loi Zero-inflated Poisson (ZIP)** est utilisée lorsqu'une distribution contient un nombre important d'observations nulles. Au même titre que la loi Poisson, la loi ZIP permet de décrire la probabilité qu'un évènement se produise un certain nombre de fois, mais cette fois-ci en séparant la distribution en deux processus : un processus de comptage de 0 avec la loi binomiale, et un processus de comptage du nombre de réalisations de l'évènement avec la loi Poisson. En reprenant les termes de la loi Poisson, sa probabilité s'écrit comme suit :

$$\mathbb{P}_{ZI}\left[N(X) = 0\right] = \pi + (1-\pi) \times e^{-\lambda}$$

$$\mathbb{P}_{ZI}\left[N(X) = k\right] = (1-\pi) \times \frac{\lambda^k}{k!} \times e^{-\lambda}$$

Avec $k \in \mathbb{N}^*$, λ le paramètre le la loi Poisson, et π la probabilité d'avoir structurellement un excès de 0.

— La **loi Zero-inflated Binomiale négative (ZIBN)**, au même titre que la loi ZIP, est utilisée en cas de nombre important d'observations nulles. Elle probabilise d'un côté le nombre d'observations nulles, et d'un autre côté le nombre de fois qu'un évènement se produit. En reprenant les termes de la loi Binomiale négative, sa probabilité s'écrit donc :

$$\mathbb{P}_{ZI}\left[N(X)=0\right] = \pi + (1-\pi) \times \mathbb{P}[N(X)=0]$$

$$\mathbb{P}_{ZI}\left[N(X)=k\right] = (1-\pi) \times \mathbb{P}[N(X)=k]$$

Où $\mathbb{P}$ est la fonction de probabilité de la loi Binomiale négative, et π la probabilité d'avoir structurellement un excès de 0.

Lois modélisant le coût moyen

Le coût moyen d'un sinistre représente la montant moyen de l'ensemble des sinistres survenus. Il s'exprime donc comme suit :

$$\textit{Coût moyen}_i = \frac{\textit{Montant total des sinitres de l'assuré } i}{\textit{Nombre de sinistres de l'assuré } i}$$

Par extension, le coût moyen du portefeuille sera :

$$\textit{Coût moyen} = \frac{\sum_{i=1}^{T} \textit{Montant total des sinitres de l'assuré } i}{\sum_{i=1}^{T} \textit{Nombre de sinistres de l'assuré } i}$$

A la différence de la fréquence qui se modélise par une loi de distribution discrète, le coût est un nombre réel, il se modélise donc par une loi de distribution continue, généralement, la loi Gamma, ou la loi Log-normale.

— La **loi Gamma** est une loi de probabilité réelle positive. Elle dépend de paramètres k et θ. Si une variable aléatoire X suit une loi Gamma ($X \sim \Gamma(k,\theta)$, sa fonction de répartition s'exprime comme suit :

$$f(x,k,\theta) = \frac{x^{k-1} \times \mathrm{e}^{-\frac{x}{\theta}}}{\Gamma(k) \times \theta^k}$$

Où $x > 0$ et Γ représente la fonction Gamma d'Euler.

— La **loi Log-normale** est également une loi de probabilité réelle positive. On dit que X suit une loi Log-normale lorsque son logarithme suit une loi Normale. En d'autres termes $X \sim Log-\mathcal{N}(\mu,\sigma^2)$ si $Y = Log(X) \sim \mathcal{N}(\mu,\sigma^2)$.
Sa densité de probabilité s'écrit comme suit :

$$f_Y(x,\mu,\sigma) = \frac{1}{x \times \sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2 \times \sigma^2}\right)$$

3.2 Les Modèles Linéaires Généralisés (GLM)

Le modèle linéaire généralisé, ou Generalized Linear Model (GLM) [1], est une approche statistique multivariée souvent utilisée en assurance pour modéliser le coût des sinistres pour de

1. Formulé par John Nelder et Robert Wedderburn en 1972

nombreux types de risques, notamment le risque santé. Cette approche a pour objet d'exprimer la relation entre une variable réponse observée Y (par exemple la fréquence ou le coût d'un sinistre) et un certain nombre de covariables $X_1, X_2, ..., X_p$, appelées variables explicatives.

Ce modèle est souvent utilisé pour la tarification car peu coûteux en temps de calcul et facile à interpréter.

Comme son nom l'indique, le GLM est une généralisation du modèle linéaire classique (ou modèle linéaire gaussien), qui repose sur des hypothèses fortes et pas toujours vérifiées.

3.2.1 Principe des modèles linéaires classiques

Dans un modèle linéaire, la variable réponse Y est conceptualisée comme la somme de sa moyenne, μ, et d'un résidu ε.

$$Y = \mu + \varepsilon$$

Le modèle linéaire classique requiert les hypothèses suivantes :

1. μ est une combinaison linéaire des variables explicatives :

$$\mu = \beta_0 + \sum_{i=1}^{p} \beta_i \times X_i$$

   Avec β_0 le coefficient individuel, appelé « intercept », $\beta_1, ..., \beta_p$ les coefficients de régression estimés à partir des données, et $X_1, X_2, ...X_p$ les variables explicatives.

2. Le résidu ε est une variable aléatoire suivant une loi normale d'espérance nulle et de variance σ_ε^2 :

$$\varepsilon \sim \mathcal{N}(0,\, \sigma_\varepsilon^2)$$

La variable réponse peut donc s'écrire comme suit (voir équation 3.7) :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i \times X_i + \varepsilon \tag{3.7}$$

Le modèle suppose donc que la variable réponse Y suit une loi normale d'espérance $\beta_0 + \sum_{i=1}^{p} \beta_i \times X_i$ et de variance constante σ_ε^2.

La notation matricielle se trouve en Annexe B.5

La normalité des résidus entraîne celle du vecteur de variables réponses. Ainsi, sous l'hypothèse d'homoscédasticité [2] des résidus, nous avons $Y \sim \mathcal{N}(X \cdot \beta,\, \sigma_\varepsilon^2 I_n)$, ce qui fait une des limites de ce modèle. En effet, la normalité de la variable réponse n'est en pratique pas toujours vérifiée. De plus, il est difficile d'affirmer la constance systématique de la variance de Y.

L'utilisation des modèles linéaires généralisés (GLM) permet de faire face à ces contraintes.

3.2.2 Principe du GLM

Les GLM comportent une large gamme de modèles qui incluent les modèles linéaires classiques comme un cas particulier. Les hypothèses de normalité de la variable réponse et de constance de la variance ne sont plus imposées.

Le GLM fait toutefois appel aux hypothèses suivantes :

2. Tous les résidus doivent avoir la même variance.

1. Le vecteur de variables réponses $Y = (Y_1, Y_2, ..., Y_n)^T$ est indépendant et identiquement distribué et sa distribution doit appartenir à la famille **exponentielle**, développée en Annexe B.6 ;
2. La moyenne du vecteur réponse Y est fonction inverse d'une combinaison linéaire des variables explicatives $X_1, X_2, ..., X_p$:

$$\mathbb{E}[Y] = \mu = g^{-1}(X \cdot \beta) \tag{3.8}$$

ou réciproquement :

$$g(\mu) = X \cdot \beta$$

avec X et β définis comme en partie 3.2.1
g étant une fonction monotone et différentiable, appelée « fonction lien ».

3.2.3 Validation du modèle

Un modèle est validé par sa qualité à expliquer les variables réponses par les variables explicatives. Cette qualité se mesure à travers le calcul de différentes mesures statistiques.

La déviance

Le calcul de la déviance a pour but de comparer le modèle calibré à un modèle dit « saturé » où les estimations seraient identiques aux observations.
Ainsi, la déviance peut s'écrire comme suit :

$$D = -2 \times \varphi \times log\left(\frac{L_{\text{saturé}}(Y)}{L(Y, \hat{\beta}, X)}\right) \tag{3.9}$$

Avec L la fonction de vraisemblance définie en partie B.7.
Il est alors clair que plus la déviance est faible, meilleur est le modèle et les variables prédites seront plus proches des variables observées.

Les critères AIC et BIC

Le critère d'information d'Akaike (AIC) est un estimateur de l'erreur de prédiction et donc de la qualité relative des modèles linéaires. L'idée principale de ce critère est d'estimer la quantité d'information perdue par un modèle donné. Par conséquent, plus l'AIC est faible, plus l'information perdue est faible et donc meilleur est le modèle. Il est basé sur la fonction de vraisemblance maximisée $\hat{L}$ et le nombre de paramètres k estimés dans le modèle. L'AIC s'écrit alors :

$$AIC = 2 \times k - 2 \times log(\hat{L}) \tag{3.10}$$

Le critère d'information bayésien (BIC) suit le même principe, en prenant additionnellement en compte le nombre de lignes n. Il s'écrit comme suit :

$$BIC = k \times log(n) - 2 \times log(\hat{L}) \tag{3.11}$$

De façon générale, l'un des critères n'est pas meilleur que l'autre. Toutefois, l'AIC peut choisir un mauvais modèle quelque soit n, alors que le BIC, se basant sur plus de paramètres, a moins de chance de choisir un mauvais modèle si n est assez grand. En revanche, pour des

bases relativement faibles, le BIC ne choisit pas toujours le meilleur modèle du fait de sa forte pénalisation.

En pratique, on choisi le modèle pour lequel ces deux critères sont les plus faibles.

Un enjeu clé lors du calibrage d'un GLM est de trouver le bon sous-ensemble de variables explicatives, de sorte que le modèle soit le moins complexe possible tout en optimisant les critères de validation. Pour cela, il existe plusieurs méthodes de sélection de variables.

3.2.4 Sélection des variables

Le calibrage d'un modèle linéaire comportant des variables liées entre elles ou peu explicatives peut entraîner des biais qui impacteraient négativement la qualité du modèle. Pour pallier ce problème, il existe des méthodes dites « pas-à-pas », ou « septwise » consistant à sélectionner les variables une à une. Les plus connues sont la méthode « forward » et la méthode « backward ».

La méthode Forward

La méthode forward est une méthode itérative qui consiste à tester l'ajout d'une nouvelle variable dans le modèle et visualiser l'impact sur les critères de performance. L'algorithme est le suivant :

1. Construire un modèle linéaire sans variable explicative ;
2. Ajouter une variable explicative au modèle ;
3. Observer le critère choisi (déviance, AIC ou BIC) : s'il est plus élevé qu'avant l'ajout, retirer la variable ajoutée, s'il est plus faible ou inchangé, la conserver ;
4. Répéter l'algorithme à partir de l'étape 2, jusqu'à ce que l'ajout d'une variable n'améliore plus le critère.

La méthode Backward

Au même titre que la méthode forward, la méthode backward est une méthode itérative qui consiste, à chaque itération, à retirer une variable et observer l'impact de ce retrait sur les critères de performance. L'algorithme est le suivant :

1. Construire un modèle linéaire avec l'ensemble des variables explicatives ;
2. Retirer une variable explicative du modèle ;
3. Observer le critère choisi : s'il est plus élevé qu'avant le retrait, remettre la variable retirée, s'il est plus faible ou inchangé, conserver ce retrait ;
4. Répéter l'algorithme à partir de l'étape 2 jusqu'à ce que le retrait d'une variable n'améliore plus le critère.

Fondamentalement, l'une des méthodes n'est pas meilleure que l'autre. Il convient d'appliquer les deux et conserver la plus optimale.

Conclusion

L'enjeu de la tarification repose sur l'estimation du risque. Cette estimation peut se faire via des modèles linéaires généralisés calibrés de façon adéquate. Le GLM est le modèle la plus utilisé en assurance pour la tarification, notamment en raison de sa simplicité. Cependant, ce dernier présente des limites. En effet, les hypothèse paramétriques ne sont, en pratique, pas toujours

vérifiées. Et si elles ne le sont pas, le modèle peut être mal adapté aux données. Des méthodes dîtes non-paramétriques permettent une modélisation des paramètres par apprentissage. Si ces méthodes sont moins utilisées dans le secteur de l'assurance pour leur interprétabilité, elles peuvent être plus précises que les GLM et ne dépendent pas des hypothèses de distribution des variables.

Chapitre 4

Les méthodes non-paramétriques

Une des principales limites de la prédiction par modèle linéaire est le besoin d'indépendance entre les variables. Lorsque deux variables interagissent entre elles, elles biaisent les résultats du modèle.

Les méthodes non-paramétriques permettent de contourner le besoin d'indépendance des variables explicatives en vue de prédire une variable réponse. Ces méthodes se basent sur l'apprentissage statistique de la machine, aussi appelé « Machine Learning », c'est-à-dire qu'elles se paufinent automatiquement grâce à l'expérience et à l'utilisation des données.

4.1 Les arbres de décision

Un arbre de décision se définit comme une méthode d'apprentissage statistique. Elle est basée sur la construction d'un arbre dont les feuilles représentent les valeurs possibles de la variable réponse, et les branches des combinaisons de valeurs possibles des variables explicatives.

Contrairement aux modèles linéaires, les arbres de décision cartographient assez bien les relations non-linéaires entre les variables pour résoudre des problèmes de classification ou de régression.

Dans un arbre de décision, on distingue différents types de noeuds

— Le premier noeud, ou la racine, où s'effectue la première division en fonction de la variable la plus significative ;
— Les noeuds intermédiaires, qui divisent à nouveau l'ensemble des données en fonction des valeurs des variables ;
— Les feuilles, qui sont situées au bas du schéma et leur fonction est d'indiquer la classification ou la valeur finale.

Le schéma 4.1 illustre l'arbre de décision en ces éléments.

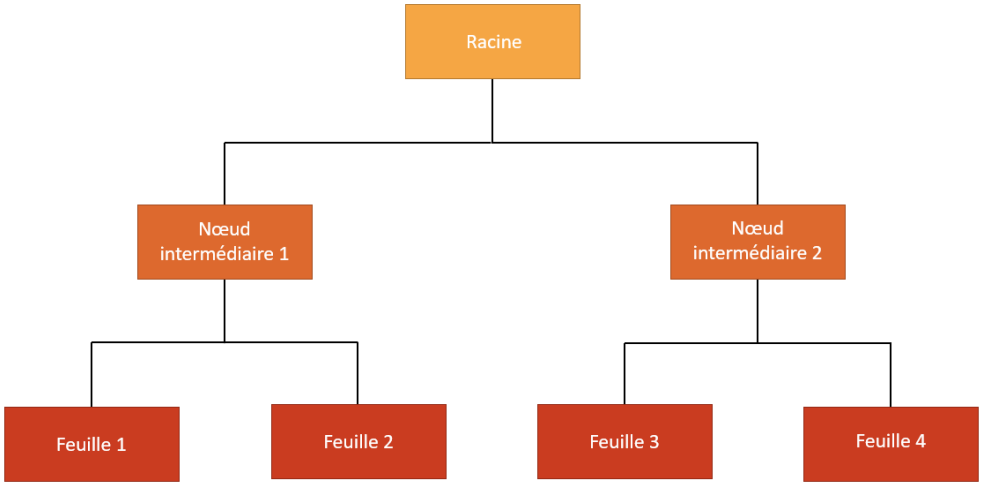

FIGURE 4.1 – Arbre de décision

Les arbres de décision ont tendance à se sur-ajuster. Bien qu'ils apprennent efficacement des données d'apprentissage, leurs résultats peuvent différer d'une simulation à une autre. Pour fiabiliser leurs résultats, on peut utiliser les forêts aléatoires.

4.2 Les forêts aléatoires

Les forêts aléatoires sont fondées sur les arbres de décision, présentés précédemment, et sont utilisées à des fins de classification ou de régression.

Une forêt aléatoire (ou Random Forest) est un ensemble d'arbres de décision combinés par *bagging*[1].

Lorsqu'on utilise le bagging, différents arbres voient différentes parties des données. Aucun arbre ne voit toutes les données d'apprentissage. Cela signifie que chaque arbre est entraîné avec différents échantillons de données pour le même problème. De cette façon, en combinant leurs résultats, certaines erreurs sont compensées par d'autres et la prédiction en est mieux généralisée.

Il y a deux étapes dans l'algorithme de la forêt aléatoire, la première est la création de la forêt (de façon aléatoire comme l'indique son nom) et l'autre consiste à faire une prédiction à partir du classificateur de la forêt aléatoire créée dans la première étape.

L'algorithme procède comme suit :

1. Sélectionner aléatoirement k échantillons dans la base de données ;
2. Construire un arbre de décision pour chaque échantillon sélectionné. Nous avons alors un résultat de prédiction pour chaque arbre créé.
3. La valeur finale s'obtient par vote majoritaire des résultats de chaque arbre pour un problème de classification. Pour un problème de régression, la valeur finale sera la moyenne des résultats de chaque arbre.

Le schéma 4.2 représente le principe d'une prédiction par forêt aléatoire lorsqu'elle est utilisée à des fins régressives.

1. Le bagging est un type spécifique de processus d'apprentissage supervisé qui utilise l'apprentissage d'ensemble pour faire évoluer les modèles d'apprentissage supervisé.

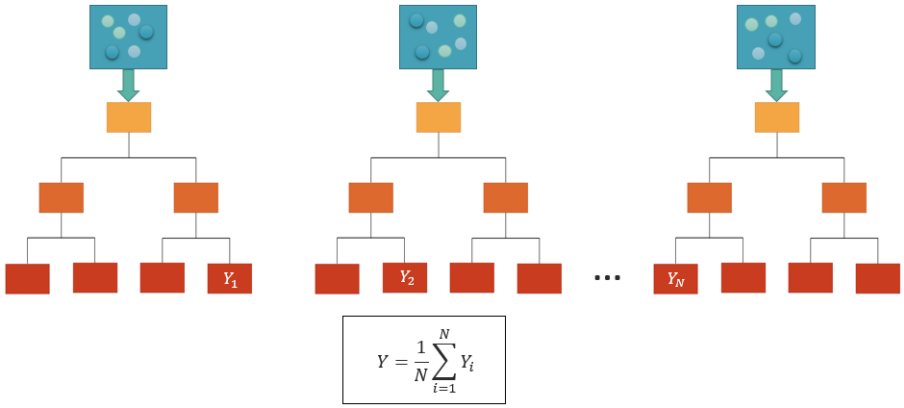

FIGURE 4.2 – Exemple de prédiction par forêt aléatoire

L'interprétation mathématique du processus de prédiction des forêts aléatoires se trouve en Annexe E.1.

La forêt aléatoire possède l'avantage d'être facile à utiliser, notamment car les paramètres par défaut donnent souvent des résultats satisfaisant, bien qu'il faille les choisir en amont pour optimiser les modèles. Les paramètres variants sont généralement le nombre d'arbres composant la forêt *ntree*, et le nombre de variables candidates à faire interagir.

De plus, un des grands problèmes du Machine Learning est le surapprentissage, c'est-à-dire le risque d'adapter le modèle trop distinctement aux données apprises, et le rendre par la suite inefficace. Le surapprentissage ne concerne pas les forêts aléatoires, puisqu'elles se basent sur une méthode de bagging.

La principale limite des forêts aléatoires est qu'un grand nombre d'arbres peut rendre l'algorithme lent et inefficace pour les prédictions en temps réel. En effet, une prédiction plus précise nécessite plus d'arbres, ce qui se traduit par un modèle plus lent.

4.3 Le Gradient Boosting

A l'instar des forêts aléatoires, le Gradient Boosting est un algorithme de Machine Learning basé sur les arbres de décisions. En revanche, à la différence des forêts aléatoires, qui utilisent le bagging, le Gradient Goosting, comme son nom l'indique, utilise le *boosting*[2]. De fait, son algorithme vise à minimiser le biais, là où la forêt aléatoire cherche à minimiser la variance.

Un modèle de Gradient Boosting consiste en un ensemble d'arbres de décision individuels, entraînés de manière itérative, de sorte que chaque nouvel arbre tente d'améliorer les erreurs des arbres précédents par la méthode de décente du gradient.

L'algorithme de décente du gradient procède comme suit :
Soit f une fonction convexe que l'on souhaite minimiser et soit Δf son gradient.

1. Initialisation : Choisir x_0 appartenant au domaine de définition de f ;
2. Calculer $f(x_i)$ pour chaque i ;
3. Mettre à jour x_i : $x_{i+1} = x_i - \eta \Delta f(x_t)$

2. Le boosting est un algorithme destiné à réduire le biais par apprentissage.

4. Répéter les étapes 2 et 3 jusqu'à ce que $|f(x_i) - f(x_{i-1})| < \varepsilon$ où $\varepsilon > 0$ est un réel choisi en amont.

Avec η un paramètre constant.

L'algorithme du Gradient Boosting est une généralisation de la méthode de décente du gradient. Son détail est donné en annexe E.2.

Le Gradient Boosting a pour avantage notamment de ne pas être fortement influencé par les valeurs aberrantes ou extrêmes. De plus, il est très utile dans l'exploitation des données et parvient à identifier de manière efficace les variables les plus importantes.

Les limites du Gradient Boosting reposent d'une part sur son interprétabilité : lorsqu'on combine plusieurs arbres, l'interprétabilité des modèles à arbre unique est perdue. De plus, si le nombre d'arbres est trop important, le gradient boosting peut être sujet au surapprentissage.

Chapitre 5

Construction d'un zonier

Les variables géographiques sont particulièrement utilisées dans la tarification IARD [1], ainsi qu'en assurance santé en raison, entre autre, d'une offre médicale relativement disparate en France. Certaines zones peuvent représenter un risque sensiblement proche et être regroupées dans une même classe. L'ensemble de ces classes est appelé *zonier*. Les zoniers peuvent être à plusieurs mailles selon le besoin. De la plus large à la plus fine :

— La maille régionale, peu utilisée car souvent trop large ;
— la maille départementale ;
— la maille code postal, qui représente le code du bureau de poste distributeur de la commune ;
— la maille code INSEE qui représente le code communal ;
— la maille code IRIS.

Le choix de la maille est une étape importante dans la construction d'un zonier car celle-ci doit assurer l'homogénéité des risques au sein d'une unité, et d'un nombre d'observations suffisant pour éviter une biais conséquent.

Le nombre de classes dans un zonier est généralement arbitraire et peut aller jusqu'à plusieurs dizaines, selon la précision souhaitée.

5.1 Le besoin d'un zonier dans la tarification

La tarification santé intégrant une variable spatiale est généralement plus précise. Cette variable se doit d'être à une maille assez fine pour être représentative du risque de manière à éviter l'antisélection. En effet, une maille départementale pourrait être trop large pour évaluer le niveau de risque d'un individu. En théorie, le coefficient linéaire associé à une zone sera d'autant plus élevé que la zone contient de prestations dans l'historique de données. Or, pénaliser linéairement tout un département revient à pénaliser chaque assuré résident dans celui-ci, y compris ceux résident dans des zones de déserts médicaux.

A titre d'exemple, une personne résident à Rousset (Bouches-du-Rhone), devant faire 30 minutes de route pour trouver le gynécologue le plus proche de sa ville de résidence, aura vraisemblablement tendance à modérer sa fréquence de visites gynécologiques, par rapport à une personne résident à Marseille, qui contient plus de 200 gynécologues [2]. Il serait donc difficile de

1. Incendies, accidents et risques divers
2. Annuaire santé d'Ameli.

décrire ces deux villes comme identiquement risquées.

Pour cela, les actuaires peuvent choisir de tarifer à la maille ville. Cependant, en se basant sur une maille aussi fine, cela créé une variable à plusieurs milliers de modalités différentes, ce qui pourrait complexifier le modèle linéaire et créer un biais non-négligeable dû à la disparité des données. Pour pallier cela, la construction d'un zonier consiste à classer chaque ville dans une catégorie selon le risque qu'elle représente, de manière à avoir une variable spatiale avec un nombre de modalités limité et ainsi obtenir un modèle optimal.

5.2 Étapes de construction d'un zonier

Dans le cadre de ce mémoire, la construction d'un zonier ne se base pas sur l'étude de la sinistralité en elle même, mais sur l'écart entre la sinistralité observée et la sinistralité prédite, appelé résidu. La modélisation de cette sinistralité inclut les variables spatiales et l'impact de ces variables se reporte sur les résidus : si un résidu est élevé, cela signifie que l'on a surestimé la sinistralité et que par conséquent la zone associée est peu risquée. Si à l'inverse un résidu est faible ou négatif, cela signifie que la sinistralité est sous-estimée et la zone est ainsi plus à risque.

Les variables explicatives peuvent être placées dans deux catégories :
- Les variables spatiales, qui concernent ici le code INSEE, et éventuellement les variables externes liées au code INSEE, comme les données sociales, démographiques, ou économiques ;
- les variables non-spatiales, qui concernent toutes les autres variables tarifaires liées aux caractéristiques de l'assuré.

Ainsi, les étapes de construction d'un zonier sont les suivantes :

1. Modéliser le risque à partir des variables explicatives non-spatiales uniquement, préalablement sélectionnées. Ce modèle est appelé **modèle contraint**, dans le sens où les coefficients linéaires issus de ce modèle seront gelés pour le modèle suivant. La sinistralité peut s'exprimer comme suit :

$$\textit{Sinistralité observée} = \textit{Sinistralité prédite} + \varepsilon^{contraint} \tag{5.1}$$

Ou linéairement, si l'on a n variables non-spatiales $X_1, X_2, ..., X_n$, et qu'on écrit la sinistralité Y :

$$Y = X_1 \times \beta_1 + ... + X_n \times \beta_n + \varepsilon^{contraint} \tag{5.2}$$

La sinistralité peut représenter la fréquence, le coût, ou la prime pure. L'élément *Sinistralité prédite* représente l'effet non-spatial. On s'intéresse donc au $\varepsilon^{contraint}$, qui contient l'information géographique à laquelle s'ajoute une part d'aléa. La figure 5.1 schématise la première étape.

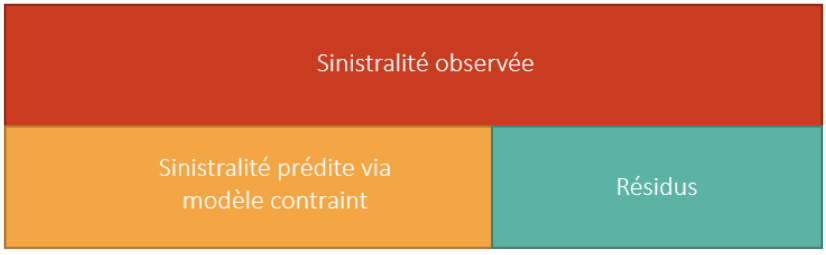

FIGURE 5.1 – Décomposition de la sinistralité selon la première étape

2. Agréger les résidus $\varepsilon^{contraint}$ à chaque zone en calculant la moyenne pondérée par l'exposition au sein de celle-ci. A l'issu de cette étape, chaque zone dispose d'un résidu. Ce résidu contient l'effet spatial en plus d'un sous-résidu.
   Pour une zone i, la sinistralité peut s'écrire comme suit :

$$\begin{aligned} \textit{Sinistralité observée}_i &= \textit{Sinistralité prédite}_i + \varepsilon_i^{contraint} \\ &= \textit{Effet non-spatial}_i + \textit{Effet spatial}_i + \varepsilon_i \end{aligned} \tag{5.3}$$

   La figure 5.2 illustre la décomposition de la sinistralité en fonction de ces éléments.

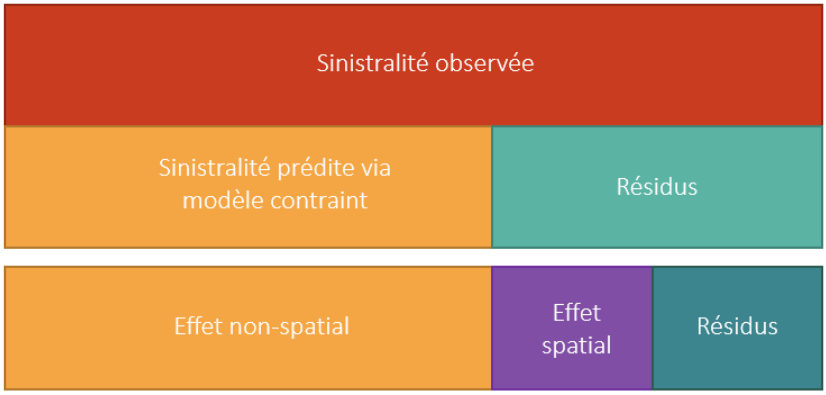

FIGURE 5.2 – Décomposition de la sinistralité selon la deuxième étape

3. Lisser les résidus $\varepsilon_i^{contraint}$. L'idée ici est d'attribuer un résidu aux zones voisines des zones observées, et dépénaliser les communes porteuses de valeurs extrêmes, de sorte que l'écart entre l'effet spatial et le résidu $\varepsilon^{contraint}$ se rapproche d'un bruit blanc.
   La sinistralité peut ainsi se décomposer comme suit :

$$\begin{aligned} \textit{Sinistralité observée} &= \textit{Sinistralité prédite} + \varepsilon^{contraint} \\ &= \textit{Effet non-spatial} + \textit{Effet spatial lissé} + \varepsilon \end{aligned} \tag{5.4}$$

   Le ε étant supposé être un bruit blanc.
   La figure 5.3 illustre les résultats attendus à l'issu de cette étape.

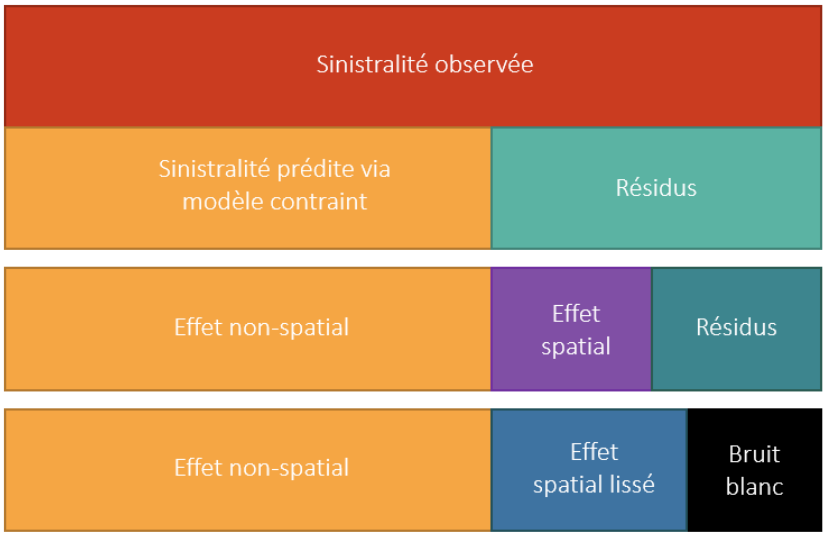

FIGURE 5.3 – Décomposition de la sinistralité selon la quatrième étape

Les méthodes de lissage seront développées plus tard dans ce chapitre.

4. Modéliser l'effet spatial. Il s'agit ici de prédire le résidu lissé « *Effet spatial lissé* $+ \varepsilon$ » de l'équation 5.4 à partir de variables géographiques de façon à ne récupérer que l'effet spatial lissé. Schématiquement, on cherche donc à capter la partie bleue du schéma 5.3, qui était jusqu'alors contenu dans le résidu.
5. Classification de l'effet spatial et réduction du nombre de classes. Les méthodes de classification seront développées plus tard dans ce chapitre.

La validation du modèle passera par une comparaison des indicateurs de performance entre l'avant et l'après intégration du zonier dans le modèle. Le modèle enrichi de la variable zonier est appelé « modèle complet ».

En assurance santé, les risques étant très différents en fonction de la garantie, il convient de construire un zonier pour chaque acte.

5.3 Lissage par la théorie de la crédibilité

Lorsque l'on construit un zonier, il y a un fort besoin de données spatiales réparties le plus uniformément possible. L'absence de sinistre dans une zone ne veut pas dire que le risque y est absent, d'autant plus si les zones voisines sont sinistrées, mais simplement que cette zone manque d'observations. De plus, si une zone contient une valeur extrême, la présence de charge importante dans cette zone ne veut pas forcément dire que celle-ci est plus risquée que d'autres, d'autant plus si les zones voisines sont peu ou pas sinistrées.

Le lissage par théorie de la crédibilité vise à donner à un individu une nouvelle valeur qui sera une pondération entre sa valeur effective et la valeur de l'ensemble du portefeuille.

Ainsi, une zone i aura une certaine valeur de résidu agrégé r_i. Cette zone se verra attribuer une nouvelle valeur de résidu r_i^*, qui s'exprime comme suit (voir équation 5.5) :

$$r_i^* = Z_i \times r_i + (1 - Z_i) \times \frac{\sum_{j=1}^{n} r_j \times e_j \times f(d_{i,j})}{\sum_{j=1}^{n} e_j \times f(d_{i,j})} \tag{5.5}$$

Avec :

— Z_i le facteur de crédibilité, compris entre 0 et 1, associé à la zone i ;
— e_i l'exposition au sein de la zone i ;
— $f(d_{i,j})$ une fonction monotone de la distance entre la zone i et la zone j ;

Le facteur de crédibilité Z_i dépend de l'exposition au sein de la zone i, et d'une constante a. Il s'écrit comme suit :

$$Z_i = \frac{e_i}{e_i + a} \tag{5.6}$$

Avec $a \in \mathbb{R}$. Le choix de a repose sur l'importance (ou le crédit) que l'on souhaite donner aux zones peu exposées. Si a est grand, les zones faiblement exposées auront un faible facteur de crédibilité, et leur valeur lissée sera de fait plus proche du terme de droite dans la formule 5.5.

Le choix de la fonction de distance $f(d_{i,j})$ dépend de l'importance que l'on souhaite donner à la distance. Plus la fonction décroît rapidement, moins les zones éloignées de la zone i auront de poids dans le calcul de r_i^*.

Cette méthode a pour avantage d'être une formule fermée, donc relativement simple à mettre en place. De plus, elle tient compte à la fois de la distance et de l'exposition des zones voisines.

L'inconvénient de cette méthode repose sur le fait que sa qualité dépend du choix, fait a priori, du coefficient a, et de la fonction de distance $f(d_{i,j})$. De plus, cette méthode n'est pas adaptée aux distribution à queue épaisse, c'est-à-dire ayant un grand nombre de valeurs extrêmes.

5.4 Méthodes de classification

Lorsqu'on construit un modèle GLM, il convient d'une part de transformer les variables quantitatives en variables qualitatives, et d'autre part de réduire le nombre de modalités des variables de sorte que le modèle soit le moins complexe possible. Pour cela, il existe des méthodes de classification consistant à regrouper les modalités ayant sensiblement le même impact entre elles. A titre d'exemple, la différence de risque entre un adulte de 35 ans et un adulte de 36 ans est sensiblement faible, il n'est donc pas aberrant de ne pas les distinguer. La construction d'un zonier vise à regrouper les zones présentant des risques similaires dans une même classe.

Il existe plusieurs méthodes de classification par apprentissage, comme la méthode des k-means, qui consiste à regrouper un échantillon dans un nombre k de groupes, fixé a priori, de telle sorte que le centroïde de chaque groupe soit le plus proche possible des éléments de son groupe respectif. Cette méthode est détaillée en Annexe C.1.1. Toutefois, cette méthode donne des résultats différents d'une simulation à une autre et, en ce sens, perd de sa fiabilité. La méthode de **classification ascendante hiérarchique** permet de pallier ce phénomène.

La classification ascendante hiérarchique (CAH) vise à partitionner une population en plusieurs classes de sorte que les individus de chaque classe soient les plus similaires possibles, et que les sous-groupes se distingue le plus possible un-à-un selon un ou plusieurs critères définis. Comme son nom l'indique, la CAH est une méthode ascendante, puisqu'elle traite chaque individu de manière itérative, et hiérarchique puisqu'elle repose sur la construction de classes, et de sous-groupes au sein de celles-ci.

L'algorithme de la CAH a pour objectif de construire une matrice de distance entre chaque individu. Plus les individus sont semblables selon un critère donné, plus la distance sera faible. Par la suite l'idée est de créer un arbre binaire (ou dendogramme) représentant le partitionnement des individus, et dont la racine correspond à la classe regroupant tous les individus. On

peut décrire l'algorithme comme suit :

On considère un ensemble de n individus à regrouper dans une matrice de distance de dimension $n \times n$.

1. Affecter à chaque individu sa propre classes, de sorte à avoir à ce niveau n classes contenant chacune un individu. Les écarts entre chaque classes sont égaux aux distances entre chaque individus.
2. Calculer les distances de chaque classe deux à deux, puis trouver les deux classes dont la distance qui les sépare est la plus faible et les fusionner, de sorte à avoir à ce niveau une classe de moins.
3. Re-calculer les distance entre chaque nouvelle classe et fusionner les deux classes dont la distance est la plus faible.
4. Répéter les étapes 2 et 3 jusqu'à n'avoir plus qu'une classe contenant n individus.

L'algorithme peut cependant s'arrêter une fois l'obtention du nombre désiré de classes.

L'étape 3 peut être réalisée de différentes manières. En effet, pour calculer la proximité entre deux classes comportant plusieurs individus, il existe plusieurs méthodes. En voici une liste non-exhaustive :

— Considérer que la distance entre une classe et une autre est égale à la plus faible distance entre tout membre de l'une, et tout membre de l'autre. Cette méthode est appelée **méthode à lien simple**.
— A l'inverse, considérer que la plus grande distance entre une classe et une autre est égale à la plus grande distance entre tout membre d'une classe et tout membre d'une autre. Cette méthode est appelée **méthode à lien complet**.
— Considérer que la distance entre une classe et une autre est égale à la distance moyenne entre tous les membres d'une classe et tous les membres de l'autre. Cette méthode est appelée **méthode à lien moyen**.
— Fusionner les classes dont le regroupement minimise la variation d'inertie intra-classe, ou, de manière équivalente, dont le regroupement maximise la variation d'inertie inter-classe. Cette méthode s'appelle **méthode de Ward** et est la plus courante dans la classification hiérarchique.

Les résultats peuvent sensiblement varier en fonction du choix de la méthode. Dans le cadre de ce mémoire, nous privilégions la méthode de Ward. Les interprétations mathématiques de cet algorithme et de ces méthodes sont données en annexe C.2.

La classification ascendante hiérarchique présente les avantages suivants :
— Relativement facile à mettre en place et à interpréter ;
— A la différence de la méthode des k-means, les résultats finaux ne dépendent pas du choix des centres initiaux.

Néanmoins, elle présente des inconvénients non-négligeables :
— Particulièrement lourde en temps de calculs avec une complexité en $O(n^2)$;
— Pas adaptée aux données volumineuses.

Il convient donc de choisir la méthode adéquate selon le jeu de données à disposition en étudiant les avantages et inconvénients de chaque méthode.

Quatrième partie

Application et présentation des résultats

Cette partie vise à mettre en pratique la théorie du zonier développée dans le chapitre 5, à savoir :

— Modéliser la sinistralité à partir des variables non-spatiales (modèle contraint) ;
— Capturer l'effet spatial contenu dans l'erreur de prédiction ;
— Classifier l'effet spatial capturé.

Dans cette partie, les modèles seront effectués sur un portefeuille regroupant le nombre et la charge de prestations de chaque assuré, pour chaque acte, y compris les assurés n'ayant pas consulté. Ce portefeuille est obtenu en agrégeant les données présentées dans le chapitre 2.

De plus, on se focalise sur la catégorie A du portefeuille (la catégorie des actifs). D'une part car il s'agit de la catégorie la plus représentée de la base effectif. D'autre part car la consommation n'est pas la même d'une catégorie à l'autre, comme on a pu le constater dans le chapitre 2. De ce fait, les assurés actifs ont globalement un comportement assez homogènes et seraient alors moins sujets à valeurs extrêmes, que ce soit en termes de coût ou de fréquence.

Ainsi, la base effectif à disposition contient 38 635 lignes, une ligne par assuré.

Étant donné le grand nombre d'actes, dans le but de ne pas dupliquer les travaux, seul l'acte *Consultations spécialistes* sera étudié dans le cadre de ce mémoire, tout en sachant que cet maille peut inclure plusieurs actes NOÉMIE [3]. Pour cet acte, les observations évoquant des majorations ont été traitées de telle sorte qu'une consultation majorée doit totaliser un acte, mais son montant vaut celui de la consultation sommé à celui de la majoration.

Le choix de cet acte s'est basé sur :

— son volume au sein du portefeuille : 56% des assurés actifs ou ayants droits d'actifs ont consulté au moins une fois un spécialiste ;
— l'hétérogénéité des coûts : à la différence de la médecine généraliste, beaucoup de médecins spécialistes ne sont pas nécessairement conventionnés, ajouté à cela les éventuelles majorations, cela laisse un large champ de coûts possibles ;
— sa répartition dans l'espace : comme vu en partie 2, cet acte est présent dans presque tous les départements, et à différents fréquences et coûts.

3. Norme Ouverte d'Échange entre la Maladie et les Intervenants Extérieurs

Chapitre 6

Modélisation de la sinistralité

Comme vu en partie 1.5, le risque santé représente le risque pour un assuré d'effectuer des dépenses médicales couvertes par la complémentaire santé. Ce risque diffère pour chaque acte, c'est pourquoi la modélisation de la sinistralité doit être distincte par type d'acte.

Ainsi, l'objectif est de modéliser la fréquence et le coût d'un acte. Il convient dans un premier temps de choisir la loi de distribution la mieux adaptée pour cet acte. Enfin, il faudra choisir le modèle GLM le plus optimal par une sélection de variables.

6.1 Modélisation de la fréquence

6.1.1 Choix de la loi modélisant la fréquence

Les lois testées pour modéliser la fréquence sont la loi Poisson, la loi Binomiale négative, la loi Zero-inflated Poisson (ZIP) et la loi Zero-inflated Binomiale négative (ZINB). Comme vu en partie 3.1.2, la loi Poisson est caractérisée par l'égalité entre l'espérance et la variance. Ainsi, dans un premier temps, il convient de regarder l'espérance et la variance de la fréquence de cet acte. Par la suite, en ajustant les distributions selon les lois testées, les indicateurs de performance permettent de choisir la loi la plus adéquate.

L'acte *Consultations spécialistes*, du poste Consultations et visites, est consommé en 2019 (année de survenance) par 56% de la population des actifs. Ceci est élevé, mais laisse un certain nombre d'observations nulles dans le portefeuille, c'est-à-dire de personnes n'ayant pas consommé cet acte. L'ajustement de la distribution de la fréquence a été effectué à l'aide du package *fitdistrplus* du logiciel *R*.

Les statistiques de cet acte pour chaque loi se trouvent dans le tableau 6.1 ci-dessous.

Statistique	Poisson	Binomiale négative	ZIP	ZINB
Espérance	2.49			
Variance	23.27			
AIC	184 311	135 455	157 984	134 281
BIC	184 319	135 472	158 001	134 306

TABLE 6.1 – Statistiques par loi de distribution de la fréquence

Pour cet acte, la moyenne est très inférieure à la variance. La fréquence de visites médicales peut varier selon l'âge, le genre, les besoins des individus, et entre autres l'offre de soins, qui ne

sont pas les mêmes pour chacun, d'où une variance plutôt haute. Cela éloigne l'hypothèse des lois Poisson et ZIP.

L'AIC pour la loi Poisson est bien plus élevé que pour la loi Binomiale négative, même lorsque les observations nulles sont ajustées via la loi ZIP. Ces deux lois ne pourront pas être retenues. Étant donné une proportion d'observations nulles modérée (44%), la loi ZINB donne, de peu, un meilleur AIC et BIC que la loi Binomiale négative.

La figure 6.1 présente l'histogramme de la fréquence, sur lequel s'ajoutent les courbes ajustées à ces lois de distributions.

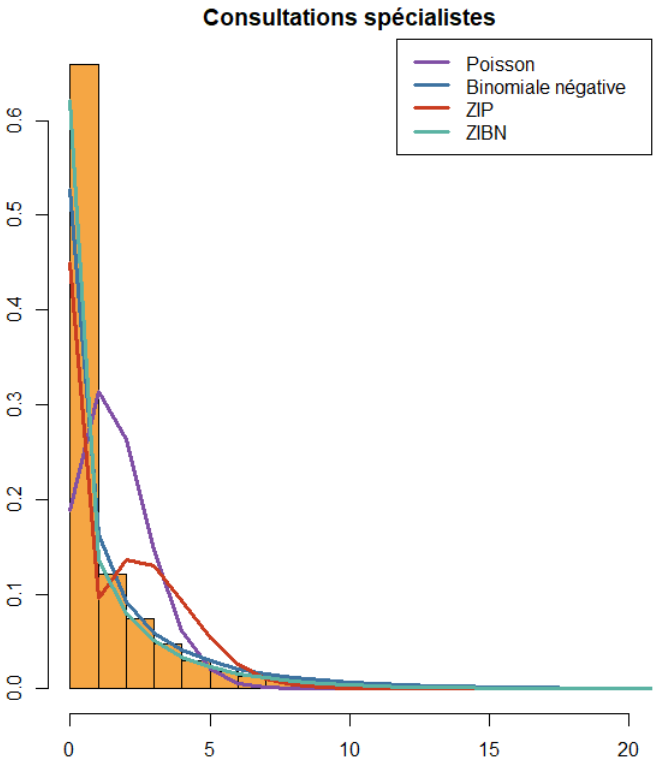

FIGURE 6.1 – Distribution de la fréquence

Dans cette figure, on peut voir que les observations nulles sont sous-estimées par les lois Poisson et ZIP. Les valeurs de 1 à 5 sont quant à elles surestimées. Ajouté à l'écart conséquent entre l'espérance et la variance, ces lois peuvent difficilement être retenues pour un poste ayant beaucoup d'observations.

La courbe bleue, correspondant à la loi Binomiale négative, semble mieux s'adapter à la distribution. La différence avec la courbe verte, correspondant à la loi ZINB, n'est pas significative. Le nombre de 0 semble être mieux estimé par la loi ZINB.

C'est donc la loi ZINB qui sera retenue.

6.1.2 Sélection des variables pour modéliser la fréquence

Dans notre modèle contraint, on cherche à expliquer la sinistralité à partir des variables tarifaires non-spatiales. Ces variables sont peu nombreuses, mais finalement assez communes dans la tarification.

Les variables testées sont les suivantes :
— L'âge (de 0 à 70 ans) ;
— Le sexe (F ou M) ;
— Le type de population (Actif, Invalide, Conjoint survivant d'actif de plus de 55 ans, Conjoint survivant d'actif de moins de 55 ans, Enfant orphelin, Enfant cotisant, ou Actif en congé parental) ;
— Le lien familial (Assuré, Conjoint ou Enfant).

Cependant, l'âge doit être classifié de sorte à ne pas complexifier le modèle, et avoir des classes d'âge les plus hétérogènes possible, tout en conservant la croissance des âges au sein des classes.

Par la suite, une sélection de variables a été effectuée pour chaque acte. Cette partie présente les résultats issus de la méthode forward.

Classification de la variable Âge

Un arbre de décision se basant sur la fréquence de consommation et l'exposition globales par âge a été construit. Cet arbre peut se construire à l'aide des packages *rpart* et *rpart.plot* du logiciel *R*. La figure 6.2 présente l'arbre de décision classifiant les âges.

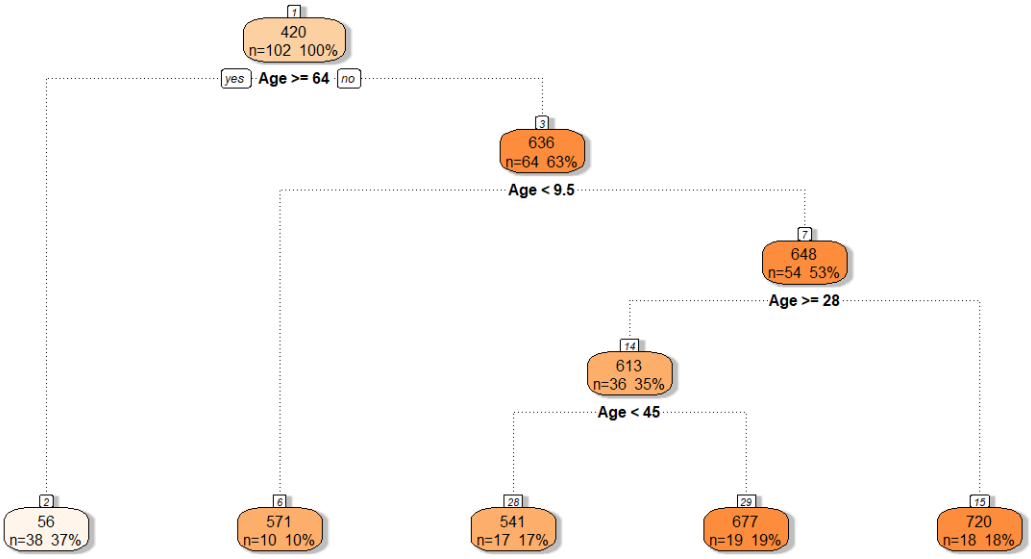

FIGURE 6.2 – Arbre de classification des âges

Cet arbre de décision généré par *R* suggère la classification suivante :
— Classe 1 : de 0 à 9 ans ;
— Classe 2 : de 10 à 27 ans ;
— Classe 3 : de 28 à 44 ans ;
— Classe 4 : de 45 à 63 ans ;
— Classe 5 : 64 ans et plus.

Chacune de ces classes est assez riche en données, bien que la classe 5 des plus de 63 ans ne représente que 3% du portefeuille.

Cependant, comme vu en partie 2, les risques ne sont pas les mêmes au sein de la classe 2 des 10-27 ans. En effet, les assurés de 10 à 16 ans consomment davantage les actes dentaires,

alors que les assurés de 17 à 27 ans consomment bien davantage les frais médicaux de ville et produits pharmaceutiques. Cette classe étant la plus volumineuse avec 32% du portefeuille, on peut se permettre de la séparer en deux de façon à distinguer les assurés de moins de 17 ans des plus de 17 ans.

Ainsi, la classification des âges est reportée dans le tableau 6.2.

Tranche d'âge	Classe	Répartition
0 - 9	1	14%
10 - 16	2	12%
17 - 27	3	19%
28 - 44	4	22%
45 - 63	5	29%
> 63	6	3%

TABLE 6.2 – Classification des âges par tranche

Cette classification sera utilisée, en substitution de l'âge brut, pour la sélection de variables à venir.

Sélection de variables

La méthode de sélection Forward, expliquée en partie 3.2.4, a été mise en pratique. Les résultats en fonction des variables sont indiqués dans le tableau 6.3 ci-dessous.

Variables	AIC	BIC	MSE
$\emptyset$	134 281	134 306	9.030
Classe d'âge	133 252	133 364	8.8923
Classe d'âge+Sexe	132 244	132 372	8.767
Classe d'âge+Sexe+Lien familial	132 012	132 192	8.7616
Classe d'âge+Sexe+Lien familial+Population	132 082	132 365	8.756

TABLE 6.3 – Sélection de variables

Les résultats indiquent une amélioration de l'AIC après l'ajout des trois premières variables, mais une légère détérioration après ajout de la variable *Population*. En effet, cette variable contient 7 modalités (puisqu'on se restreint à la catégorie A. Cf : tableau 2.2), mais 99% des observations ont pour modalité « ACTIFS ». Par conséquent, le pouvoir discriminant de cette variable est assez faible. Nonobstant l'amélioration de la MSE qu'elle engendre, sa complexité pénalise les critères d'information.
La variable *Population* ne sera pas conservée.

Par ailleurs, il est légitime de s'interroger sur une éventuelle corrélation entre la variable *Lien familial* et *Classe d'âge*. D'une part, nous avons vu en partie 2 que les conjoints sont en moyenne plus âgés que les assurés principaux, mais surtout, il est clair que les deux premières classes d'âge (de 0 à 16 ans) correspondent parfaitement avec la modalité Enfant.

En croisant les deux variables, on observe le taux d'appartenance de part et d'autre des modalités de ces deux variables (voir tableau 6.1.2).

	Assuré	Autre bénéficiaire	Conjoint	Enfant
1	0.00	0.00	0.00	13.70
2	0.00	0.00	0.00	12.20
3	8.20	0.00	0.20	11.40
4	20.70	0.00	1.50	0.00
5	27.00	0.00	2.50	0.00
6	2.00	0.00	0.60	0.00

TABLE 6.4 – Croisement entre les variables *Classe d'âge* et *Lien familial*

Ce tableau indique une correspondance presque parfaite entre les classes d'âge 1, 2 et 3, et le lien Enfant. De même, une importante correspondance est observable entre la classe d'âge 5 et le lien Conjoint.

Le V de Cramer (expliqué en Annexe B.4.4) permet de mesurer le niveau de dépendance entre deux variables. Plus la dépendance entre deux variables est forte, plus il s'approche de 1. Plus cette dépendance est faible, plus il s'approche de 0.

Ici V de Cramer entre la classe d'âge et le lien est de 0.52, ce qui est trop élevé pour considérer les deux variables comme indépendantes, et suffisant pour estimer une dépendance.

En conséquence, la variable *Lien familial* sera également retirée.

Finalement, les variables non-spatiales conservées pour la modélisation de la fréquence de l'acte *Consultations spécialistes* sont la classe d'âge et le sexe.

6.1.3 GLM fréquence : modèle contraint

Les variables et la loi de distribution ayant été choisies, il est possible de modéliser la fréquence de survenance en utilisant un modèle linéaire généralisé.

La modélisation de la fréquence pour l'acte *Consultations spécialistes* a été effectuée en utilisant la loi ZINB à l'aide de la fonction *zeroinfl* du package *pscl* sur le logiciel *R*.

L'AIC obtenu est de 132 244, et le BIC de 132 372. Ce dernier est plus élevé car pénalisé par le nombre élevé de lignes.

Le tableau 6.5 présente les coefficients associés à chaque modalité de variable pour la modélisation de la fréquence. Il s'agit des coefficients donnés par le logiciel *R* via la fonction *zeroinfl* du package *pscl*.

Variables	Modalité	Coefficient	p-value
Intercept	(intercept)	0.84882	$<$ 2e-16
Classe d'âge	2	-0.35873	$<$ 2e-16
Classe d'âge	3	-0.48693	$<$ 2e-16
Classe d'âge	4	-0.10277	0.000116
Classe d'âge	5	0.03687	0.127915
Classe d'âge	6	-0.13507	0.010509
Sexe	M	-0.21282	$<$ 2e-16

TABLE 6.5 – Coefficients GLM de la loi ZINB

L'intercept contient la modalité 1 pour la variable *Classe d'âge* et F pour la variable *Sexe*.

La colonne Coefficient correspond au coefficient linéaire associé aux modalités respectives. Elle représente les β_i de la formule 3.7. Ainsi, plus un coefficient est élevé, plus la modalité associée a d'influence sur la variable réponse (ici la fréquence). S'il est négatif, la modalité associée a une influence négative sur la variable réponse. A titre d'exemple ici, les hommes ont une influence négative sur la fréquence puisqu'ils consomment en moyenne moins que les femmes, comme vu en partie 2, c'est pourquoi leur coefficient linéaire est négatif.

La p-value permet d'accepter l'hypothèse de significativité d'une modalité. Si elle est inférieure à 0.05, on peut accepter l'hypothèse évoquant que la modalité est significative, avec 5% de chance de se tromper.

Ainsi, l'âge et le sexe sont très influents sur la fréquence. Précisément, les classes d'âge 2 et 3 (soit la tranche d'âge de 10 à 27 ans) influent négativement sur la fréquence. C'est-à-dire qu'elles font baisser significativement la fréquence globale par rapport à la classe 1 (0 à 9 ans) qui influe fortement positivement, puisque cet acte inclut les actes de pédiatrie.

Pour se conforter dans la pertinence de ces coefficients, on peut confronter la fréquence moyenne prédite par modalité de variable à la fréquence moyenne observée par modalité (voir figures 6.3 et 6.4)

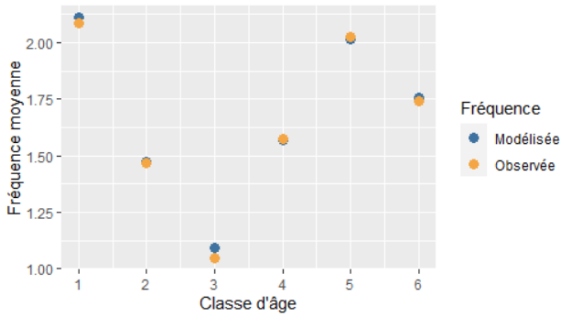

FIGURE 6.3 – Fréquence moyenne par classe d'âge modélisée vs observée

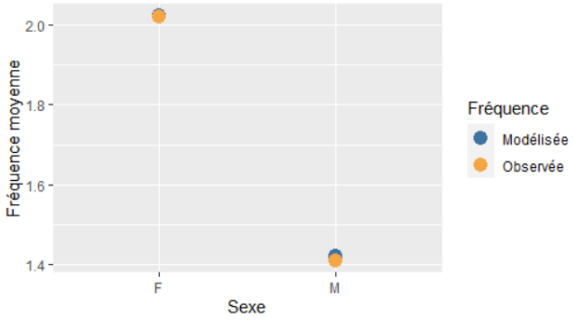

FIGURE 6.4 – Fréquence moyenne par sexe modélisée vs observée

Ces graphiques appuient la cohérence des coefficients GLM, mais aussi de la classification des âges effectuée en partie 6.1.2 puisque d'une classe à l'autre, la fréquence moyenne varie fortement.

Ainsi, la fréquence a été modélisée via le modèle contraint pour l'acte *Consultations spécialistes*. La suite vise à modéliser le coût.

6.2 Modélisation du coût

Dans le cadre de ce mémoire, parler de coût revient à parler de frais réels d'une prestation.

Pour modéliser le coût, le schéma est globalement le même que celui utilisé pour la fréquence. Toutefois, le choix de la loi de distribution peut être influencé par des tests statistiques.

Dans cette sous-partie, on se restreint au portefeuille des assurés ayant consommé au moins une fois l'acte étudié.

6.2.1 Choix de la loi modélisant le coût

Deux lois seront testées pour modéliser le coût : la loi Gamma et la loi Log-normale. Pour choisir la loi la plus adéquate, on se base sur les statistiques des tests de Kolmogorov-Smirnov et Cramer-Von Mises (qui permettent de tester l'ajustement d'une distribution observée à une distribution théorique continue), mais aussi sur les critères d'information.

Du fait d'un grand nombre de données et de valeurs redondantes, on peut s'attendre à ce que la p-value de ces tests tende rapidement vers 0. Or, cette statistique ne doit pas être décisive dans le choix de notre loi. En effet, des études constatent que l'observation de la p-value perd de sa pertinence lorsque celle-ci est issue de tests effectués sur des données importantes [1] car celle-ci s'approchera souvent de 0.

La modélisation du coût de l'acte *Consultations spécialistes* est délicate, dans la mesure où une certaine partie des médecins spécialistes sont conventionnés et ont par conséquent des tarifs encadrés.

En conséquence, certains montants de coûts moyens seront plus représentés que d'autres, tant et si bien que l'ajustement d'une loi peut être éloigné de l'aspect théorique de celle-ci.

Toutefois, comme vu en partie 2, le coût de cet acte reste relativement disparate, aussi bien au niveau des zones qu'à celui des individus.

Les tests de Kolmogorov-Smirnov et Cramer-Von Mises ont été effectués à l'aide de la fonction *gofstat* du package *fitdistrplus* du logiciel *R*. Ces tests sont développés en Annexe D.1.

Les statistiques obtenues à l'issu de ce test sont reportées dans le tableau 6.6 ci-dessous.

Statistique	Gamma	Log-normale
Statistique Kolmogorov-Smirnov	0.0997	0.0955
Cramer-Von Mises	33.0480	32.5219
AIC	175 297	174 318
BIC	175 313	174 334

TABLE 6.6 – Statistiques par loi de distribution du coût

Les statistiques vont unanimement en faveur de la loi Log-normale.

Or, en observant la distribution des montants, on voit que les deux lois sont difficilement ajustables. La figure 6.5 présente la distribution des coûts moyens sous forme d'histogramme, à laquelle s'ajoutent les courbes de densité des lois ajustées.

1. Lin, M., Lucas, H. C., Jr., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. Information Systems Research

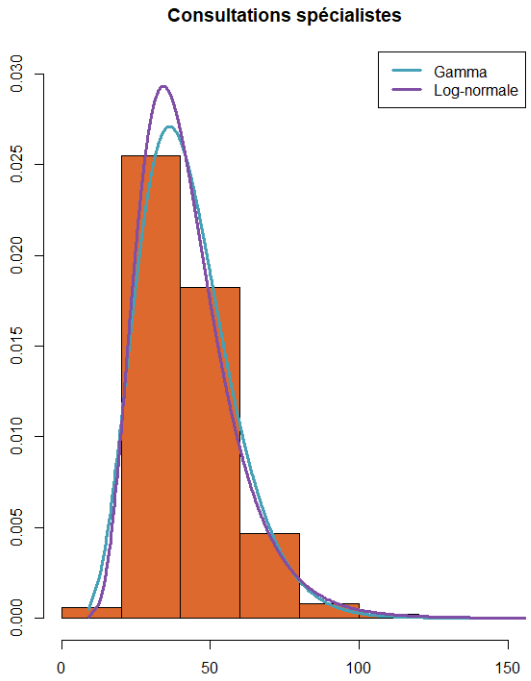

FIGURE 6.5 – Distribution du coût [2]

Les deux distributions théoriques sont difficilement ajustables à notre échantillon.

La différence n'est pas visuellement considérable, ce qui confirme que les statistiques de tests reportées dans le tableau 6.6 soient si proches. On remarque que la loi Log-normale atteint sa valeur maximale avant la loi Gamma, et est ainsi mieux ajustée aux montants les plus représentés.

La plus grande bande de l'histogramme correspond à une fourchette particulièrement représentée dans le portefeuille. Le tarif d'une consultation de médecin spécialiste conventionné de secteur 1 s'élève à 25,00€ [3]. Ce montant additionné aux éventuelles majorations, on retrouve 30% des coûts moyens compris entre 25,00 € et 30,00 €.

C'est la loi Log-normale qui sera retenue pour la modélisation du coût de l'acte *Consultations spécialistes*.

6.2.2 Sélection des variables pour modéliser le coût

La sélection des variables a été effectuée par les méthodes Forward et Backward, développées dans la partie 3.2.4. Les variables conservées étant les mêmes à l'issu des deux méthodes, nous ne détaillerons que les résultats de la méthode Backward.

Comme pour la fréquence, les variables testées seront : la classe d'âge (détaillée en partie 6.1.2), le sexe, le lien familial, et le type de population.

Les statistiques du modèle GLM, utilisant la loi Log-normale pour modéliser le coût, évoluent à chaque retrait de variable. Ces statistiques sont reportées dans le tableau 6.7.

2. La figure 6.5 représente la distribution des valeurs de chaque montant. L'aire des rectangles somme donc à 1.

3. Améli, « Consultations en métropole : vos remboursements »

Variables	AIC	BIC	MSE
Classe d'âge+Sexe+Population+Lien	17 590	17 734	286.9
Classe d'âge+Sexe+Population	17 595	17 714	287.1
Classe d'âge+Sexe+Lien	17 609	17 705	287.3
Classe d'âge+Sexe	17 608	17 672	287.5
Classe d'âge	17 647	17 703	287.9
$\emptyset$	18 239	18 255	294.7

TABLE 6.7 – Sélection de variables

On constate que le retrait de la variable *Lien* améliore le BIC. Ce qui signifie que cette variable pénalise le modèle. Cela peut s'expliquer par le fait que le coût d'une consultation spécialiste ne change pas, ou peu, selon que le bénéficiaire soit un assuré principal, un conjoint ou un enfant. Bien qu'elle améliore dans une moindre mesure l'erreur quadratique, il ne serait pas actuariellement optimal de conserver cette variable.

Le constat et la conclusion sont les mêmes pour la variable *Population*.

Les variables conservées pour la modélisation du coût de l'acte *Consultations spécialistes* sont donc : *Classe d'âge* et *Sexe*.

6.2.3 GLM coût : modèle contraint

Pour réaliser un GLM sur le coût avec la loi Log-normale, d'un point de vue modélisation, il suffit d'utiliser la fonction *glm* du package *MASS* du logiciel *R*, en prenant pour famille de distribution la loi Normale avec *log* pour fonction lien, et cherchant à expliquer le logarithme du coût par les variables explicatives retenues.

Ainsi, les coefficients obtenus à l'issu de ce modèle sont reportés dans le tableau 6.8 ci-dessous.

Variables	Modalité	Coefficient	p-value
Intercept	(intercept)	3.671799	$<$ 2e-16
Classe d'âge	2	-0.097803	$<$ 2e-16
Classe d'âge	3	-0.028008	0.00251
Classe d'âge	4	0.012594	0.13273
Classe d'âge	5	0.078419	$<$ 2e-16
Classe d'âge	6	0.154786	$<$ 2e-16
Sexe	M	-0.030263	2.09e-09

TABLE 6.8 – Coefficients GLM de la loi Log-normale

A travers la table 6.8, il apparaît que la modalité « M » de la variable *Sexe* influe négativement sur le coût, même si son coefficient linéaire est très faible. En observant de près, on constate que, pour cet acte, les hommes ont des frais réels moyens de 41 €, et les femmes de 42 €, ce qui explique ce coefficient soit (très faiblement) négatif, et pertinent étant donné la quantité de données.

La classe d'âge 2, correspondant à la tranche d'âge de 9-16 ans, influe aussi négativement sur le coût, et pour cause : cette classe d'âge a le coût le plus faible, c'est pourquoi son coefficient linéaire est le plus faible, et le plus significatif. De surcroît, on remarque qu'à partir de cette classe, les coefficients augmentent avec l'âge.

Comme avec l'âge, on peut se conforter dans la pertinence de ces coefficients en observant le coût moyen modélisé par modalité de variable face au coût moyen observé (voir figures 6.6 et

6.7)

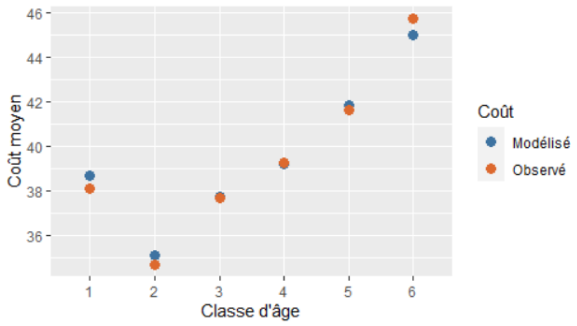

FIGURE 6.6 – Coût moyen par classe d'âge modélisé vs observé

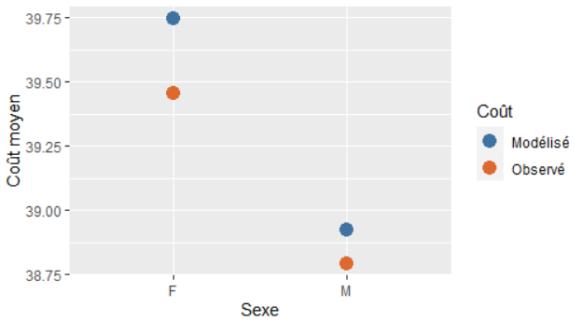

FIGURE 6.7 – Coût moyen par sexe modélisé vs observé

De même, ces graphiques appuient la cohérence des coefficients GLM obtenus. Bien qu'un léger décalage soit observable, la tendance est la même d'une modalité à l'autre.

Conclusion

A l'issu de ce chapitre, la sinistralité a été modélisée en vue d'estimer la fréquence et le coût. Si la pertinence de la modélisation du coût reste discutable pour l'acte *Consultations spécialistes*, étant donné la difficulté à trouver une distribution théorique semblable, nous verrons plus tard que celle-ci peut considérablement être améliorée après enrichissement d'une variable géographique.

L'erreur de modélisation qui résulte de notre modèle, appelée résidu, contient un élément explicatif qui n'a pas été pris en compte dans nos précédents modèles : l'effet spatial. Cet effet peut être capté en vue d'améliorer les modèles contraints. Le chapitre suivant portera sur cette erreur de modélisation.

Chapitre 7

Étude des résidus

Ce chapitre détaille les étapes finales de construction d'un zonier. Pour relativiser le besoin de granularité dans l'information géographique, ce zonier sera construit de part et d'autre sous deux mailles géographiques : une maille départementale, et une maille ville (ou code postal).

7.1 Agrégation et analyse des résidus

Dans cette sous-partie sera étudié le résidu issu du modèle contraint. Il s'agit à ce niveau d'un résidu ligne à ligne. Celui-ci représente la différence entre la sinistralité observée et la sinistralité prédite, qu'on appelle résidu classique.

Ainsi, le $i^{ème}$ résidu se calcule comme suit (voir équation 7.1) :

$$r_i = \textit{Sinistralité observée}_i - \textit{Sinistralité prédite}_i \tag{7.1}$$

La sinistralité peut concerner la fréquence, le coût, ou la prime pure.

Les résidus de chaque assuré sont par la suite agrégés à chaque zone, selon la maille choisie. Pour les agréger à une zone, il convient de les pondérer par l'exposition au sein de celle-ci. Ainsi, pour une zone k ayant p individus, le résidu R_k de cette zone s'exprime comme suit (voir équation 7.2) :

$$R_k = \frac{\sum_{i=1}^{p} r_{k,i}}{\sum_{i=1}^{p} e_{k,i}} \tag{7.2}$$

Avec $e_{k,i}$ l'exposition du $i^{ème}$ individu de la zone k, et $r_{k,i}$ le résidu du $i^{ème}$ résidu de la zone k.

Cependant, selon la maille choisie, toutes les zones ne seront pas habitées ou sinistrées. Au sein de notre population d'actifs, toutes les régions sont habitées par au moins un assuré, mais 1 département de l'hexagone ne l'est pas.

Par ailleurs, à l'issu de notre simulation de codes postaux (Cf : Partie 2.2.4), 2 501 codes postaux sont exposés, soit environ 40% des codes postaux métropolitains.

7.1.1 Résidus Fréquence

Dans un premier temps, nous observons les résidus issus du modèle GLM fréquence contraint, c'est-à-dire n'intégrant aucune variable explicative géographique.

L'acte *Consultations spécialistes* est représenté dans 92 départements différents, et 1853 codes postaux. Nous allons observer la distribution des résidus de la fréquence agrégés à la maille départementale dans un premier temps, puis à la maille code postal dans un second temps.

Maille départementale

En traçant chaque résidu fréquence par ordre croissant, les valeurs extrêmes de ceux-ci ressortent davantage. La figure 7.1 représente la distribution des résidus par département.

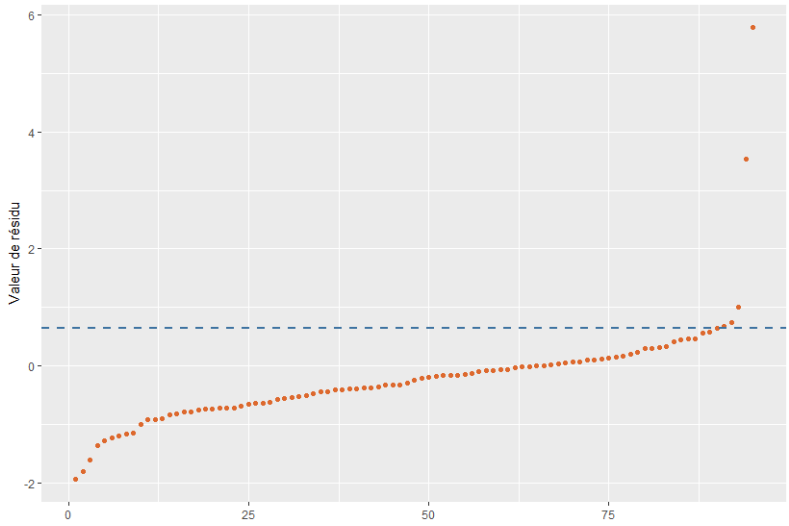

FIGURE 7.1 – Distribution des résidus de la fréquence à la maille départementale

La ligne pointillée bleue correspond au quatre-vingt-quinzième centile. Les points au dessus correspondent donc aux résidus supérieurs à 95% de l'ensemble des résidus.

Le tableau 7.1 présente les statistiques sur ces résidus.

Valeur min	Quantile 1	Médiane	Moyenne	Quantile 3	Valeur max
-1.94	-0.67	-0.25	-0.20	0.08	5.78

TABLE 7.1 – Statistiques des résidus de la fréquence à la maille départementale

La valeur maximale correspond au dernier point de la figure 7.1, qui se distingue nettement des autres. Ce dernier correspond au Territoire de Belfort, qui a une exposition de 6, et au sein duquel un assuré a effectué 17 actes de consultations spécialistes, et deux en ont effectué 10. Cela révèle le problème des zones faiblement exposées mais sinistrées. Le manque d'expérience ne permet pas de mutualiser chaque zone à la même échelle. Un lissage devra donc être effectué pour capter plus justement le risque spatial intrinsèque.

En effectuant la même étude à la maille code postal, les constats sont sensiblement les mêmes, à une proportion différente.

Maille code postal

La figure 7.2 présente la distribution des résidus fréquence à la maille code postal.

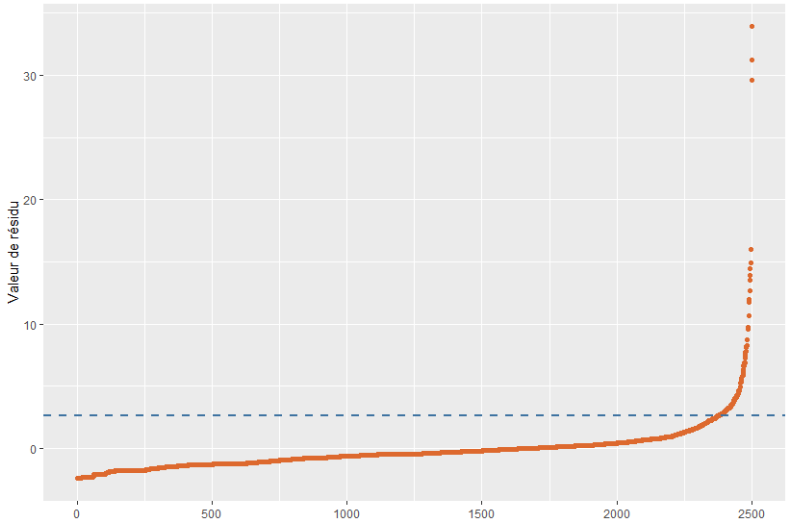

FIGURE 7.2 – Distribution des résidus de la fréquence à la maille code postal

Pour cette maille, on distingue les résidus qui dépassent le 95e centile, et lorsqu'ils le dépassent, ils s'en éloignent rapidement. Les statistiques de ces résidus sont données dans le tableau 7.2 ci-dessous.

Valeur min	Quantile 1	Médiane	Moyenne	Quantile 3	Valeur max
-2.43	-1.21	-0.45	-0.15	0.22	33.91

TABLE 7.2 – Statistiques des résidus de la fréquence à la maille code postal

La valeur maximale correspond à une ville n'ayant qu'un résident qui a effectué 36 actes de consultations spécialistes. Il n'est pas judicieux de considérer que cet individu représente le risque global de cette zone.

Plus généralement, la moyenne des expositions des villes dont le résidu dépasse le 99e centile est de 1.14. Mathématiquement :

$$\mathbb{E}[e|R > R_{99\%}] = 1.14$$

Ainsi, le biais lié à la faible exposition est d'autant plus conséquent que la maille géographique est fine. En effet, les résidus agrégés à une zone seront plus ou moins neutralisés par l'exposition selon que celle-ci soit importante.

Avant le lissage, les valeurs extrêmes doivent être retraitées. En effet, malgré la faible exposition qu'elles représentent, ces dernières peuvent à la fois « résister » au lissage, et parasiter ce dernier en étendant leur biais sur leurs voisins.

Ainsi, les valeurs supérieures au 98e centile seront retirées, tout en sachant que la moyenne des expositions des zones concernées est de 1.50 pour la maille code postal, et de 11.95 pour la maille départementale.

En dehors des valeurs extrêmes, on constate que plus de la moitié des résidus sont négatifs, ce qui signifie que la sinistralité par zone a été surestimée. Ce phénomène peut conduire d'une autre manière à l'antisélection. En effet, sur-tarifer un « bon risque » pourrait pousser ce dernier à se tourner vers un assureur concurrent. Ainsi, le zonier a aussi pour but de classifier ces bons risques et ajuster leur prime pure en conséquence.

Ainsi, après avoir étudié et traité l'erreur de modélisation de la fréquence, il convient d'en faire de même pour le coût.

7.1.2 Résidus coût

Pour le calcul des résidus du coût par zone, il convient d'appliquer la formule 7.2, mais cette fois-ci en ne tenant compte que de l'exposition des personnes ayant consommé au moins une fois l'acte. Ainsi, au sein d'une zone k ayant p individus, le résidu R_k s'écrit :

$$R_k = \frac{\sum_{i=1}^{p} r_{k,i} \times \mathbb{1}_{\{l'individu\ i\ a\ consommé\}}}{\sum_{i=1}^{p} e_{k,i} \times \mathbb{1}_{\{l'individu\ i\ a\ consommé\}}} \tag{7.3}$$

Maille départementale

En traçant chaque résidu « coût » par ordre croissant, les valeurs extrêmes de ceux-ci ressortent davantage. La figure 7.3 représente la distribution des résidus par département.

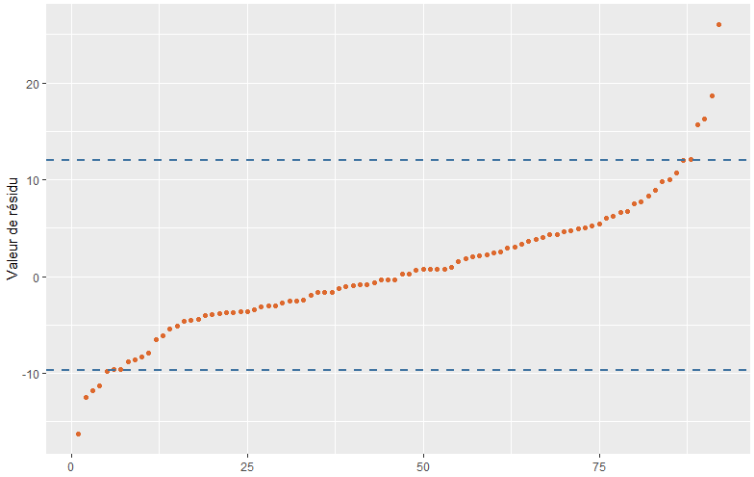

FIGURE 7.3 – Distribution des résidus coût à la maille départementale

La ligne pointillée bleue du haut correspond au 95^{e} centile, celle du bas au 5^{e} centile. La queue de distribution semble à première vue moins lourde que pour la fréquence.

Le tableau 7.3 présente les statistiques sur ces résidus.

Valeur min	Quantile 1	Médiane	Moyenne	Quantile 3	Valeur max
-16.23	-3.65	-0.045	0.596	4.43	26.02

TABLE 7.3 – Statistiques des résidus coût à la maille départementale

On observe ici un problème de valeurs extrêmes à gauche et à droite. Le résidu minimal de -16.23 se trouve dans l'Aube (10), qui contient 4 assurés, dont un a effectué 4 actes aux frais réels moindres. Son coût moyen a par conséquent été surestimé, ce qui explique ce résidu négatif. Dans l'ensemble, la moyenne de l'exposition des départements dont les résidus sont en dessous du 5^{e} est de 3.665. En revanche, au dessus du 95^{e} quantile, la moyenne des expositions est de 156.12, ce qui signifie que ces résidus ne sont pas tous des valeurs aberrantes.

Pour minimiser la perte d'information, seules les valeurs supérieures au 99^{e} centile et inférieures au deuxième centile seront retirées, soit les valeurs minimales et maximales.

Effectuons à présent la même étude à la maille code postal.

Maille code postal

La figure 7.4 présente la distribution des résidus coût.

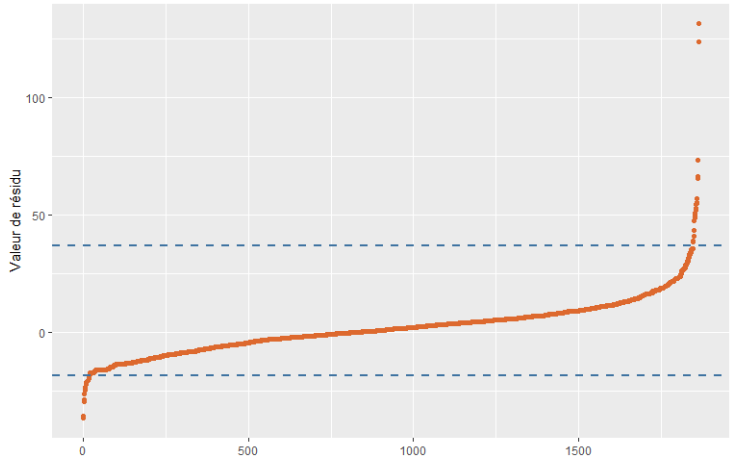

FIGURE 7.4 – Distribution des résidus coût à la maille code postal

Pour cette maille, on distingue plus clairement les valeurs extrêmes à gauche, comme à droite. Les lignes pointillées bleues correspondent au deuxième centile pour celle du bas, et avant dernier centile pour celle du haut. Le tableau 7.4 reporte les statistiques des résidus coût à la maille code postal.

Valeur min	Quantile 1	Médiane	Moyenne	Quantile 3	Valeur max
-36.79	-5.24	1.23	1.86	7.17	131.36

TABLE 7.4 – Statistiques des résidus coût à la maille code postal

Le résidu minimal de -36.79 correspond à une ville ayant un assuré avec une exposition de 0.32. Cet assuré a effectué un acte à 25€, mais sa faible durée d'affiliation fait de cette zone une valeur résiduelle extrême. Le résidu maximal de 131.36 correspond à une ville dont l'unique assuré résident a effectué un acte particulièrement coûteux.

Plus globalement, la moyenne des expositions de villes ayant un résidu en dessous du 5^{e} centile est de 1.22, et pour les villes au dessus du 95^{e} centile, cette moyenne est de 2.56. Ces dernières sont considérées comme valeurs extrêmes et seront retirées en vue d'effectuer le lissage spatial.

7.2 Lissage spatial des résidus

Après obtention des résidus par zone, deux problèmes sont rencontrés :
- Selon la maille géographique étudiée, toutes les zones ne seront pas exposées ou sinistrées, ce qui ne veut pas dire que le risque au sein de ces zones est nul.
- Certaines zones peu exposées peuvent se voir agréger un résidu non-représentatif de leur effet spatial.

Pour pallier ces limites, il convient d'attribuer à chaque zone une valeur de résidu plus ou moins semblable à ses voisins.

Afin d'attribuer une importance plus considérable aux zones fortement exposées, dont le résidu agrégé est a fortiori plus proche du risque réel selon la loi des grands nombres, on utilise la

théorie de la crédibilité.

Pour rappel, la formule de crédibilité vise à donner à chaque résidu R_i une nouvelle valeur R_i^*. Pour un total de p zones, R_i^* s'exprime comme suit :

$$R_i^* = Z_i \times R_i + (1 - Z_i) \times \frac{\sum_{j=1}^{p} R_j \times e_j \times f(d_{i,j})}{\sum_{j=1}^{p} e_j \times f(d_{i,j})}$$

Avec

— e_i la somme des expositions des assurés résidents dans la zone i
— $f(d_{i,j})$ une fonction monotone de la distance entre la zone i et la zone j
— Z_i le facteur de crédibilité associé à la zone i.

Pour un lissage adéquat, il convient de choisir le bon facteur de crédibilité Z. En effet, un facteur de crédibilité trop proche de 1 reviendrait à ré-attribuer à une zone un nouveau résidu bien trop proche de son résidu initial, auquel cas le lissage aurait peu d'intérêt. Par ailleurs, choisir un coefficient trop proche de 0 reviendrait à donner à chaque valeur un résidu dépendant de l'ensemble des autres, ce qui revient à un lissage uniforme, auquel cas le zonier en lui-même perdrait de son intérêt.

Ainsi, pour chaque zone i, le facteur de crédibilité Z_i, qui dépend de l'exposition e_i de la zone, s'exprime comme suit (voir équation 7.4) :

$$Z_i = \frac{e_i}{e_i + a} \tag{7.4}$$

Où a est un réel à fixer. L'enjeu repose donc sur le choix du coefficient a.

Pour mieux comprendre l'importance de ce choix, un lissage des résidus à la maille code postal a été effectué sous différentes intensités. La figure 7.5 reporte cartographiquement les résidus du modèle fréquentiel de l'acte *Consultations spécialistes* lissés à la maille ville selon différents niveaux de lissage.

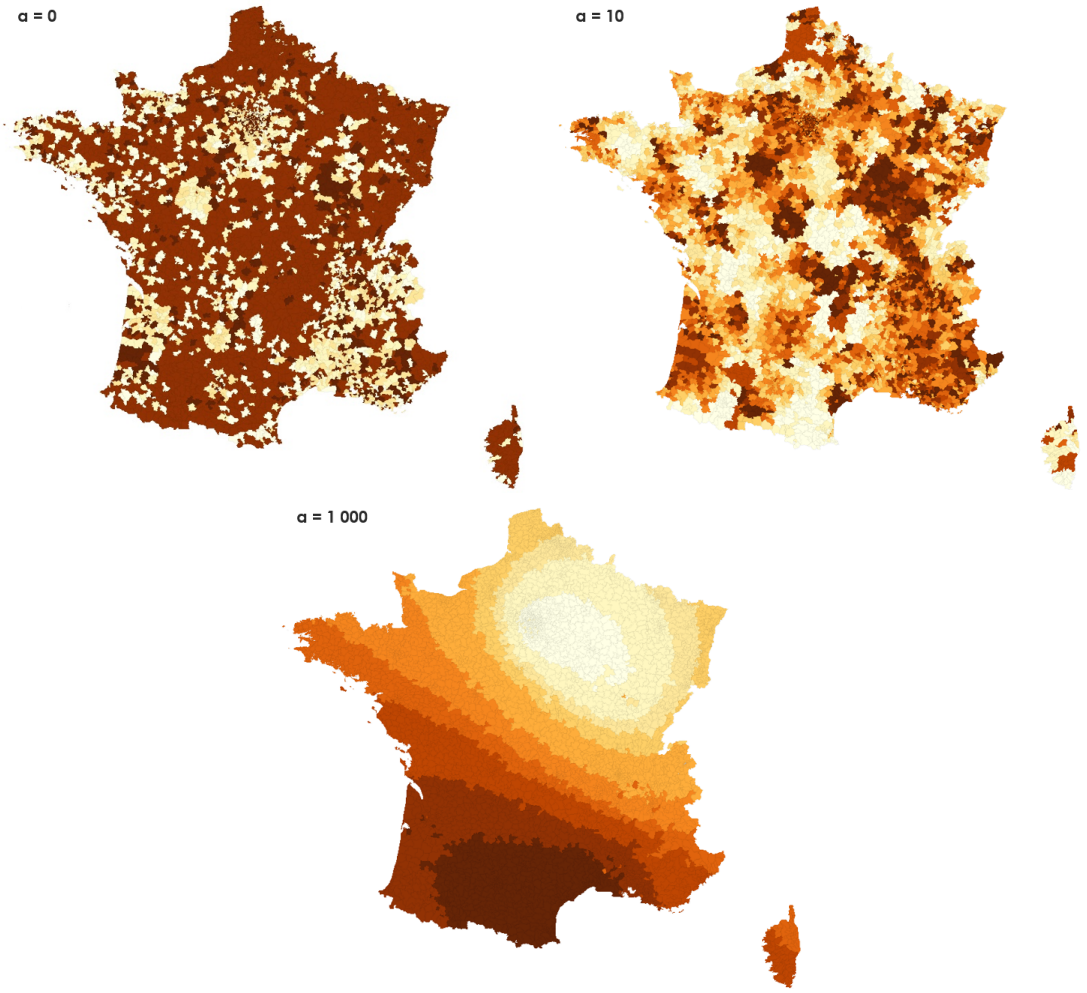

FIGURE 7.5 – Cartographie des résidus à la maille ville pour un lissage d'intensité $a = 0$, $a = 10$, et $a = 1000$

On constate qu'un lissage maximal (ou crédibilité totale) uniformise les résidus pour, à terme, ne plus distinguer aucune zone.

Pour la valeur de a, nous avons choisi le rapport entre la variance intra-zone moyenne et la variance inter-zone moyenne, la zone pouvant être le code postal ou le département. Ce choix s'est basé sur la littérature actuarielle [1].

Soit $\Theta = (\Theta_1, ..., \Theta_\ell)$ l'ensemble des zones, chaque zone Θ_k contenant un certain nombre $n_k \geq 1$ d'individus. Le coefficient a s'écrit comme suit (voir équation 7.5) :

$$a = \frac{\mathbb{E}\left[Var(r|\Theta)\right]}{Var(\mathbb{E}[r|\Theta])} \tag{7.5}$$

Le numérateur de l'équation 7.5 peut s'écrire de façon empirique comme suit (voir équation

1. Etienne Arbogast, Audrey Mahuzier, *Performance de la crédibilité*, 2014, p.5

7.6) :

$$\mathbb{E}\left[Var(r|\Theta)\right] = \frac{1}{\ell}\sum_{k=1}^{\ell} Var(r|\Theta_k) = \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{1}{n_k-1}\left(\sum_{j=1}^{n_k}\left(r_{k,j} - \frac{1}{n_k}\sum_{i=1}^{n_k} r_{k,i}\right)^2\right) \tag{7.6}$$

Le dénominateur de l'équation 7.5 peut quant à lui s'écrire de façon empirique comme suit (voir équation 7.7) :

$$\begin{aligned} Var(\mathbb{E}[r|\Theta]) &= \frac{1}{\ell-1}\sum_{k=1}^{\ell}\left(\mathbb{E}[r|\Theta_k] - \frac{1}{\ell}\sum_{j=1}^{\ell}\mathbb{E}[r|\Theta_j]\right) \\ &= \frac{1}{\ell-1}\sum_{i=1}^{\ell}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} r_{k,i} - \frac{1}{\ell}\sum_{j=1}^{\ell}\frac{1}{n_j}\sum_{i=1}^{n_j} r_{j,i}\right)^2 \end{aligned} \tag{7.7}$$

Enfin, la fonction de distance $f(d_{i,j})$ utilisée est l'inverse de la racine carrée de la distance euclidienne entre i et j, $f(d_{i,j}) = \frac{1}{\sqrt{d_{i,j}}}$, où la distance euclidienne (dont la formule est en Annexe C.1.2) est basée sur les coordonnées géographiques de leur centroïde respectif. Ce choix se fait de sorte à donner plus de crédit aux zones proches de i.

7.2.1 Lissage spatial des résidus fréquence

En appliquant la formule 7.5, à la maille départementale, on obtient un coefficient $a = 7.70$. En lissant chaque zone conformément à la formule 5.5, et en retraçant les résidus lissés sur une carte, on obtient la figure 7.6 pour la fréquence de l'acte *Consultations spécialistes*. Les nuances de couleurs sont ajustées par quantile à 20%.

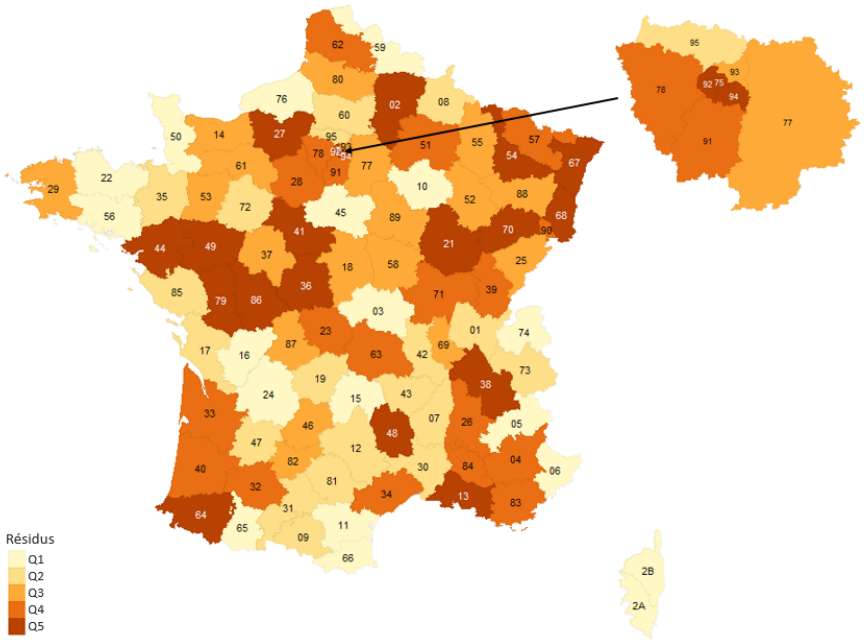

FIGURE 7.6 – Résidus fréquence lissés à la maille départementale

Manifestement, les zones à forts résidus sont principalement les départements disposant d'un site de l'entreprise E (notamment le 75, le 91 et le 38), ce qui est interprétable : les résidus

représentent la partie de la sinistralité inexpliquée par les variables non-spatiales, celle-ci est plus importante dans les zones fortement exposées, en l'occurrence les zones disposant de sites dans le cadre d'un portefeuille santé d'entreprise.

Un point est notable : le département 37, pourtant disposant d'un site et fortement exposé, n'est pas aussi foncé que ses voisins. Cela s'explique par le fait qu'une importante partie des assurés résidents ont consommé moins que le modèle contraint l'a prédit. Par ailleurs, en observant la fréquence moyenne par département, le département 37 se situe sous la médiane, ce qui en fait une zone à « bon risque » du point de vue assurantiel, c'est-à-dire une zone où la consommation est correctement estimée, ou éventuellement surestimée.

Pour la maille code postal, le coefficient obtenu est $a = 4.9$. En lissant chaque résidu conformément à la formule 5.5, et en retraçant les résidus lissés sur une carte, on obtient la figure 7.7 pour la fréquence de l'acte *Consultations spécialistes* à la maille ville. Les nuances de couleurs sont ajustées par quantile à 20%.

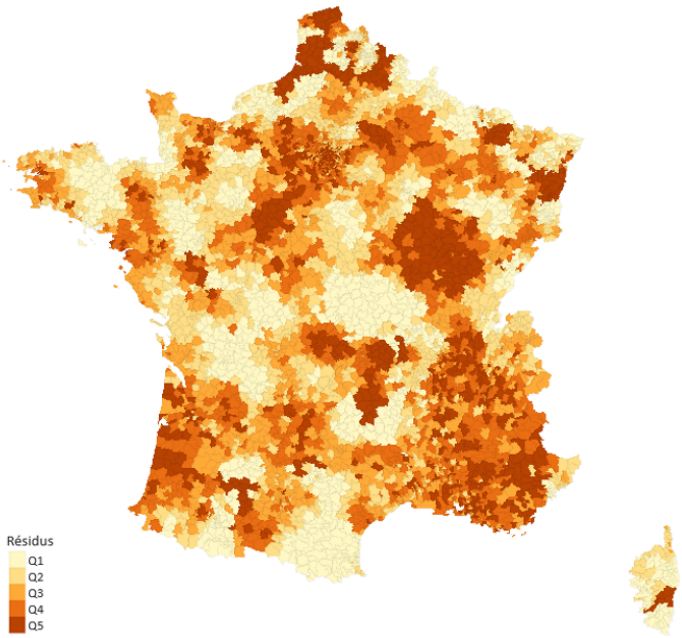

FIGURE 7.7 – Résidus fréquence lissés à la maille code postal

Malgré le lissage, il y a plusieurs zones de « déserts résiduels », c'est-à-dire des zones n'ayant aucun résidu lissé. Ces déserts font suite à un lissage sur des zones de déserts d'exposition. Typiquement, la région du Languedoc-Rousillon contient très peu d'observation, et par extension, les villes n'ayant pas, ou peu de voisins aux résidus non-nuls auront un lissage moindre. Par ailleurs, la côte sud-est est fortement représentée à travers cette carte.

Cependant, il est important de retenir ici qu'il s'agit des résidus « bruts » d'un effet inexpliqué, et que cette carte ne représente pas le risque spatial seul. De plus, les zones plus claires sont des zones dont les résidus appartiennent au premier quantile à 20%, donc des résidus négatifs (cf : tableau 7.2), soit des zones pour lesquelles la fréquence a été surestimée, ce qui en fait des zones à risque d'antisélection.

7.2.2 Lissage spatial des résidus coût

Maille départementale

En appliquant la formule 7.5 à la maille départementale, le coefficient obtenu est $a = 11.56$. En lissant chaque zone conformément à la formule 5.5, et en retraçant les résidus lissés sur une carte, on obtient la figure 7.8 pour le coût de l'acte *Consultations spécialistes*. Les nuances de couleurs sont ajustées par quantiles à 20%.

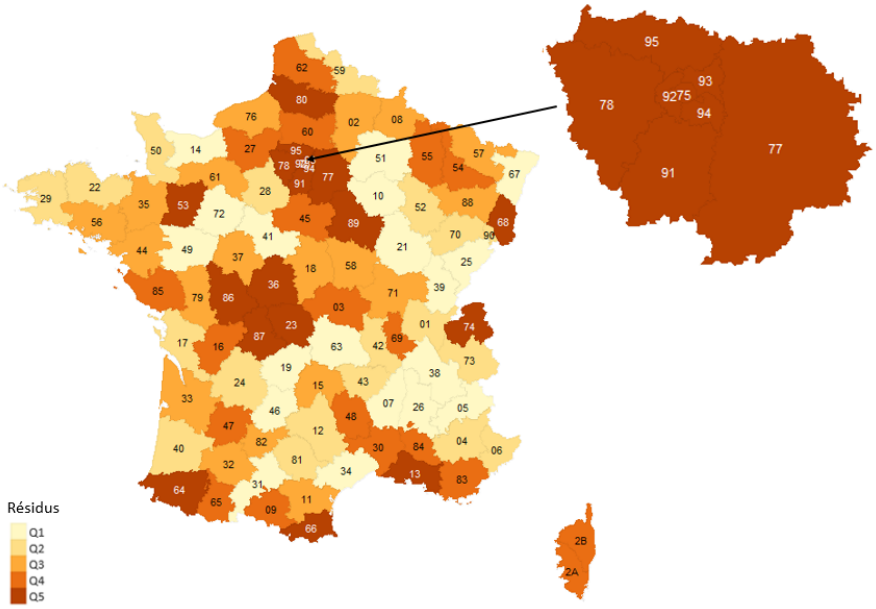

FIGURE 7.8 – Résidus coût lissés à la maille départementale

Comme on pouvait l'anticiper, l'Ile-de-France ressort fortement sur cette carte de résidus coût, ce qui signifie que les coûts moyens modélisés pour les résidents de cette zone ont été sous-estimés. L'Ile-de-France étant la région la plus urbaine, il n'est pas inconcevable que les frais réels moyens d'un spécialiste soit en moyenne plus cher qu'en province. De surcroît, au sein de notre portefeuille, les départements de Paris et des Hauts-de-Seine font parti des plus chers en termes de coûts moyens. Le constat est le même sur la côte méditerranéenne.

Maille code postal

Pour la maille code postal, le coefficient obtenu est $a = 3.5$. Le lissage spatial des résidus est effectué à la maille code postal et reporté sur la figure 7.9 ci-dessous. Les nuances de couleur sont ajustées par quantile à 20%.

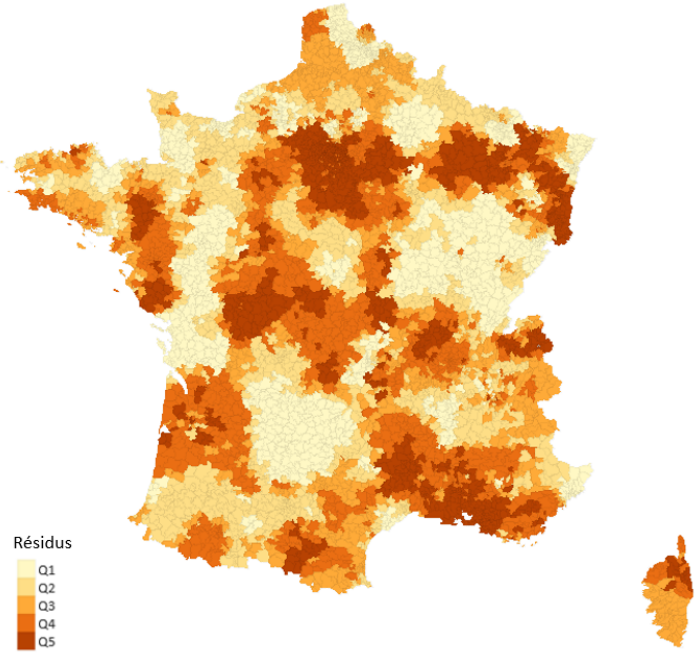

FIGURE 7.9 – Résidus coût lissés à la maille code postal

Au même titre qu'à la maille départementale, les zones apparentes sont principalement les zones urbaines, notamment Paris, Lyon, Marseille, Bordeaux, ainsi que leurs villes voisines. Hormis au Centre-Val de Loire, on retrouve une certaine similitude avec la carte représentant le niveau de vie médian par commune [2], ce qui témoigne d'une corrélation entre le coût et le niveau de vie d'une zone.

7.3 Modélisation de l'effet spatial par Machine Learning

Pour rappel, le résidu peut se théoriser comme suit :

$$Résidu = Effet\ spatial + \varepsilon$$

Où ε est la partie du résidu qui ne contient pas d'effet spatial.

Pour modéliser l'effet spatial, on cherche à expliquer les résidus par des variables géographiques, externes ou non.

Ces variables sont à différentes mailles. Les variables géographiques retenues sont reportées dans le tableau 7.5 ci-dessous.

2. Nicolas Certes, 2017, « Quel est le niveau de vie dans votre commune ? », Le Figaro

Variable	Type de variable	Maille
Nombre de médecins généralistes	Médicale	Ville
Nombre d'infirmiers	Médicale	Ville
Nombre de radiologues	Médicale	Ville
Nombre de chirurgiens-dentistes	Médicale	Ville
Nombre de kinésithérapeutes	Médicale	Ville
Nombre de sage-femmes	Médicale	Ville
Nombre de psychologues	Médicale	Ville
Nombre de naissances en 2019	Sociale	Ville
Nombre de décès en 2019	Sociale	Ville
Nombre de logements	Sociale	Ville
Nombre de chômeurs	Sociale	Ville
Nombre de travailleurs	Sociale	Ville
Niveau de vie (€)	Sociale	Département
Taux de pauvreté	Sociale	Département
Part des ménages fiscaux imposés	Sociale	Département
Superficie (km^2)	Géographique	Ville
Population totale	Géographique	Ville
Présence d'un site	Interne	Département

TABLE 7.5 – Liste des variables géographiques

L'intérêt de prendre des variables qui n'ont pas forcément de rapport direct avec l'acte étudié repose sur la possibilité d'un éventuel rapport indirect. A titre d'exemple, s'il y a plus de médecins généralistes dans une ville, le respect du parcours de soins sera plus simple pour consulter un spécialiste par la suite. Cela peut s'étendre à d'autres causes directes ou indirectes impliquant une consultation de médecin spécialiste.

Ainsi, ce sont ces variables qui permettront de modéliser l'effet spatial par apprentissage statistique.

Les modèles de Machine Learning testés seront les forêts aléatoires (ou Random Forest), à l'aide du package *randomForest* du logiciel *R*, et le gradient boosting machine (ou GBM) à l'aide du package *gbm*.

Par ailleurs, **la modélisation de l'effet spatial par Machine Learning ne concernera que la maille code postal**. En effet, la base de résidus à la maille départementale contient autant d'observations qu'il existe de département dans l'hexagone, soit moins d'une centaine. D'un point de vue métier, ce nombre de lignes peut être considéré insuffisant pour entraîner un modèle, à la différence de la base à maille code postal, qui en contient 6 000. De plus, la loi des grands nombres veut que, si une zone contient un grand nombre d'individus, son résidu s'approche de son effet spatial. Les départements étant (logiquement) plus exposés que les villes, on peut considérer que leurs résidus correspondent à leur effet spatial.

Les algorithmes Random Forest et GBM dépendent tous deux de paramètres décisifs à leur performance, choisi a priori.

Le Random Forest nécessite un choix judicieux du nombre d'arbres *ntree*, et du nombre de variables candidates *mtry*. Si un nombre d'arbres trop élevé ne détériorera pas le modèle [3], il peut le rendre excessivement lourd en temps de calculs. De plus, le nombre de variables candidates à chaque division influe fortement quant à la qualité du modèle.

3. Leo Breiman et Adele Cutler, *Random Forests*, « Random forests does not overfit. You can run as many trees as you want. »

Le GBM peut quant à lui faire l'objet d'un surapprentissage en cas de nombre trop élevé d'arbres.

Pour effectuer ces choix, on peut avoir recours à une validation croisée, dont la théorie est développée en Annexe E.4. Cette méthode vise à diviser la base de données en k blocs dont $k-1$ serviront itérativement de base d'apprentissage pendant que le dernier servira de base de test sur laquelle on récupère l'erreur de modélisation. L'erreur globale sera alors la moyenne des k erreurs récupérées. La figure 7.10 schématise le principe de la validation croisée.

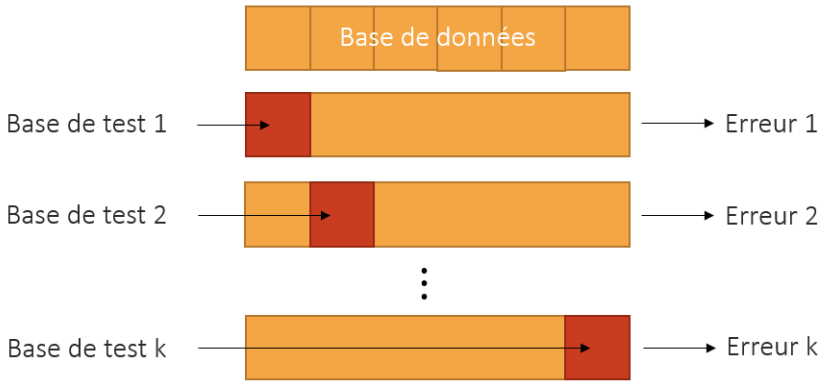

FIGURE 7.10 – Algorithme de la validation croisée

7.3.1 Modélisation de l'effet spatial fréquence

Une validation croisée a été effectuée pour les modèles Random Forest et GBM. Un nombre de blocs $k = 5$ peut être suffisant pour une estimation stable de l'erreur [4].

Pour le RandomForest, il faut dans un premier temps choisir un nombre d'arbres qui stabilise le modèle. Ainsi, en fixant $mtry = 1$, on observe la MSE évoluer en fonction du nombre d'arbres (voir figure 7.11).

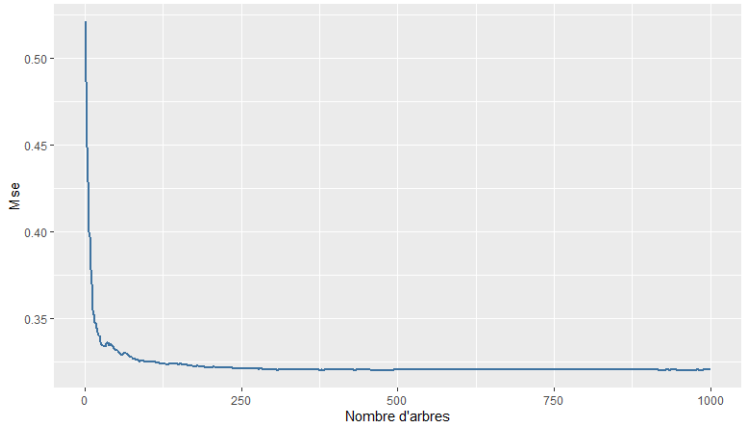

FIGURE 7.11 – Random Forest : MSE par nombre d'arbres

La MSE semble se stabiliser à partir de 200 arbres. Pour garder une certaine marge, nous

4. Jason Brownlee, *How to Configure k-Fold Cross-Validation*

fixerons 500 arbres.

Grâce à la validation croisée, le choix du nombre de variables candidates a été effectué. De la même manière, on observe la MSE évoluer en fonction de la valeur de *mtry* (voir figure 7.12).

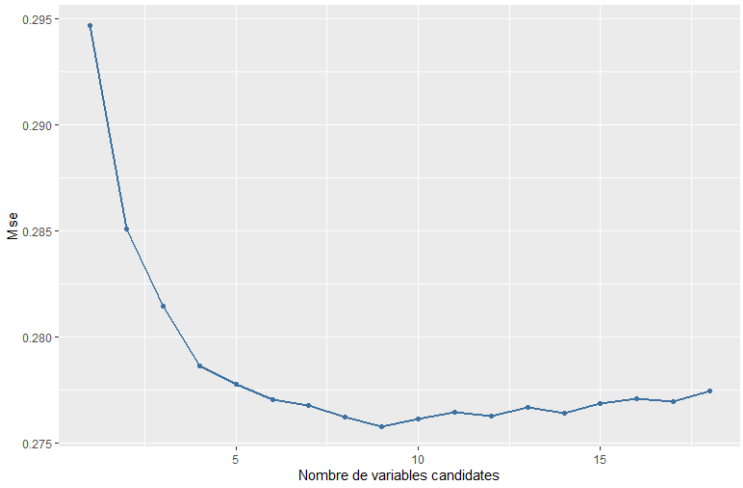

FIGURE 7.12 – Random Forest : MSE par nombre de variables candidates

Le nombre de variables candidates qui minimise la MSE est $mtry = 9$, c'est celui qui sera conservé en vue de modéliser l'effet spatial de la fréquence. Le tableau 7.6 ci-dessous reporte les indicateurs de performance de ce modèle.

Indicateur	Base d'apprentissage	Base de test
RMSE	0.4836	0.5446
MSE	0.2339	0.2966
MAE	0.3746	0.4234

TABLE 7.6 – Performance du modèle Random Forest pour la prédiction de l'effet spatial fréquence

Pour le GBM, le choix du nombre d'arbres est important pour éviter le surapprentissage. Une validation croisée à 5 blocs a été effectuée afin de choisir le nombre d'arbres qui minimise l'erreur quadratique moyenne.

La figure 7.13 montre l'évolution de la MSE, calculée sur la base de test et sur la base d'apprentissage, en fonction du nombre d'arbres.

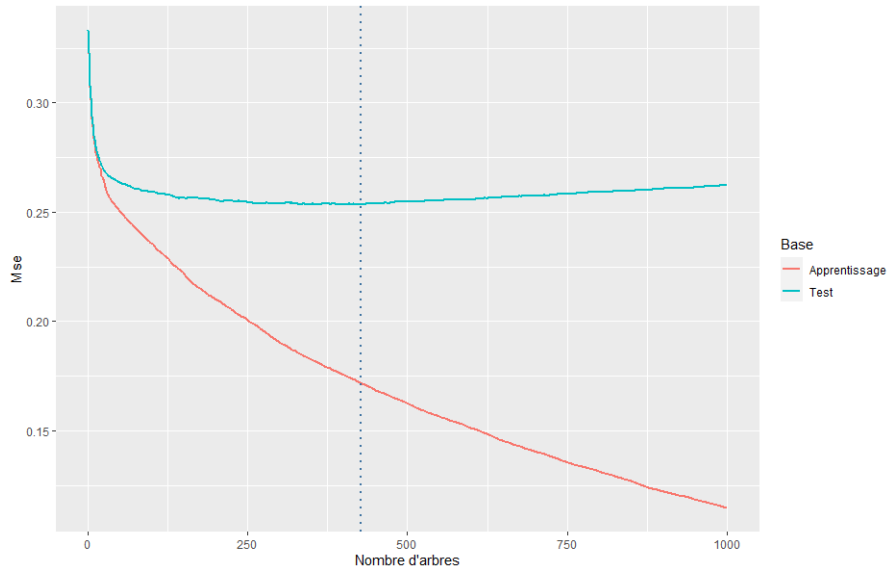

FIGURE 7.13 – GBM : MSE par nombre d'arbres

Naturellement, l'erreur de prédiction sur la base d'apprentissage est décroissante en fonction du nombre d'arbres puisque le modèle prédit à partir de données qu'il connaît, et dont il tente pour chaque arbre de minimiser l'erreur. En revanche, la MSE est minimale sur la base de tests pour 427 arbres. C'est le nombre d'arbres qui sera retenu pour la modélisation de l'effet spatial fréquence.

Le tableau 7.7 reporte les indicateurs de performances du modèle retenu.

Indicateur	Base d'apprentissage	Base de test
RMSE	0.4849	0.5429
MSE	0.2351	0.2948
MAE	0.3750	0.4219

TABLE 7.7 – Performance du modèle GBM pour la prédiction de l'effet spatial fréquence

Les deux modèles donnent des résultats sensiblement proches. Le GBM donne des résultats légèrement meilleurs sur la base de test.

Le niveau de contribution de chaque variable pour nos deux modèles est donné en Annexe G.1.

Les figures 7.14 et 7.15 représentent cartographiquement l'effet spatial capté pour chaque modèle. Les nuances de couleurs sont ajustées par quantile à 20%.

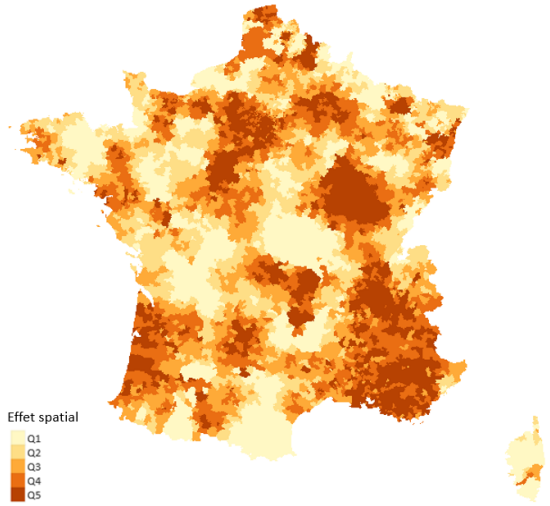

FIGURE 7.14 – Effet spatial fréquence modélisé par Random Forest

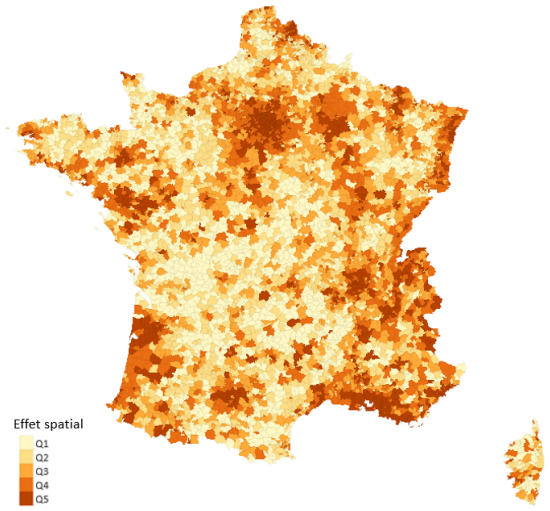

FIGURE 7.15 – Effet spatial fréquence modélisé par GBM

Il semble que le Random Forest tienne mieux compte du lissage, alors que le GBM est davantage influencé par les variables communales. Il y a moins de déserts résiduels côté GBM, et ces derniers sont très peu étendus. Le système de bagging du Random Forest tient compte des données ayant résisté au lissage, le GBM, grâce à sa méthode de boosting, ajuste ses valeurs de sorte que l'erreur résultant d'une mauvaise prédiction ne se reproduise pas.

Les zones les plus apparentes sont les grandes agglomérations (typiquement, l'Île-de-France, la Provence-Alpes-Côte-d'Azur, le Rhône-Alpes) mais également les zones disposant d'un site, ainsi que leurs zones voisines. Une très large étendue de rouge est observable dans la côte Sud-Est. Les Bouches-du-Rhône et le Var étant des départements à forte consommation moyenne par assuré, avec une importante exposition au sein de ceux-ci, cette sur-consommation a été lissée vers un grand nombre de voisins relativement éloignés étant donné la prise en compte de l'exposition dans la formule de lissage 5.5.

7.3.2 Modélisation de l'effet spatial coût

Par des méthodes analogues à celles développées dans la sous-partie précédente, on parvient à retrouver les paramètres optimaux pour la modélisation de l'effet spatial du coût.

Pour choisir le nombre d'arbres du Random Forest, on gèle dans un premier temps le nombre de variables candidates à 1, et modélise l'effet spatial avec 1000 arbres. La figure 7.16 montre l'évolution de la MSE par nombre d'arbres.

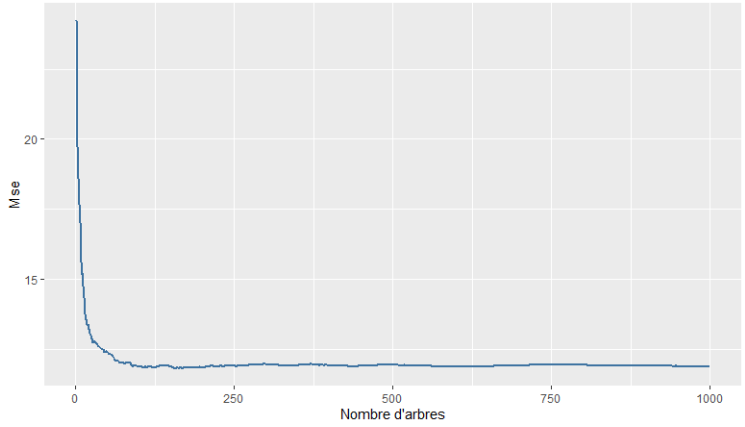

FIGURE 7.16 – Random Forest MSE par nombre d'arbres

La MSE semble se stabiliser à partir de 250 arbres. Pour conserver une certaine marge, nous prendrons 500 arbres.

Par la suite, il convient de choisir le nombre de variables candidates. Pour ce faire, nous aurons recours à une validation croisée à 5 blocs. La figure 7.17 montre l'évolution de la MSE en fonction du nombre de variables candidates.

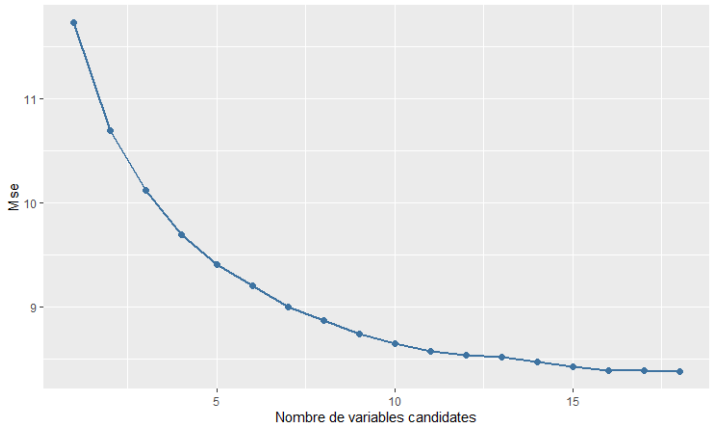

FIGURE 7.17 – Random Forest : MSE par nombre de variables candidates

Il semble que le nombre de variables candidate optimal soit le nombre maximal, soit $mtry = 18$.

Ainsi, pour le Random Forest, les paramètres choisis seront $ntree = 500$ et $mtry = 18$. Le tableau 7.8 reporte les indicateurs de performance de ce modèle.

Indicateur	Base d'apprentissage	Base de test
RMSE	1.175	2.889
MSE	1.381	8.351
MAE	0.856	2.101

TABLE 7.8 – Performance du modèle Random Forest pour la prédiction de l'effet spatial coût

Pour le GBM, le choix du nombre d'arbres doit passer par une validation croisée à 5 blocs,

de sorte à éviter le surapprentissage. La figure 7.18 montre l'évolution de la MSE par nombre d'arbres pour la base d'apprentissage et la base de test.

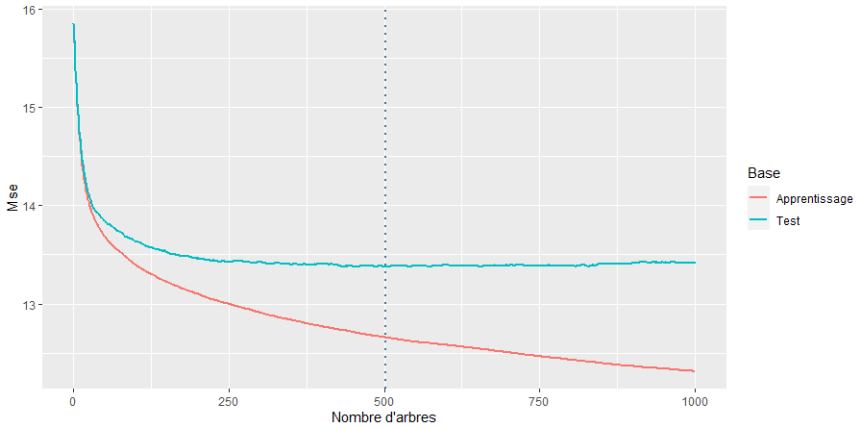

FIGURE 7.18 – GBM : MSE par nombre d'arbres

La MSE est minimale pour la base de test avec un nombre d'arbres $n.trees = 502$. C'est ce nombre d'arbres qui sera retenu pour notre GBM. Le tableau 7.9 reporte les indicateurs de performance de ce modèle.

Indicateur	Base d'apprentissage	Base de test
RMSE	2.953	3.216
MSE	8.716	10.344
MAE	2.265	2.418

TABLE 7.9 – Performance du modèle GBM pour la prédiction de l'effet spatial coût

Vraisemblablement, le Random Forest prédit bien mieux les résidus que le GBM, mais cette performance est relative. La qualité d'un modèle repose ici sur son habilité à capter l'effet spatial contenu dans les résidus.

Le niveau de contribution des variables pour nos deux modèles est donné en Annexe G.1.

Les figures 7.19 et 7.20 représentent cartographiquement l'effet spatial capté pour ces deux modèles. Les nuances de couleurs sont ajustées par quantile à 20%.

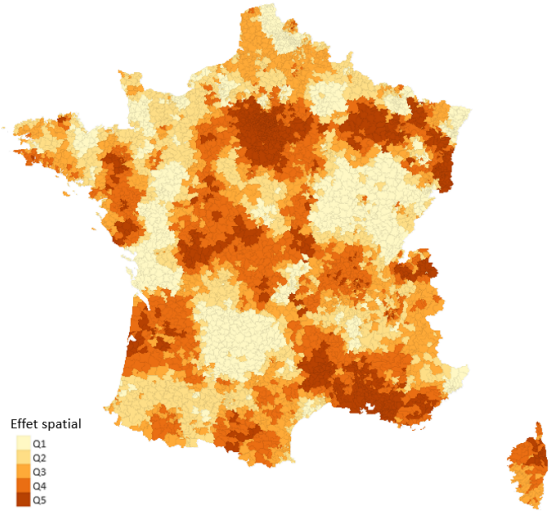

FIGURE 7.19 – Effet spatial coût modélisé par Random Forest

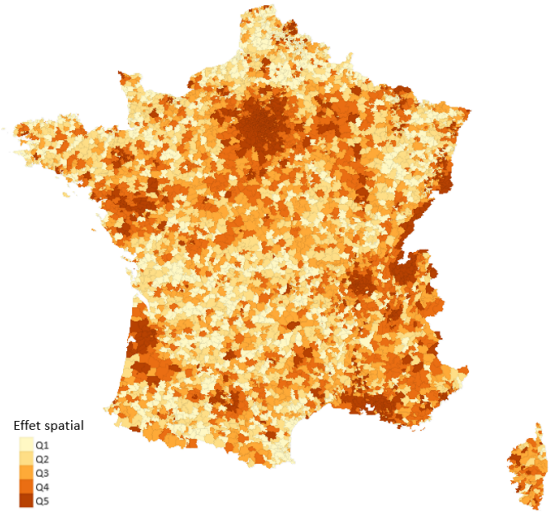

FIGURE 7.20 – Effet spatial coût modélisé par GBM

La modélisation de l'effet spatial est très différente d'un modèle à l'autre. Côté Random Forest, le résidu a été prédit de façon très protocolaire, si bien que la différence entre l'effet spatial modélisé, et les résidus bruts (cf : Figure 7.9) est très légère. Le GBM établit cependant mieux la relation vraisemblable entre les variables géographiques et le risque spatial : certains résidus, même après lissage, ne représentent pas la réalité du risque spatial. Le modèle GBM a pu atténuer les résidus élevés dûs aux cas isolés. Cependant, cela ne signifie pas que l'effet spatial du GBM représente mieux le risque spatial de ce portefeuille, mais simplement qu'il n'est pas influencé par l'exposition.

L'effet spatial est à présent modélisé. Il convient par la suite de le classifier, de façon à réduire le nombre de modalités et optimiser le modèle GLM enrichi de cette information classifiée.

7.4 Classification de l'effet spatial

La classification de l'effet spatial a été effectuée à l'aide de la classification ascendante hiérarchique appliquée à la méthode de Ward. Cette méthode vise à classifier les résidus de sorte que la variance intra-classe soit la plus faible possible, et la variance inter-classe la plus élevée possible. En d'autres termes, il faut que les classes contiennent des éléments semblables (ici, la valeur des effets spatiaux), et que ces classes soient distinctes unes-à-unes.

Chaque zone se verra ensuite attribuer sa classe respective, et par extension, chaque assuré se verra enrichi d'un nouvel attribut : la classe de sa zone de résidence, qui s'appellera *Zonier*.

Le nombre k de classes doit être choisi de sorte que la variable créée soit la plus optimale possible pour notre modèle complet. Choisir le k qui maximise la variance inter-classe reviendrait à choisir $k = +\infty$, tandis que choisir celui qui minimise la variance intra-classe reviendrait à prendre $k = 1$.

Ainsi, le k préféré sera le seuil à partir duquel la variance intra-classe converge.

7.4.1 Classification de l'effet spatial fréquence

La classification a été effectuée pour la maille départementale, pour laquelle on parle de résidus (ou résidus « bruts »), et la maille code postal, pour laquelle on parle d'effet spatial (ou résidus « net »).

Maille départementale

Dans un premier temps, observons l'évolution de la variance intra-classe des résidus de la maille départementale, en faisant varier le nombre de classes entre 1 et 15 (voir figure 7.21).

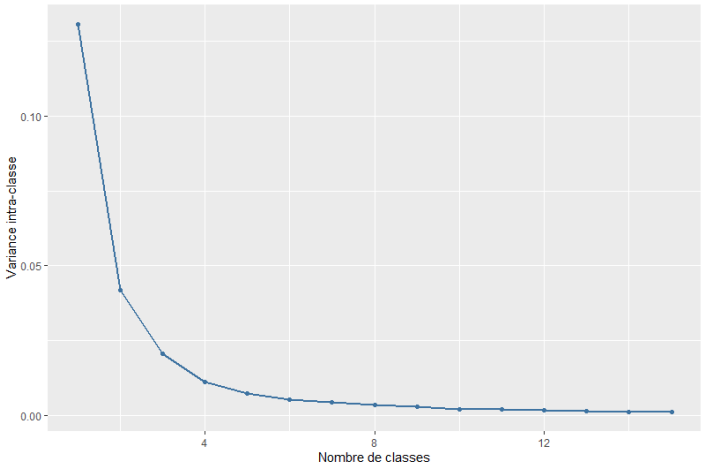

FIGURE 7.21 – Variance intra-classe des résidus pour la maille départementale

La variance intra-classe semble se stabiliser à partir de 10 classes.

En réattribuant itérativement la variable *Zonier*, créée après chaque choix de classe, il est possible de regarder évoluer les performances de notre modèle enrichi de cette variable, le modèle complet. Voir tableau 7.4.1.

Nombre de classes	AIC
2	132 013
3	131 949
4	131 925
5	131 902
6	131 886
7	131 887
8	131 887
9	131 885
10	131 875
11	131 877
12	131 875
13	131 870
14	131 874
15	131 875

TABLE 7.10 – Évolution de l'AIC du modèle complet en fonction du nombre de classe k, maille départementale

L'AIC cesse de décroître à partir de 10 classes. **C'est le nombre de classes $k = 10$ qui sera choisi pour le zonier à la maille départementale**.

La figure 7.22 représente cartographiquement les classes de zones, tout en considérant que la zone 1 est la plus « sûre » du point de vue assurantiel, et la zone 10 la plus « risquée ».

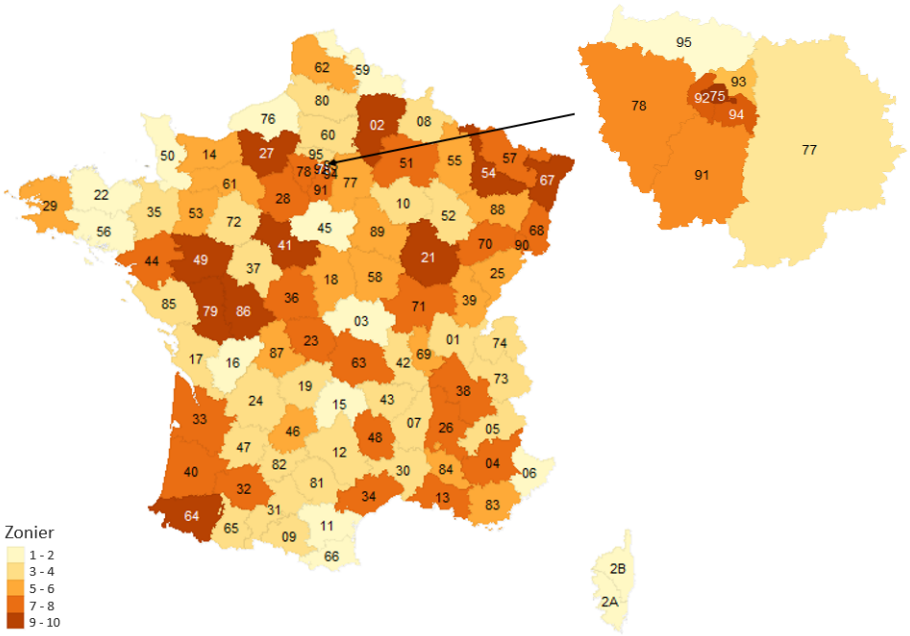

FIGURE 7.22 – Zonier fréquence à la maille départementale

A travers cette carte 7.22, on peut conclure que le zonier à la maille départementale augmentera la fréquence prédite des assurés résidents à Paris (75), dans la Côte-d'Or, dans les Bouches-du-Rhône (13), et plus globalement dans toutes les zones oranges/rouges. Par ailleurs, ce zonier diminuera la fréquence prédite des assurés résidents dans les zones claires, comme l'Allier (03) ou la Charente (16).

Pour se faire une idée plus détaillée, effectuons les mêmes études à la maille code postal.

Maille code postal

La figure 7.23 présente l'évolution de la variance moyenne intra-classe de l'effet spatial maille code postal en fonction du nombre de classes k, en faisant varier ce dernier entre 1 et 20.

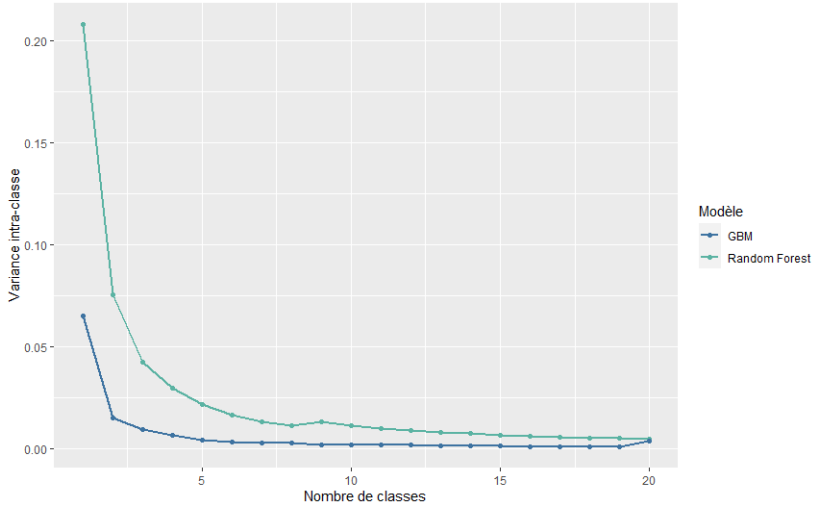

FIGURE 7.23 – Variance intra-classe de l'effet spatial pour chaque modèle, maille code postal

Pour les deux modèles, la variance intra-classe semble être stable à partir de 8 classes. Il est possible de choisir des classes supérieures, mais un nombre trop important de classes pourrait complexifier, voire détériorer nos modèles.

En réattribuant itérativement la variable *Zonier*, créée après chaque choix de classe, il est possible de regarder évoluer les performances de notre modèle enrichi de cette variable.

Nombre de classes	Random Forest	GBM
2	131 942	132 122
3	131 305	132 118
4	131 304	132 122
5	131 266	132 106
6	131 010	132 108
7	130 989	132 108
8	130 952	132 107
9	130 948	132 100
10	130 865	132 101
11	130 867	132 104
12	130 870	132 100
13	130 854	132 109
14	130 852	132 108
15	130 855	132 108
16	130 841	132 106
17	130 840	132 109
18	130 841	132 111
19	130 844	132 114
20	130 844	132 114

TABLE 7.11 – Évolution de l'AIC du modèle complet en fonction du nombre de classes k, maille code postal

Au même titre que la variance intra-classe, l'AIC semble se stabiliser à partir de $k = 10$ classes. En effet, la variance intra-classe moyenne étant sensiblement proche pour les k supérieurs, les individus au sein de celles-ci ne se distinguent guère beaucoup mieux avec 11 classes qu'avec 10, d'où la faible évolution de l'AIC au delà de ce seuil. **C'est le nombre de classes $k = 10$ qui sera choisi pour le zonier fréquence à la maille code postal**.

Les figures 7.24 et 7.25 représentent cartographiquement les classes de zones pour les deux modèles, en considérant que la zone 1 est la plus « sûre » d'un point de vue assurantiel, et la zone 10 la plus « risquée ».

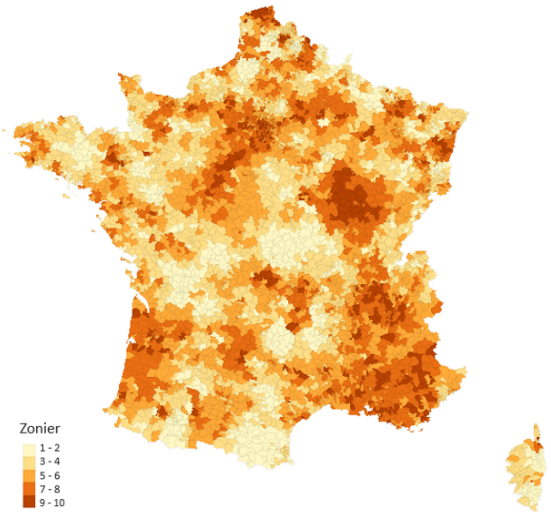

FIGURE 7.24 – Zonier fréquence maille code postal par Random Forest

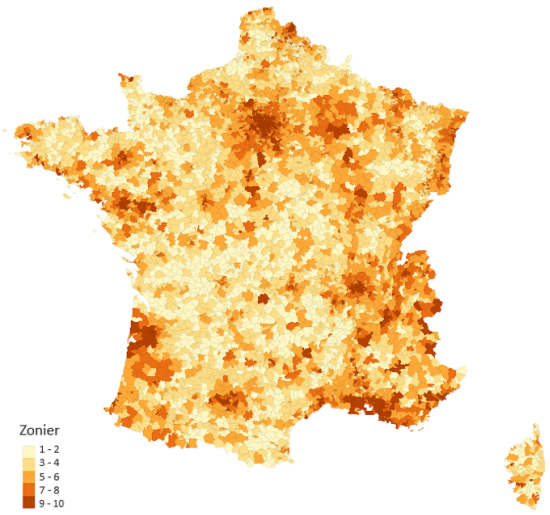

FIGURE 7.25 – Zonier fréquence maille code postal par GBM

Dans l'ensemble, les mêmes départements sont « pénalisés » qu'à la maille départementale, mais certaines villes au sein de ceux-ci le seront moins que d'autres. Il en va de même pour les zones « dépénalisées ». Le chapitre suivant confrontera les zoniers à ces deux mailles. De plus, en fonction du modèle, certaines villes seront plus ou moins pénalisées. A titre d'exemple, le Nord n'est pas considéré comme une zone à risque par le modèle GBM, alors qu'il l'est par le Random Forest.

7.4.2 Classification de l'effet spatial coût

Le choix du nombre de classes k dans lesquelles on souhaite regrouper l'effet spatial coût se fait par une méthode analogue à celle de l'effet spatial fréquence, développée précédemment.

Maille départementale

Le développement et la justification du choix du nombre de classes k, dans lesquelles seront regroupés les départements, se trouve en Annexe G.2.

Le nombre de classes à partir duquel la variance intra-classe converge, tout en optimisant l'AIC du modèle complet, est $k = 9$.

La figure 7.26 présente cartographiquement le zonier du coût à la maille départementale.

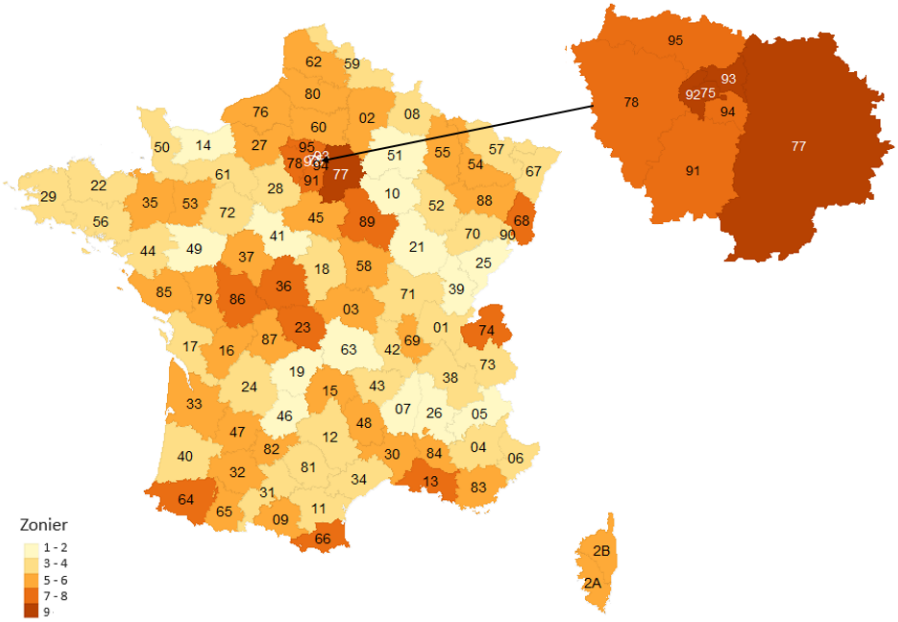

FIGURE 7.26 – Zonier coût à la maille départementale

Seuls quatre départements se trouvent dans la classe 9 qui est la plus risquée du point de vue de l'assureur, et ces quatre départements se trouvent tous en Ile-de-France. Par ailleurs, les départements de cette région se situent tous entre la classe 7 et 9. L'Ile-de-France semble être une zone à risque de sous-tarification dans le cadre de notre portefeuille.

Maille code postal

Le choix du nombre de classes k du zonier coût à la maille code postal a été effectué de manière analogue au zonier fréquence. Les détails se trouvent en Annexe G.2.

Ainsi, pour la classification de l'effet spatial coût, le nombre de classes choisi sera $k = 11$.

Les figures 7.27 et 7.28 représentent cartographiquement les zoniers construits respectivement via le modèle Random Forest, et le modèle GBM.

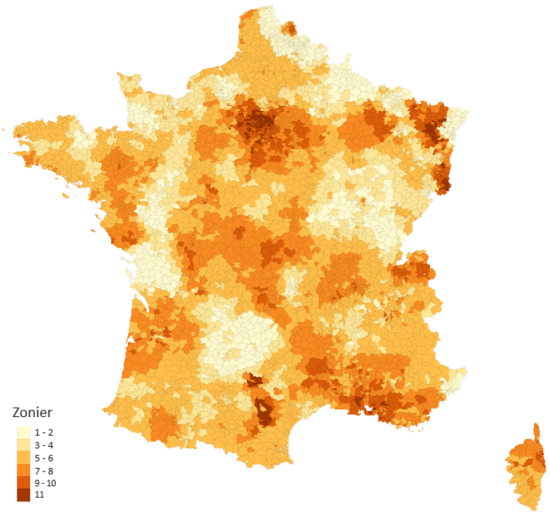

FIGURE 7.27 – Zonier coût maille code postal par Random Forest

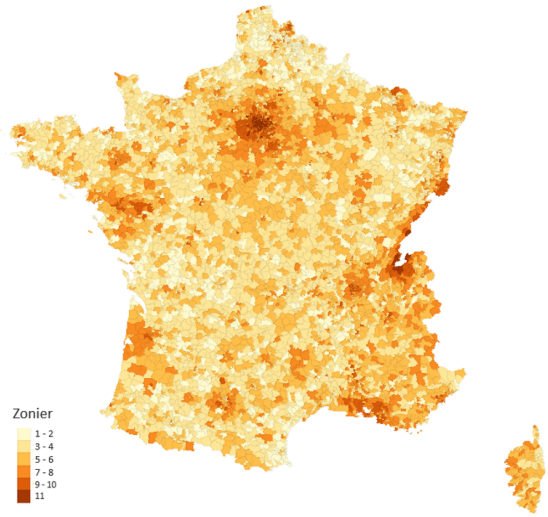

FIGURE 7.28 – Zonier coût maille code postal par GBM

A l'instar du zonier fréquence, les zones les plus apparentes sont les grandes agglomérations telles Paris, Marseille, Lyon, Bordeaux, Nantes.

Au vu d'une importante concentration de zones oranges/rouges côté Random Forest, on peut imaginer que les coûts modélisés avec ce zonier augmenteront conséquemment, mais ces coûts seront relativement homogène dans l'espace. Côté GBM, il est clair que les grandes agglomérations se verront augmenter leur coût moyen modélisé avec ce zonier.

7.5 Blancheur de l'effet non-expliqué

L'effet spatial ayant été extrait du résidu brut, on a la formule suivante :

$$\textit{Résidu} = \textit{Effet spatial} + \varepsilon$$

Ce ε est un bruit que nous allons étudier dans cette section, conformément au schéma 5.4.

Pour s'assurer du pouvoir explicatif de l'effet spatial, il est pertinent d'observer le « sous-résidu » qu'il génère. En effet, la qualité d'un modèle se distingue lorsque l'effet non-expliqué est non-explicable, et que l'erreur de modélisation qui en découle représente une part d'aléa. C'est ce qu'on appelle un **Bruit blanc**.

Mathématiquement, un processus $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ est un bruit blanc si :

— $\mathbb{E}[\varepsilon] = 0$

— $Var(\varepsilon_k) = \sigma^2 \ \forall k$, où $\sigma^2 \in \mathbb{R}$

— $Cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$

Dans la pratique, l'aspect aléatoire d'un résidu se mesure à travers un test de blancheur Portmanteau, et plus précisément le test de Ljung-Box, qui teste l'autocorrélation d'un échantillon avec pour hypothèse nulle H_0 : « Il n'y a pas d'autocorrélation des erreurs d'ordre supérieur à 1 », et sa gaussienneté peut s'apprécier subjectivement à travers un diagramme quantile-quantile (ou *QQ-plot*).

Ainsi, étudier l'effet spatial modélisé revient à observer sa qualité à absorber suffisamment d'information pour blanchir l'erreur de modélisation. L'effet spatial n'ayant pas été modélisé

pour la maille départementale, on observera les résidus découlant des modèles Random Forest et GBM.

7.5.1 Bruit fréquence

L'erreur de modélisation de l'effet spatial fréquence par Random Forest (resp. GBM), qu'on appellera ε_{RF} (resp. ε_{GBM}), est représenté graphiquement sur la figure 7.29 (resp. 7.30).

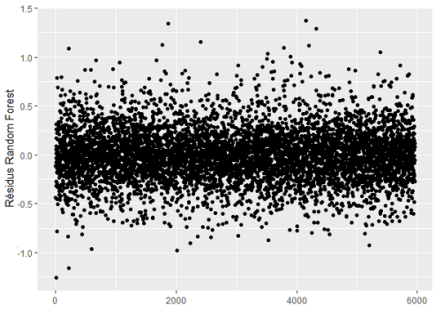

FIGURE 7.29 – Erreur de modélisation fréquence du Random Forest

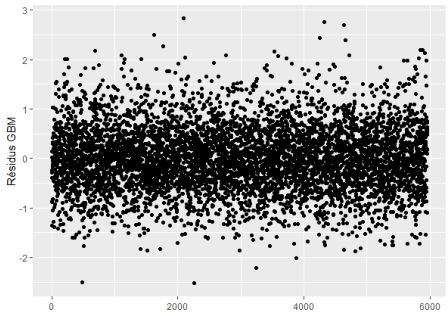

FIGURE 7.30 – Erreur de modélisation fréquence du GBM

Visiblement, les résidus semblent aléatoires et centrés.
Les tests de Ljung-Box écartent l'hypothèse d'autocorrélation des résidus (voir tableau 7.13)

	Random Forest	GBM
p-value	0.64	0.27

TABLE 7.12 – p-value du test de blancheur Ljung-Box

En observant le diagramme *QQ-plot* on peut subjectivement valider ou rejeter la gaussienneté, des résidus. Voir figures 7.35 et 7.36.

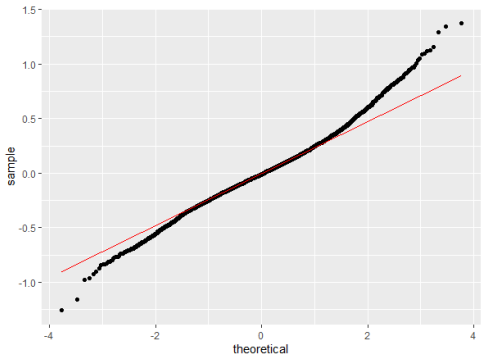

FIGURE 7.31 – Diagramme QQ-plot de l'erreur du Random Forest

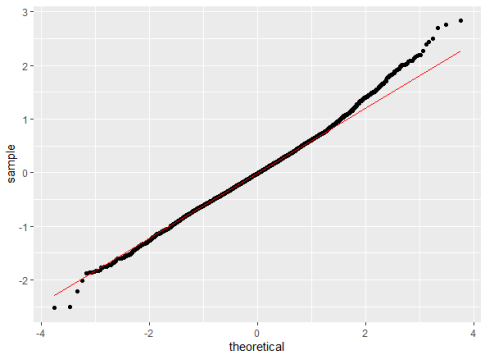

FIGURE 7.32 – Diagramme QQ-plot de l'erreur du GBM

La qualité d'ajustement se mesure à travers l'ajustement des points noirs à la droite de régression rouge. Visuellement, il semblerait que le GBM blanchit mieux les résidus fréquence, ce qui s'explique par le fait qu'il apprend par boosting, non pas par bagging : la différence est conséquente car le GBM cherche à minimiser le biais, là où le bagging cherche à minimiser la variance, en conséquence, l'erreur de modélisation du boosting sera moins biaisée, et par extension moins

explicable. En tout état de cause, les deux résidus s'ajustent relativement correctement à une loi gaussienne.

Ainsi, on peut considérer que l'erreur résiduelle ε entre l'effet spatial et les résidus « bruts » s'approche d'un bruit blanc.

7.5.2 Bruit coût

L'erreur de modélisation de l'effet spatial coût modélisé par Random Forest (resp. GBM) est représentée graphiquement sur la figure 7.33 (resp. 7.34).

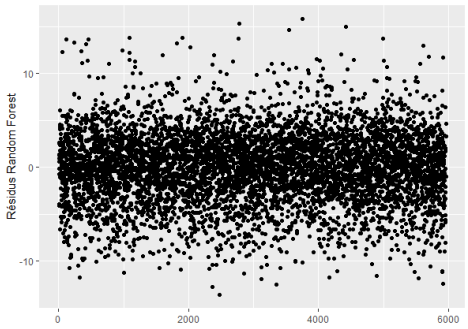

FIGURE 7.33 – Erreur de modélisation coût du Random Forest

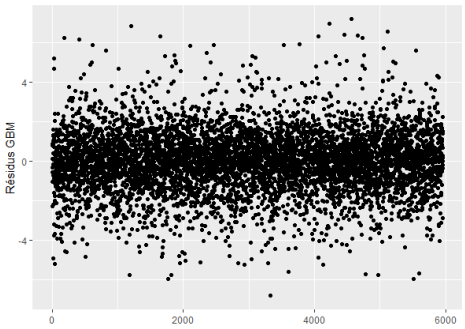

FIGURE 7.34 – Erreur de modélisation coût du GBM

Les tests de Ljung-Box écartent l'hypothèse d'autocorrélation des résidus (voir tableau 7.13)

	Random Forest	GBM
p-value	0.82	0.21

TABLE 7.13 – p-value du test de blancheur Ljung-Box

En observant le diagramme *QQ-plot* on peut subjectivement valider ou rejeter la gaussienneté, des résidus. Voir figures 7.35 et 7.36.

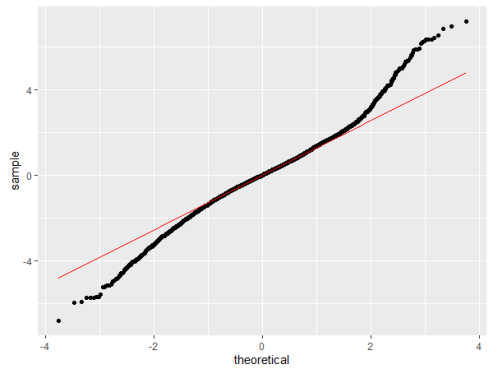

FIGURE 7.35 – Diagramme QQ-plot de l'erreur du Random Forest

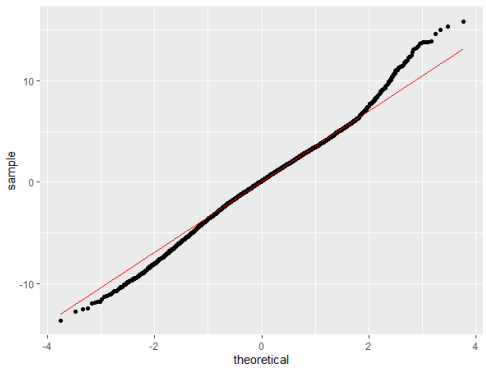

FIGURE 7.36 – Diagramme QQ-plot de l'erreur du GBM

Comme pour la fréquence, il semblerait que le GBM blanchit mieux les résidus, puisque son algorithme de boosting vise à réduire l'erreur résiduelle, mais dans l'ensemble les deux modèles

donnent des résultats satisfaisant.

Conclusion

A l'issu de ce chapitre, une classification des zones, appelée *Zonier* a été construite pour l'acte *Consultations spécialistes*. Les résultats sont globalement en faveur du zonier à la maille ville construit par les forêts aléatoires. Ce zonier, plus complexe, est sujet à incertitudes dues à la faible exposition au sein de certaines villes, incertitude minimisée par la maille départementale qui dispose de plus d'informations intra-zone, mais moins d'informations inter-zone.

Chapitre 8

Apports, limites et perspectives du zonier

Ce chapitre a pour objectif de conforter le besoin d'une variable discriminante à la zone dans un modèle de tarification santé collective, mais aussi de prendre du recul sur les études effectuées dans ce mémoire.

Le portefeuille sur lequel ont été effectuées les études et modèles a été enrichi d'une variable, la variable *Zonier*, respectivement aux actes, mailles, et sinistralités (fréquence ou coût) pour lesquels ont été construits ces zoniers. Ces variables ont pour but de donner une indication de la zone de résidence de chaque assuré concernant son niveau et type de risque. La figure 8.1 illustre le spectre du risque spatial.

FIGURE 8.1 – Spectre du risque spatial classifié par le zonier

Dans un zonier à 10 classes, les dernières représentent un risque de sous-tarification dans la mesure où leur prime pure, et par extension leur risque, ont été sous-estimés. Tandis que les premières classes représentent un effet spatial négatif, donc une surestimation de la prime pure qui, si elle s'étend spatialement, pourrait, dans le cadre d'un contrat de complémentaire santé collective, inciter l'entreprise à se tourner vers un assureur concurrent. C'est donc un risque d'antisélection.
Le zonier a donc pour but d'ajuster ces deux risques dans la tarification.

Dans ce chapitre, l'apport de la variable *Zonier* créée pour l'acte *Consultations spécialistes* sera étudié.

La performance de notre modèle de prédiction initial, appelé « modèle contraint », sera confrontée à celle du même modèle enrichi de la variable zonier, appelé « modèle complet ».
Pour rappel, le zonier a été construit à la maille départementale et à la maille code postal. Les résultats issus de ces deux mailles seront étudiés et confrontés.

8.1 Comparaison des modèles

Les variables *Zonier fréquence* et *Zonier coût* ont été ajoutées au portefeuille. Pour vérifier leur pertinence, il est possible d'observer de part et d'autre les résultats obtenus via le modèle contraint, et ceux obtenus via le modèle complet, à savoir le modèle intégrant la variable construite.

8.1.1 Comparaison des modèles GLM fréquence

Le tableau 8.1 présente les indicateurs de performance de chaque modèle ajusté de manière optimale, à savoir le modèle contraint intégrant les variables *Classe d'âge* et *Sexe*, le modèle complet dont le zonier a été construit à la maille départementale, le modèle complet dont le zonier a été construit à la maille ville par Random Forest, et le modèle complet dont le zonier a été construit par GBM.

Modèle	AIC	MSE	MAE
Modèle contraint	132 244	8.7694	1.7282
Modèle complet maille départementale	131 875	8.7118	1.7184
Modèle complet maille ville Random Forest	130 865	8.5611	1.7001
Modèle complet maille ville GBM	132 101	8.7552	1.7242

TABLE 8.1 – Indicateurs de performance pour chaque modèle GLM fréquence

Le modèle contraint est le moins performant, puisqu'il contient moins d'informations permettant d'expliquer la fréquence.

Le zonier construit par le modèle Random Forest est, de loin, le plus performant, tant et si bien que la variable *Zonier* en est plus discriminante que la variable *Classe d'âge*, ce qui peut introduire la question de corrélation entre ces variables.

Ainsi, en testant la dépendance entre la entre la variable *Zonier* et la variable *Classe d'âge* via les tests de χ^2 et du V de Cramer, on obtient un V de Cramer de 0.02, avec une p-value de 0.001, ce qui permet de valider l'hypothèse d'indépendance entre la classe d'âge et la classe de zone de résidence.

Cette performance peut cependant être remise en cause par le fait que le zonier construit par Random Forest est consécutif à une prédiction sur une base d'apprentissage, et donc s'adapte plus au comportement du portefeuille qu'au risque spatial intrinsèque. A l'inverse, le GBM prédit un effet spatial sensiblement éloigné du comportement des assurés du portefeuille au profit des variables spatiales mais, sur le long terme, est moins sujet à volatilité que le Random Forest.

En outre, le zonier à la maille départementale améliore considérablement le modèle, et peut être plus fiable sur le long terme dans la mesure où la variable département n'a pas été simulée, contrairement au code postal. De plus, la population intra-départementale est moins sujette à évolution que la population intra-ville.

Comme effectué avec le modèle GLM contraint en partie 6.1.3, une observation des coefficients linéaires respectifs de chaque modalité donne une indication de la pertinence de la nouvelle variable. Effectuons cette observation en prenant le zonier le plus performant, celui du Random Forest. Voir tableau 8.2.

Variables	Modalité	Coefficient	p-value
Intercept	(intercept)	1.43100	< 2e-16
Classe d'âge	2	-0.34431	< 2e-16
Classe d'âge	3	-0.48815	< 2e-16
Classe d'âge	4	-0.11526	< 2e-16
Classe d'âge	5	0.03699	0.523
Classe d'âge	6	-0.15391	0.00307
Sexe	M	-0.22284	< 2e-16
Zonier	2	0.2274	0.00769
Zonier	3	0.44329	8.59e-07
Zonier	4	0.59515	8.70e-14
Zonier	5	0.69728	< 2e-16
Zonier	6	0.85423	< 2e-16
Zonier	7	0.84349	< 2e-16
Zonier	8	0.93365	< 2e-16
Zonier	9	1.08602	< 2e-16
Zonier	10	1.33359	< 2e-16

TABLE 8.2 – Coefficients GLM fréquence : modèle complet

L'intercepte contient la modalité 1 pour la variable *Classe d'âge*, F pour la variable *Sexe*, et 1 pour la variable *Zonier*.

Le coefficient du zonier augmente avec le numéro de classe. Cela signifie que plus la classe est élevée, plus la fréquence l'est, ce qui est logique puisque le zonier a été construit comme tel. Par ailleurs, chaque modalité du zonier est significative au vu de la faible p-value.

8.1.2 Comparaison des modèles GLM coût

Le tableau 8.3 présente les indicateurs de performance de chaque modèle ajusté de manière optimale.

Modèle	AIC	MSE	MAE
Modèle contraint	17 608	287.5	12.69
Modèle complet maille départementale	16 507	272.8	12.26
Modèle complet maille ville Random Forest	16 152	268.6	12.14
Modèle complet maille ville GBM	17 114	279.9	12.52

TABLE 8.3 – Indicateurs de performance pour chaque modèle GLM coût

Tout comme pour la fréquence, le modèle préféré est le modèle complet avec zonier construit par Random Forest, qui est mieux adapté à notre portefeuille. A titre indicatif, la variable *Zonier* ajoutée a un pouvoir discriminant plus fort que la classe d'âge.

Par ailleurs, les résultats du modèle complet avec zonier départemental sont plus que satisfaisants.

Le constat concernant les coefficients linéaires associés aux modalités de la variable *Zonier* est sensiblement le même que pour la fréquence (cf : Partie 8.1.1), à savoir une croissance de l'influence de sur la fréquence d'une classe à l'autre, et une importante significativité de ces dernières. Ces coefficients ne seront par conséquent pas détaillés.

8.1.3 Autres aspects de comparaison

L'influence de la variable *Zonier* peut s'apprécier à travers des conséquences directes quant aux modèles de prédiction.

Fréquence intra-classe

La fréquence prédite par le modèle contraint, qu'on appellera $Fréquence_{contraint}$ n'est, en théorie, pas influencée par la zone puisque ce modèle ne dépend d'aucune information géographique.

A l'inverse, la fréquence prédite par le modèle complet, qu'on appellera $Fréquence_{complet}$, est ajustée en fonction de chaque classe de zone. En conséquence, on peut s'attendre à ce que la fréquence moyenne intra-zone découlant du modèle contraint soit la même , et la fréquence moyenne intra-zone découlant du modèle complet soit croissante, ou tout du moins variable d'une zone à l'autre.

Mathématiquement, soit $\Theta = (\Theta_1, ..., \Theta_k)$ l'ensemble des zones. On aurait alors :

$$\mathbb{E}[Fréquence_{contraint}|\Theta_i] = \mathbb{E}[Fréquence_{contraint}|\Theta_j] = \mathbb{E}[Fréquence_{contraint}]$$

Et :

$$\mathbb{E}[Fréquence_{complet}|\Theta_i] \neq \mathbb{E}[Fréquence_{complet}|\Theta_j]$$

Pour tout $i \neq j$.

La figure 8.2 confronte la fréquence intra-classe prédite via le modèle contraint face à la fréquence intra-classe prédite via le modèle complet [1], et à la fréquence intra-classe observée.

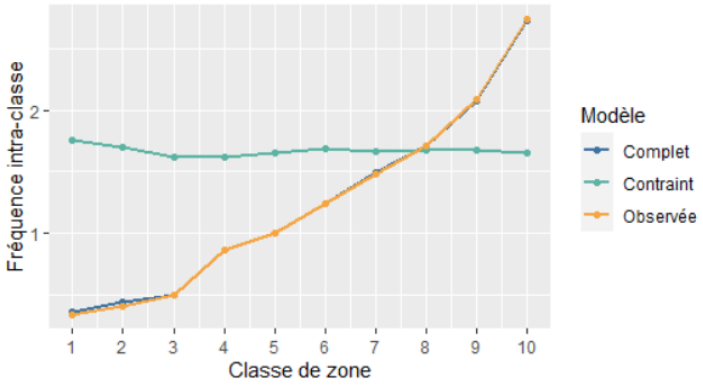

FIGURE 8.2 – Fréquence moyenne intra-classe par modèle vs fréquence moyenne intra-classe observée

On voit qu'avec le modèle contraint, la fréquence moyenne intra-classe est presque constante. Tandis que celle prédite par modèle complet est presque confondue à la fréquence intra-classe observée, et croissante d'une zone à l'autre. Cela donne une indication sur la considérabilité du zonier et de son influence. Une partie des zones étaient surestimées par le modèle contraint (de la classe 1 à la classe 7), tandis que l'autre était sous-estimée. Le zonier parvient à ajuster ces erreurs.

1. Le zonier choisi pour cette observation est celui qui a le plus d'influence sur la modélisation de la fréquence, soit le zonier construit par Random Forest.

Cela témoigne également d'une indépendance entre la zone et l'âge. En effet, la présence d'une corrélation importante entre ces deux variables pourrait faire varier la fréquence intra-classe du modèle contraint, ce qui n'est pas le cas ici.

Le constat se vérifie également pour le zonier coût.

Coût moyen intra-classe

A l'instar de la fréquence, le coût moyen intra-classe prédit par modèle contraint peut être confronté au coût moyen intra-classe prédit par modèle complet, ainsi qu'aux observations réelles. Le tableau 8.3 présente le coût moyen intra-classe pour chaque classe de zone.

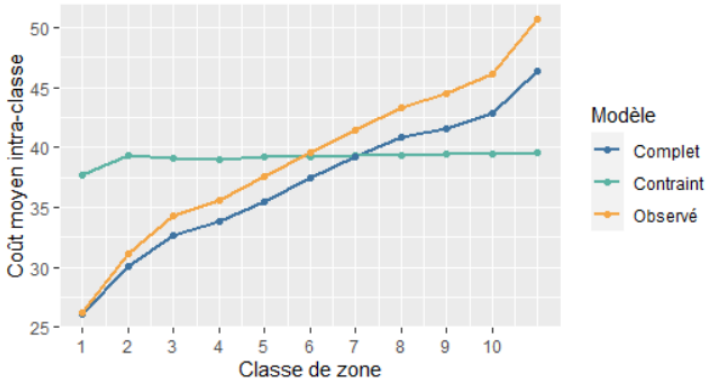

FIGURE 8.3 – Coût moyen intra-zone par modèle vs coût moyen intra-classe observé

Le coût moyen prédit ne varie que peu d'une zone à l'autre pour le modèle contraint, tandis qu'il croît pour le modèle complet, en suivant la tendance des observations, comme on pouvait s'y attendre. Cela démontre un ajustement de l'estimation du risque par zone à travers le zonier. Cela prouve également une absence de corrélation entre la zone et les autres variables explicatives : aucune zone ne contient plus ou moins de personnes âgées, et les personnes plus ou moins âgées n'ont pas de zone préférée.

8.2 Limites et perspectives de ce zonier

Nonobstant une conséquente amélioration des modèles de prédiction consécutifs à l'intégration du zonier, tant pour la sous-estimation que la surestimation de la sinistralité, quelques limites ou points sont discutables quant à son fondement :

- Le zonier est construit en se basant sur un portefeuille santé d'entreprise, et s'applique en conséquence à un portefeuille, ou en tout état de cause, à cette entreprise. De plus, les sites de l'entreprise ne couvrent pas toutes les zones de la France à parts égales. Ainsi, le zonier pourrait être inefficace s'il était appliqué à un portefeuille santé d'une autre entreprise ;
- Le modèle dépend d'hypothèses fortes qui ne sont, en pratique, pas toujours vérifiées, comme l'indépendance du coût et de la fréquence ;
- Le régime Alsace-Moselle est différent du régime général, mais, dans l'impossibilité de distinguer l'appartenance à un régime, a été traité de la même manière ;
- Le zonier à la maille ville se base sur des codes postaux simulés, bien que cette simulation respecte une certaine cohérence quant à la démographie intra-départementale ;
- Le zonier départemental est construit sur les résidus bruts ;

— Les données datent de 2019 [2], année où les prémisses de la réforme 100 % Santé ne sont pas déployés. En conséquence, les résultats du zonier sont susceptibles d'être obsolètes. Toutefois, l'approche de ce zonier a été ici appliquée sur un acte qui n'est pas concerné par le 100 % Santé, mais les possibles conséquences indirectes de cette réforme sur cet acte pourraient être observées (à titre d'exemple, l'augmentation de la consommation dans le poste optique pourrait conduire à une augmentation des consultations ophtalmologiques).
— De plus, les évènements liés à la pandémie de la COVID-19 survenue en 2020 ont bouleversé le marché de l'assurance santé, notamment en raison des reports de soins [3]. Ainsi, en adoptant cette approche sur des données plus récentes, si elle peut être efficace, les résultats en seraient sensiblement différents.

Par ailleurs, d'autres questions peuvent remettre en cause la théorie d'un zonier, comme le fait que l'adresse renseignée dans un contrat d'assurance n'est pas nécessairement l'adresse de résidence, ou que la consommation de certains soins peut être indépendant de la zone de résidence (typiquement, les cures thermales peuvent avoir lieu dans des zones très éloignées de la résidence).

Ainsi, si certaines de ces limites pourraient être corrigées avec des données enrichies (disposant de plus d'informations discriminantes, géographiques ou non), d'autres apportent des perspectives d'approfondissement. En effet, effectuer ces études sur des données post-100 % Santé et/ou post-COVID-19 permettrait de rendre compte de l'impact de l'effet spatial dans le contexte sanitaire actuel et de l'émergence de la télé-médecine. De surcroît, la crise sanitaire a conduit à une importante montée du télétravail [4], par le biais duquel les salariés d'une entreprise seraient amenés à travailler dans des zones potentiellement éloignées de leur siège social, ce qui pourrait amener les assureurs à tarifer selon la répartition géographique des assurés plutôt que le siège social.

Dans le cadre de la tarification santé collective, la méthodologie afférente à ces études pourrait en outre avoir des fins d'optimisation des outils de tarification existants, comme l'intégration d'un coefficient correcteur visant à ajuster la prime pure d'un portefeuille santé collective en fonction de la répartition de celui-ci dans les classes de zones construites.

2. Année de survenance.
3. Gwendal Perrin, 2020, « Covid-19 : vers un fort rattrapage des dépenses en santé », L'Argus de l'assurance.
4. Activité professionnelle effectuée à distance du lieu où le résultat du travail est attendu.

Conclusion

Dans le besoin d'approcher au mieux le risque santé à travers les modèles de tarification, tout en conservant une certaine mutualisation, l'influence de l'information géographique doit être évaluée de façon rigoureuse pour être intégrée aux modèles de prédiction de primes d'assurance.

Dans ce mémoire, nous avons proposé une approche de construction de zoniers à mailles différentes, dont nous avons étudié les performances et l'apport dans les modèles de prédiction. Ce zonier a été construit à partir de l'effet géographique contenu dans l'erreur de prédiction de la sinistralité. L'approche appliquée consiste, dans un premier temps, à modéliser la sinistralité par un GLM sans variable géographique. Par la suite, il convient de récupérer les résidus découlant de ce modèle et de les agréger à chaque zone. L'étape suivante vise à appliquer un lissage spatial aux résidus afin d'homogénéiser leur répartition géographique. Ces résidus lissés doivent ensuite être modélisés par Machine Learning à l'aide de variables géographiques, de manière à ne récupérer que l'effet spatial. L'approche s'achève par une classification de l'effet spatial modélisé.

Cet effet spatial a été capté et classifié à l'aide de méthodes de Data Science de manière à faire de l'information géographique une variable qualitative avec un nombre de facteurs modéré pour ne pas complexifier le modèle qui l'intégrera et répondre au besoin de mutualisation.

Cette approche a été appliquée sur un acte particulièrement représenté dans notre portefeuille, l'acte *Consultations spécialistes*, mais pourrait être capitalisé sur d'autres actes pourvu que ceux-ci disposent de suffisamment d'observations. De plus, étant donné la différence de risque spatial entre la fréquence et le coût, notamment pour cet acte, les zoniers ont été construits distinctement pour la fréquence et le coût.

Les résultats statistiques indiquent une amélioration des modèles de tarification intégrant le zonier. Le modèle le plus efficace est le zonier à la maille ville construit par Random Forest, son système de bagging ajustant mieux l'effet spatial au comportement des assurés de ce portefeuille. Le zonier maille ville construit par GBM, de par son système de boosting, résiste mieux aux valeurs importantes qui ne seraient pas en adéquation avec les données géographiques. De ce fait, ce modèle donne de moins bons résultats pour ce portefeuille, mais il est probable qu'il garde sa robustesse sur le long terme, à la différence du zonier Random Forest qui perdrait de son apport sur un autre portefeuille.

Grâce aux lissages spatiaux, les zoniers ne présentent pas de « sauts » entre zones adjacentes, c'est-à-dire de passages d'une zone neutre à une zone risquée ou inversement. Cela signifie que deux entités spatialement voisines, sous réserve qu'elles aient sensiblement les mêmes caractéristiques de risque non-géographiques, n'auraient pas de primes pures radicalement différentes. Dans le cadre d'un portefeuille santé collective, ce lissage maintient donc une cohérence dans le calcul de la prime pure collective.

Le zonier départemental quant à lui propose également des résultats favorables, et possiblement plus fiables sur le long terme dans la mesure où, s'agissant d'un portefeuille santé d'entre-

prise, la population intra-départementale est moins sujette à volatilité.

Si l'enjeu stratégique de la prise en compte de l'information géographique dans les modèles de tarification n'est pas remise en cause par ce mémoire, puisque tous les zoniers construits proposent des résultats plus ou moins optimaux, il peut ici être discuté. En effet, notre zonier se base sur un portefeuille santé d'une entreprise et il est possible que les résultats soient sensiblement différents sur le portefeuille santé d'une autre entreprise. Par ailleurs, la démarche dépend d'hypothèses fortes comme l'indépendance entre la fréquence et le coût, qui ne sont, en pratique, pas toujours vérifiées.

Malgré ces limites, l'apport méthodologique de ce mémoire pourrait être étudié en vue d'optimiser les outils de tarification santé collective existants. A titre d'exemple, ajuster une prime pure collective par le biais d'un coefficient correcteur calculé en fonction de la répartition du portefeuille (ou des sites de l'entreprise) dans les zones classifiées.

De plus, les données externes fiabilisent l'impact de l'effet spatial sur le comportement des assurés, et plus généralement permettent, pour les portefeuilles en cours de création, un apport de données non négligeable, par exemple pour un assureur qui souhaiterait lancer un nouveau produit. En outre, ce zonier permettrait d'ajuster les primes pures dans les zones déserts médicaux, pour lesquels des opérations marketing de télémédecine pourraient être mises en place.

Enfin, une perspective d'approfondissement de cette étude consisterait à appliquer les méthodes afférentes à ce mémoire sur une maille géographique plus fine (code INSEE voire code IRIS), en vue d'élargir les axes de segmentation du risque.

Annexe A

Présentation des données

A.1 Grilles des garanties par poste

Le tableau A.1 ci-dessous présente l'ensemble des garanties couvertes par la complémentaire santé de l'entreprise E.

Poste	Acte	Libellé acte
Frais médicaux de ville	Auxiliaires médicaux	Déplacement Soins infirmiers Majoration Jour Férié Soins Infi Majoration Nuit Soins infirmiers Soins infirmiers Déplacement Kinésithérapie Majoration Jour Férié Soins Infi Majoration Nuit Soins infirmiers Soins infirmiers Déplacement Kinésithérapie Majoration Jour Férié Kinésithér Indemnité forfaitaire déplacemt Majoration Nuit Kinésithérapie kinésithérapie Déplacement Orthophonie Orthophonie Pédicurie Sage femme Majoration Jour Férié Orthoptie Orthoptie
	Analyse	Examen de Laboratoire au cabinet Déplacement laboratoire Indemnités Kilométriques labo Examen de laboratoire Prélèvement
	Autres soins	Acte de spécialité Forfait soins Majoration de Nuit Actes de Spéc Acte de Spécialité KE

		Petite Chirurgie et Spécialité Maj Nuit acte de spécialité Petite Chirurgie et Spécialité 1
	Transport	Transport hors hospitalisation Transport hospitalisation chirur Transport maison de convalescence Transport hospitalisation médicale
	Radiologie	Electroradiologie (12,6) Ostéodensitométrie Acceptée Scanner Electroradiologie Electroradiologie (10,6) Electroradiologie (8,7)
Hospitalisation	Frais de séjour	Lit accompagnant hospi. Chir. Lit accompagnant hospi. Médicale Cham. part. ambulat. maison conv Chambre partic. hospi. Chirurgie Chambre partic. maison conval. Chambre particulière Psychiatrie Chambre particulière maternité Frais de séjour hospi. Chirurgie Frais de séjour maison conval.
	Hospitalisation médicale	Cham. part. Chirurgie ambulat. Cham. part. Médecine ambulatoire Chambre partic. hospi. Médicale Frais médicaux obstétriques Frais médicaux maison de conval. Frais médicaux en Psychiatrie Frais médicaux hospi. Médi. Frais de séjour Psychiatrie Frais de séjour hospi. Médicale Honraires surveillance Maison Honoraires surveillance Psy Honoraires surv hospi. Médic. Acte spécialité hospi. Image. Acte de spécialité hospi. Médi. Anesthésie hospi. Médicale Pharmacie Maison Convalescence Pharmacie en Psychiatrie Pharmacie hospi. Médicale Electroradiologie Electroradiologie hospi. Médi. Franchise actes lourds 24 euros ACCUEIL TRAITEMENT DES URGENCES
	Hospitalisation chirurgicale	Honoraires Anesthésie Honoraires Chirurgie Frais médicaux hospi. Chir

		Honoraires Survei. K Acte de spécialité Chirurgie Acte de spécialité Chirurgie Franchise actes lourds 24 euros Electroradiologie
	Forfait journalier	Forfait hospitalier Chirurgical Forfait hospitalier conval. Forfait hospitalier IMP Forfait hospi. Psychiatrie Forfait hospi. Médecine
Consultations et visites	Consultations généralistes	Consultation Généraliste Majo coordination généraliste Majoration j. férié généraliste Majoration minuit six heures Majoration nuit généraliste
	Consultations spécialistes	Consultation Cardiologue Consultation Neuropsychiatre Consultation Professeur Consultation Spécialiste Majo coordination cardiologue Majo coordination spécialiste Majoration J. férié Cardiologue Majoration j. férié Neuropsychiatre Majoration j. férié spécialiste Majoration nuit Cardiologue Majoration nuit Neuropsychiatre Majoration nuit spécialiste Majoration pédiatre
	Médecine alternative	Médecine alternative
	Visites	Dép géné ferié crit. médicaux Dép géné nuit crit. médicaux Dép Généraliste crit. médicaux Déplacement Généraliste Indemnités Kilométriques Général Maj j. férié spécialiste MAJ. nuit spécialiste Majoration Jour Férié Généraliste Majoration minuit six heures Majoration nuit Généraliste Visite Généraliste Visite Spécialiste
	Implant	Implantologie non remboursée
	Orthodontie	Orthodontie Acceptée Orthodontie Non Remboursable
	Parodontologie	Parodontologie
		Bridge céramique Couronne céramique ou adjonction

Prothèses dentaires

Dentaire

		Inter de bridge céramique Bridge métallique Couronne métal ou adjonction Inter de bridge métallique HLF175 Inlay core RAC 0 Couronne sur implant Implantologie Acceptée Prothèse dentaire non remboursée BR150-HLF280 Proth.amov.métal. Prothèse amovible résine Prothèse transitoire Réparation s/ prothèse ou pilier Prothèse transitoire NR
	Soins dentaires	Inlay/Onlay Acte de spécialité Dentaire Electroradiologie Petite Chirurgie et Spécialité Scellement/Détartrage Majoration J .férié Cons dent. MajNuit Consultation dentaire Consult. Généraliste Dentaire Consult. Spécialiste Dentaire Soins dentaires endodontie Soins dentaires Electroradiologie dentaire
Appareillage	Appareillage	Appareillage Grand appareillage Aliment nutritif Orthopédie Aide auditive de classe II Prothèse acoustique Refusée Aide auditive de classe I Répar. app. acoustique Acceptée
Optique	Montures - Forfaits	Monture - de 16 ans Monture + de 16 ans
	Verres	Verres - de 16 ans Verres + de 16 ans
	Lentilles	Lentilles Acceptées Lentilles Refusées
	Intervention Myopie	Intervention Myopie
Pharmacie	Pharmacie	Honoraire Dispensation Complexe Honoraire dispensation Age Honoraire de dispens. méd spécif Honoraire médicament remboursable Honoraire Dispensation G.C.niv.4 Honoraire de Dispensation niv.4 Honoraire Dispensation G.C.niv.7

		Honoraire de Dispensation niv.7
		Pharmacie Hospitalière
		Honoraire Dispensation G.C.niv.2
		Honoraire de Dispensation niv.2
		Pansements
		Pharmacie 30%
		Pharmacie 65%
		Pharmacie 15%
		Vaccin accepté

Table A.1 – Tableau des garanties

Annexe B

Modélisation de la prime pure et GLM

B.1 Démonstration de l'expression de la prime pure

Pour estimer la prime pure P, on cherche à minimiser l'écart quadratique MSE entre cette dernière et la charge de sinistre S estimée. Mathématiquement, cela donne l'équation B.1 ci-dessous :

$$MSE = \mathbb{E}\left[(S-P)^2\right] \tag{B.1}$$

Or, P étant déterministe :

$$\begin{aligned} \mathbb{E}\left[(S-P)^2\right] &= \mathbb{E}\left[S^2 - 2\times S\times P + P^2\right] \\ &= \mathbb{E}[S^2] - \mathbb{E}[S]^2 + \mathbb{E}[S]^2 - 2\times P\times \mathbb{E}[S] + P^2 \\ &= Var(S) + \mathbb{E}[(S-P)]^2 \end{aligned} \tag{B.2}$$

La variance $Var(S)$ étant positive, on veut donc minimiser $\mathbb{E}[(S-P)]^2$. D'où $P = \mathbb{E}[S]$.

B.2 Le modèle « Occurrence × Charge totale »

En alternative au modèle « Coût × Fréquence », il existe un modèle consistant à estimer d'une part la probabilité p_i pour un assuré i de consommer au moins un acte médical couvert, et d'autre part le montant total de l'ensemble de ses prestations perçues. C'est le modèle « Occurrence × Charge totale ».

Pour estimer cette probabilité p_i, il suffit de poser une variable de Bernoulli :

$$\mathbb{1}_i = \begin{cases} 1 & \text{si l'assuré } i \text{ effectue au moins une dépense médicale.} \\ 0 & \text{sinon.} \end{cases}$$

Par ailleurs, posons p_i la probabilité que l'assuré i effectue au moins une dépense médicale, mathématiquement :

$$\mathbb{P}\left[\mathbb{1}_i = 1\right] = p_i$$

Soit S_i la charge totale de prestations versées à l'assuré i. Les propriétés de l'espérance nous donnent l'équation B.3 ci-dessous.

$$\mathbb{E}[S_i] = p_i \times \mathbb{E}[S_i|\mathbb{1}_i = 1] + (1-p_i)\times \mathbb{E}[S_i|\mathbb{1}_i = 0] \tag{B.3}$$

Or, il est clair que $\mathbb{E}[S_i|\mathbb{1}_i = 0] = 0$ puisque si un assuré n'effectue aucune dépense, sa charge de prestation est nulle.

D'où l'égalité de l'équation B.5 ci-dessous :

$$\mathbb{E}[S_i] = p_i \times \mathbb{E}[S_i|\mathbb{1}_i = 1] \tag{B.4}$$

La charge totale de prestations S à verser par l'assureur pour un portefeuille de taille T sera la somme des charges de prestations de tous les assurés. L'estimation de cette charge se présente donc comme suit (voir équation B.5) :

$$\mathbb{E}[S] = \sum_{i=1}^{T} p_i \times \mathbb{E}[S_i|\mathbb{1}_i = 1] \tag{B.5}$$

Cette estimation permet ainsi de définir la prime pure.

Ce modèle est une généralisation des modèles modifiés en zéro [1]. Il est notamment utilisé lorsque le portefeuille comporte un grand nombre d'observations nulles.

B.3 Espérance et variance de la loi Poisson

Soit N une variable aléatoire suivant une loi Poisson de paramètre λ. Nous avons alors pour tout entier k :

$$\mathbb{P}\left[N = k\right] = \frac{\lambda^k}{k!} \times e^{-\lambda} \tag{B.6}$$

Par définition de l'espérance, on a donc :

$$\begin{aligned} \mathbb{E}[N] &= \sum_{k=1}^{\infty} k \times \mathbb{P}[N = k] = \sum_{k=1}^{N} k \times \left(\frac{\lambda^k}{k!} \times e^{-\lambda}\right) \\ &= e^{-\lambda} \times \sum_{k=1}^{\infty} k \times \left(\frac{\lambda^k}{k!}\right) = \lambda \times e^{-\lambda} \times \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \end{aligned} \tag{B.7}$$

Rappel d'un développement limités pour $x > 0$: $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ d'où si l'on pose $j = k - 1$:

$$\mathbb{E}[N] = \lambda \times e^{-\lambda} \times \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda \times e^{-\lambda} \times e^{\lambda} = \lambda \tag{B.8}$$

De plus, la variance de N s'écrit comme suit :

$$\begin{aligned} Var(N) &= \mathbb{E}[N^2] - \mathbb{E}[N]^2 = \mathbb{E}[N(N-1) + N] - \mathbb{E}[N]^2 \\ &= \mathbb{E}[N(N-1)] + \mathbb{E}[N] - \mathbb{E}[N]^2 \end{aligned} \tag{B.9}$$

Or, le terme $\mathbb{E}[N(N-1)]$ peut se calculer de façon analogique avec les équations précédentes (voir équation B.10) :

$$\begin{aligned} \mathbb{E}[N(N-1)] &= \sum_{k=2}^{\infty} k \times (k-1) \times \mathbb{P}[N = k] \\ = \sum_{k=2}^{\infty} k \times (k-1) \times \frac{\lambda^k}{k!} \times e^{-\lambda} &= \lambda^2 \times e^{-\lambda} \times \sum_{k=2}^{\infty} k \times (k-1) \times \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 \end{aligned} \tag{B.10}$$

1. Modèles développés par Lambert (1992) et Greene (1994)

Donc d'après l'équation B.9, la variance donne :

$$Var(N) = \lambda^2 + \lambda - \lambda^2 = \lambda \tag{B.11}$$

Ainsi, on a bien $\mathbb{E}[N] = Var(N) = \lambda$.

B.4 Mesure de dépendance

Le modèle « Coût × Fréquence » dépend d'une hypothèse forte qui est l'indépendance entre la fréquence et le coût. En pratique, cette hypothèse ne peut être parfaitement vérifiée puisqu'un assuré risque d'être davantage attiré par des soins aux coûts modérés, mais le besoin de soins de qualité, et donc aux coûts importants, fait varier la crédibilité de cette hypothèse. Ainsi, il existe une multitude de tests de mesure de dépendance entre deux variables, mesurée par un coefficient de corrélation. Dans le cadre de ce mémoire, on se focalise sur le test de corrélation de Pearson.

B.4.1 Coefficient de corrélation

Soit (X, Y) un couple de variables aléatoires. Le coefficient de corrélation r décrit le niveau de relation linéaire entre X et Y. Ce coefficient r est compris entre -1 et 1. Plus il s'approche de 1, plus les variables X et Y sont linéairement dépendantes, s'il s'approche de -1, on dit que les variables X et Y sont anti-corrélées, c'est-à-dire que X décroît en fonction de Y. S'il s'approche de 0, on ne peut pas assurer de dépendance entre les variables X et Y.

B.4.2 Le test de corrélation de Pearson

Soit $X = (X_1, ..., X_n)$ et $Y = (Y_1, ..., Y_n)$ deux variables aléatoire décrivant le même ensemble, et $\bar{X}$ et $\bar{Y}$ leur moyenne respective. La formule du coefficient de corrélation de Pearson du couple (X, Y) est donnée en équation B.12.

$$\hat{r} = \frac{\sum_i^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2 \times \sum_i^n (Y_i - \bar{Y})^2}} \tag{B.12}$$

Si ce coefficient donne une indication du niveau de dépendance entre deux variables, sa significativité dépend du nombre d'observations n. Le test de corrélation de Pearson mesure la significativité de son coefficient à travers la p-value.

La p-value est une probabilité permettant de mesurer la significativité des résultats d'un test statistiques, ici, le test de corrélation de Pearson. Elle permet de valider l'hypothèse nulle H_0, qui dans le cadre de ce test s'intitule : « Le coefficient de corrélation n'est pas statistiquement significatif ». Cette hypothèse est validée selon un certain seuil α, généralement fixé à 5%. En d'autres termes, si la p-value est inférieure à α, on rejette l'hypothèse H_0 en faveur de l'hypothèse alternative « Le coefficient de corrélation est statistiquement significatif ».

A titre d'exemple, si à l'issu d'un test de corrélation, le coefficient est de 0.95 et la p-value de 0.01, on rejette cette fois-ci l'hypothèse nulle, et considère ce coefficient de corrélation comme significatif : les variables sont significativement dépendantes puisque le coefficient est proche de 1.

Le calcul de la p-value fait l'objet d'études poussées qui dépassent le cadre de ce mémoire, mais son utilisation sera importante pour consolider la cohérence des modèles mis en place.

Ce test présente cependant la limite de ne pouvoir donner d'indication que sur la dépendance, et pas sur l'indépendance. Ainsi, les résultats issus de ce test ne peuvent assurer l'indépendance entre deux variables.

B.4.3 Le test du χ^2

Le test du χ^2 est un test d'hypothèse qui permet de comparer la loi d'une distribution de données observée à une loi théorique. On peut se servir du test du χ^2 en science des données pour vérifier qu'une variable soit expliquée par une autre, et à quel niveau.

Il est utilisé pour tester la dépendance entre deux variables qualitatives. La statistique de ce test compare la valeur observée à la valeur théorique.

Supposons un échantillon de taille N, et que cet échantillon soit réparti en k classes mutuellement exclusives, avec des nombres observés respectives $O_1, ..., O_k$. L'hypothèse nulle donne la probabilité p_i de tomber sur la i^{e} classe. La valeur théorique s'exprime $m_i = p_i \times N$ (voir équation B.13) :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - m_i)^2}{m_i} \tag{B.13}$$

Où O_i est la valeur observée et m_i la valeur théorique.

L'hypothèse nulle H_0 de ce test est la suivante : « Il n'y a pas de relation de dépendance entre les deux variables ».

B.4.4 Le test du V de Cramer

Le V de Cramer est en théorie plus fiable que le χ^2 car sans biais, et peut donc s'appliquer peut importe la taille de l'échantillon. On dit que le V de Cramer est insensible à l'effectif de la population. Il est calculé à partir du χ^2 par la formule suivante (voir équation B.14) :

$$V = \sqrt{\frac{\chi^2}{n \times (min(c,l)1)}} \tag{B.14}$$

Avec :

- N le nombre d'observations
- min(c,l) = le plus petit côté du tableau de contingence (le minimum entre le nombre de lignes et le nombre de colonnes)

Plus le V de Cramer est grand, plus la dépendance est forte. Il prend une valeur entre 0 et 1, il dépasse rarement 0.20. Ainsi un V de Cramer proche de 0.20 signifie que les deux variables sont fortement dépendantes.

B.5 Notation matricielle du modèle linéaire

Dans le cas où l'on cherche à expliquer un vecteur de n variables réponses à partir de p variables explicatives, il convient alors de poser :

Y $= (Y_1, Y_2, ..., Y_n)^T$ le vecteur de variables réponses ;

X $= \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & ... & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & ... & X_{2,p} \\ \vdots & \vdots & \vdots & ... & \vdots \\ 1 & X_{n,1} & X_{n,2} & ... & X_{n,p} \end{bmatrix}$ la matrice de variables explicatives ;

$\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ le vecteur de coefficients de régression du modèle ;
$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ le vecteur de résidus du modèle. Les composantes étant gaussiennes, ε est un vecteur gaussien d'espérance le vecteur nul, et de variance $\sigma_\varepsilon^2 I_n$ avec I_n la matrice identité de rang n.

Le vecteur Y peut donc s'écrire sous forme de produit matriciel (voir équation B.15) :

$$Y = X \cdot \beta + \varepsilon \tag{B.15}$$

B.6 La famille exponentielle

Lorsque la distribution d'une variable aléatoire Y_i appartient à la famille exponentielle, sa densité s'écrit sous la forme :

$$f(Y_i, \theta_i, \varphi) = exp\left\{\frac{Y_i \times \theta_i - b(\theta_i)}{a_i(\varphi)} + c(Y_i, \varphi)\right\} \tag{B.16}$$

Où :
$\theta_i \in \mathbb{R}$ est un paramètre de la moyenne ;
$\varphi > 0$ est un paramètre de dispersion ou paramètre de la variance
a_i, b et c sont des fonctions continues dérivables telles que : $\begin{cases} a_i : \mathbb{R} \longrightarrow \mathbb{R}^* \\ b \in C^2(\mathbb{R}) : \mathbb{R} \longrightarrow \mathbb{R} \\ c : \mathbb{R}^2 \longrightarrow \mathbb{R} \end{cases}$

Le tableau B.1 présente quelques exemples de lois appartenant à la famille exponentielle, et la valeur des paramètres associés.

	$a_i(\varphi)$	$b(\theta_i)$	$c(Y_i, \varphi)$
Normale	$\frac{\varphi}{\theta_i}$	$\frac{\theta_i^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\varphi} + log(2\pi\varphi)\right)$
Gamma	$\frac{\varphi}{\theta_i}$	$-log(-\theta_i)$	$\left(\frac{1}{\varphi} - 1\right) \times log(Y_i) - log(\Gamma(\frac{1}{\varphi}))$
Poisson	$\frac{\varphi}{\theta_i}$	e^{θ_i}	$-log(Y_i!)$

TABLE B.1 – Exemples de lois de la famille exponentielle

On voit que la valeur de $a_i(\varphi)$ reste inchangée pour toutes les lois.

De plus, l'espérance et la variance de la variable Y_i dépendent des paramètres θ_i et φ :

$$\mu_i = \mathbb{E}[Y_i] = b'(\theta_i) \tag{B.17}$$

$$Var(Y_i) = b''(\theta_i) \times a_i(\varphi) \tag{B.18}$$

La démonstration des équations B.17 et B.18 est disponible en Annexe B.8.
L'espérance est fonction de θ_i, le paramètre de la moyenne, et que la variance dépend de φ, le paramètre de la variance, et de θ_i. Les expressions B.17 et B.18 sont donc cohérentes et interprétables.

L'équation B.17 permet d'établir la relation B.19 ci-dessous :

$$\theta_i = b'^{-1}(\mu_i) = h(\mu_i) \tag{B.19}$$

Avec h la fonction inverse de la dérivée de b : $h = b'^{-1}$

De plus, par l'équation 3.8, on peut établir une relation avec les coefficients de régression et les variables explicatives. Ainsi, θ_i est fonction du vecteur $\beta = (\beta_0, \beta_1, ..., \beta_p)$ et de la $i^{\text{ème}}$ variable explicative $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p})$ (voir équation B.20) :

$$\theta_i(\beta, X_i) = h(\mu_i) = h\left(g^{-1}\left(\beta_0 + \sum_{k=1}^{p} \beta_k \times X_{i,k}\right)\right) \tag{B.20}$$

Pour estimer les coefficients de régression du modèle, on utilise généralement la méthode du maximum de vraisemblance.

B.7 Estimation des coefficients de régression par maximum de vraisemblance

La méthode du maximum de vraisemblance vise à rechercher les paramètres maximisant une fonction, appelée **fonction de vraisemblance**.

Dans le cas d'un GLM, la fonction de vraisemblance L dépend des paramètres de moyennes $\theta = (\theta_1, ..., \theta_n)$ et des variables réponses $Y = (Y_1, ..., Y_n)$, et s'écrit comme le produit des fonctions de densité de chaque variable réponse :

$$L(Y, \theta) = \prod_{i=1}^{n} f(Y_i, \theta_i)$$

Grâce à l'équation B.20, on peut l'exprimer en fonction des coefficients de régression et des variables explicatives :

$$L(Y, \beta, X) = \prod_{i=1}^{n} f\left(Y_i, h\left(g^{-1}\left(\beta_0 + \sum_{k=1}^{p} \beta_k \times X_{i,k}\right)\right)\right)$$

La fonction **log-vraisemblance** est simplement la fonction du logarithme népérien appliquée à la fonction de vraisemblance :

$$log(L(Y, \beta, X)) = \sum_{i=1}^{n} log\left(f\left(Y_i\,,\, h\left(g^{-1}\left(\beta_0 + \sum_{k=1}^{p} \beta_k \times X_{i,k}\right)\right)\right)\right)$$

Pour chaque i, le β_i optimal est celui qui annule la dérivée de la fonction log-vraisemblance. Il est alors possible d'estimer le coefficient β_i pour chaque i en résolvant l'équation B.21 ci-dessous :

$$\frac{\partial log(L(Y, \beta, X))}{\partial \beta_i} = 0 \quad \forall i \in \{0, ..., p\} \tag{B.21}$$

On obtient alors le vecteur des coefficients estimés $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$. La résolution de l'équation B.21 se fait rarement algébriquement. En pratique, on utilise des méthodes d'optimisation numériques comme l'algorithme de Newtom-Raphson.

B.8 Démonstration de la relation entre la moyenne et les paramètres de la famille exponentielle

Soit Y une variable aléatoire dont la distribution appartient à la famille exponentielle. La fonction génératrice des moments M_Y se développe comme suit (voir équation B.22) :

$$\begin{aligned}
M_Y(t) = \mathbb{E}[e^{t\times Y}] &= \int_{\mathbb{R}} exp\left\{t\times y + \frac{y\times\theta - b(\theta)}{a(\varphi)} - c(y,\theta)\right\} \mathrm{d}y \\
&= exp\left\{-\frac{b(\theta)}{a(\varphi)}\right\} \int_{\mathbb{R}} exp\left\{t\times y + \frac{y\times\theta}{a(\varphi)}\right\} \mathrm{d}y \\
&= \frac{exp\left\{-\frac{b(\theta)}{a(\varphi)}\right\}}{exp\left\{\frac{-b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}} \int_{\mathbb{R}} exp\left\{\frac{a(\varphi)\times t\times y + y\times\theta}{a(\varphi)} - c(y,\theta) - \frac{b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\} \mathrm{d}y \\
&= \frac{exp\left\{-\frac{b(\theta)}{a(\varphi)}\right\}}{exp\left\{\frac{-b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}} \int_{\mathbb{R}} exp\left\{\frac{y\times(a(\varphi)\times t+\theta) - b(a(\varphi)\times t+\theta)}{a(\varphi)} - c(y,\theta)\right\} \mathrm{d}y \\
&= \frac{exp\left\{-\frac{-b(\theta)}{a(\varphi)}\right\}}{exp\left\{\frac{-b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}} \\
&= exp\left\{\frac{-b(\theta) + b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}
\end{aligned} \tag{B.22}$$

Si l'on fixe $t = 0$, l'équation B.22 nous permet de déterminer $\mathbb{E}[Y]$ (voir équation B.23) :

$$\begin{aligned}
\mathbb{E}[Y] = M'_Y(t)|_{t=0} &= \left(exp\left\{\frac{-b(\theta) + b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}\right)'\Bigg|_{t=0} \\
&= exp\left\{\frac{-b(\theta) + b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}\left(\frac{b'(a(\varphi)\times t+\theta)\times a(\varphi)}{a(\varphi)}\right)\Bigg|_{t=0} \\
&= exp\left\{\frac{-b(\theta) + b(\theta)}{a(\varphi)}\right\} \times b'(\theta) \\
&= b'(\theta)
\end{aligned} \tag{B.23}$$

De même, en s'appuyant sur les équations précédentes, on peut calculer la variance de Y :

$$Var(Y) = (M''_Y(t) - M'_Y(t)^2)|_{t=0} \tag{B.24}$$

Le terme $M'_Y(t)$ est calculé dans l'équation B.23. Reste à calculer le terme $M''_Y(t)$ (voir équation B.25) :

$$\begin{aligned}
M''_Y(t) &= \left(exp\left\{\frac{-b(\theta) + b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\}\right)''\Bigg|_{t=0} \\
&= exp\left\{\frac{-b(\theta) + b(a(\varphi)\times t+\theta)}{a(\varphi)}\right\} \times \left[b'(a(\varphi)\times t+\theta)^2 + b''(a(\varphi)\times t+\theta)\times a(\varphi)\right]\Big|_{t=0} \\
&= exp\left\{\frac{-b(\theta) + b(\theta)}{a(\varphi)}\right\} \times \left[b'(\theta)^2 + b''(\theta)\times a(\varphi)\right] \\
&= b'(\theta)^2 + b''(\theta)\times a(\varphi)
\end{aligned} \tag{B.25}$$

Ainsi, la variance s'écrit comme dans l'équation B.26 ci-dessous :

$$\begin{aligned} Var(Y) &= b'(\theta)^2 + b''(\theta) \times a(\varphi) - b'(\theta)^2 \\ &= b''(\theta) \times a(\varphi). \end{aligned} \tag{B.26}$$

Annexe C

Méthodes de classification

C.1 Méthode des k-means

C.1.1 Principe de la méthode des k-means

La méthode des k-means est une des méthodes de classification par apprentissage les plus utilisées dans cadre du traitement de données volumineuses. Son objectif est de regrouper un ensemble de données dans différents sous-groupes. Le nombre de sous-groupes, k, doit être choisi a priori.

Cet algorithme de classification sépare les données dans les groupe les mieux adaptés sur la base d'un apprentissage statistique. Les données sont séparées en k sous-groupes différents, qui sont généralement construits pour être suffisamment distincts les uns des autres. Cette distinction se mesure à travers l'éloignement dans l'espace en distance euclidienne.

Initialement, chaque groupe a un centre, appelé centroïde, et un point de données est classé dans un certain groupe en fonction de la proximité de ses caractéristiques avec le centroïde. Par la suite, les centroïdes sont re-calculés, et l'opération est répétée de manière itérative jusqu'à ce que la distance euclidienne entre les points de chaque groupe et leur centroïde respectif soient minimisés, on dit alors que l'algorithme converge.

Ainsi, l'algorithme des k-means peut s'écrire comme suit :

1. k points de la base de données sont choisis de manière aléatoire pour être le centroïde de leur classe respective.
2. Calculer la distance entre chaque point de la base de données et les k centroïdes. On choisit généralement la distance euclidienne pour calculer la distance entre chaque point de la base et les centroïdes initialisés.
3. Pour former les k sous-groupes, chaque point de la base de données est assigné au centroïde le plus proche.
4. Re-calculer les centroïdes en faisant la moyenne de tous les points de données assignés à chaque classe afin de réduire la variance au sein de ces classes.
5. Si la valeur du nouveau centroïde a changé, répéter l'algorithme à partir de l'étape 2 jusqu'à ce que :
   — le centroïde reste inchangé à l'issu de l'étape 4 ;
   — la somme des distances entre les points de chaque classe et leur centroïde respectif est la même ;
   — l'ensemble des points de données affectés à chaque classe reste inchangé.

L'interprétation mathématique de cet algorithme est développée en annexe C.1.2.

Cette méthode a pour avantages :
— Sa simplicité. L'algorithme des k-means est relativement facile à mettre en place. De plus, il est assez peu complexe en terme de calculs puisqu'il n'a que les distances et les moyennes à calculer.
— Elle est adapté aux données importantes.
— Elle converge systématiquement, bien que le nombre d'itérations nécessaires puisse être plus ou moins conséquent.

Toutefois, elle présente en outre quelques limites :
— Le nombre de sous-groupes k doit être choisi arbitrairement.
— La structure des sous-groupes peut varier en fonction des centroïdes initialement et aléatoirement choisis.
— Bien qu'adaptée aux données importantes, la méthode n'est pas adaptée pour regrouper un grand nombre de données dans une même classe.
— Les valeurs aberrantes doivent être traités en amont, sans quoi elles peuvent être choisies comme centroïdes initiaux, ou être intégrées dans un sous-groupe au lieu d'être ignorées.
— Mal adaptée lorsque le nombre de critères est trop important.

C.1.2 Aspects mathématiques

Définition d'une distance

On appelle d une distance sur un ensemble E, une application de $E \times E$ dans $\mathbb{R}_+$ vérifiant les propriétés suivantes :
— La symétrie : $\forall (a,b) \in E \times E,\ d(a,b) = d(b,a)$;
— La séparation : $\forall (a,b) \in E \times E,\ d(a,b) = 0 \iff a = b$;
— L'inégalité triangulaire : $\forall (a,b,c) \in E \times E \times E,\ d(a,c) \leq d(a,b) + d(b,c)$;

Définition de la distance euclidienne

Soit d une distance sur un espace vectoriel normé $(E, \|.\|)$. Soient $X = (x_1, ..., x_n)$ et $Y = (y_1, ..., y_n)$ deux points de E. La distance euclidienne entre X et Y s'écrit comme suit :

$$d(X,Y) = \|X - Y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Algorithme des k-means

Soit $\{X_1, X_2, ..., X_n\}$, un ensemble de n points. L'objectif est de partitionner cet ensemble en k ($\leq n$) sous-groupes $S = (S_1, ..., S_k)$.

L'algorithme s'écrit mathématiquement comme suit :

1. Choisir k points dans $\{X_1, X_2, ..., X_n\}$ de façon aléatoire. Ces points seront les centroïdes initiaux et sont notés $m_1^{(1)}, m_2^{(1)} ..., m_k^{(1)}$.
2. A la $t^{\text{ième}}$ itération, calculer la distance entre chaque point et chaque centroïde :

$$\forall i \in \{1, ..., n\}, \forall l \in \{1, ..., k\},\ d(X_i, m_l^{(t)}) = \|X_i - m_l^{(t)}\|$$

3. Pour chaque l, réunir dans le sous-groupe S_l les points les plus proches du centroïde $m_l^{(T)}$:

$$\forall l \in \{1, ..., k\},\ S_l^{(T)} = \left\{ X_i\,,\ \ \|X_i - m_l^{(T)}\| \leq \|X_i - m_{l^*}^{(T)}\|\ \forall l^* \in \{1, ..., k\} \right\}$$

4. Déterminer la valeur des prochains centroïdes $m_l^{(T+1)}$ en calculant la moyenne des points au sein de chaque sous-groupe :

$$m_l^{(t+1)} = \frac{1}{\|S_l^{(t)}\|} \times \sum_{X_i \in S_l^{(t)}} X_i$$

5. Répéter l'algorithme à partir de l'étape 2 jusqu'à ce que :
   — $m_l^{(t+1)} \approx m_l^{(t)}$
   — $\|S_l^{(t+1)}\| \approx \|S_l^{(t)}\|$
   — $S_l^{(t+1)} \approx S_l^{(t)}$

C.2 Classification ascendante hiérarchique

Définition d'une partition

Soit E un ensemble quelconque. Un ensemble P de sous-ensembles de X est une partition de E si, et seulement si les conditions suivantes sont vérifiées :
— $\emptyset \notin P$;
— $\bigcup_{A \in P} A = E$
— $\forall A, B \in P, A \neq B \implies A \cap B = \emptyset$

Définition d'une hiérarchie

Soit $E = \{x_1, ..., x_n\}$ un ensemble d'individus, et soit $H = \{H_1, ..., H_k\}$ un ensemble de parties de E. H est une hiérarchie de E si, et seulement si les conditions suivantes sont vérifiées :
— $\emptyset \in H$
— $\forall x \in E,\ \{x\} \in H$
— $E \in H$
— $\forall A, B \in H$: soit $A \subset B$, soit $B \subset A$, soit $A \cap B = \emptyset$.

Définition d'un diamètre

Soit E un espace quelconque et d une distance sur E. Le diamètre d'une partition P de E est une application de E dans $\mathbb{R}_+$ définie comme suit :

$$D(P) = sup\{d(a, b) : a, b \in P\}$$

Algorithme de la classification ascendance hiérarchique

Soit $E = \{X_1, X_2, ..., X_n\}$, un ensemble de n points à classifier. On cherche à construire une suite de partitions de E, $P_1 \subset P_2 \subset ... \subset P_{n-1} \subset P_n$, avec P_t contenant $n - t + 1$ classes pour chaque $t \in \{1, ..., n\}$. Soit d la distance appliquée à cette méthode.

1. Initialisation : $P_1 = \{\{X_1\}, ..., \{X_n\}\}$

2. Pour $t \in \{1, ..., n-1\}$:
   Calculer le diamètre de chaque union des éléments de P_t, et conserver l'union dont le diamètre est minimal pour P_{t+1} :

$$P_t = \{C_1^{(t)}, ..., C_{n-t+1}^{(t)}\}, C_i^{(t)} \subset E \ \forall i \in \{1, .., n-t+1\}$$

   Alors :

$$P_{t+1} = \{C_i^{(t)} \cup C_j^{(t)}, C_l^{(t)} \ \forall l \neq i, j\}$$

   Avec $C_i^{(t)}$ et $C_j^{(t)}$ choisis tels que :

$$D(C_i^{(t)} \cup C_j^{(t)}) \leq D(C_p^{(t)} \cup C_q^{(t)}) \ \forall p, q \in \{1, ..., n-t+1\}$$

Ainsi, il est clair que $P_n = \bigcup_{C \in P_{n-1}} C = E$ par définition de la partition.

La distance d diffère selon la méthode. Voici une liste non-exhaustive des diamètres appliquées à ces indices de distance :

— Pour la **méthode à lien simple** : $D(A \cup B) = min\{d(a, b), a \in A, b \in B\}$
— Pour la **méthode à lien complet** : $D(A \cup B) = max\{d(a, b), a \in A, b \in B\}$
— Pour la **méthode à lien moyen** : $D(A \cup B) = \frac{1}{|A| \times |B|} \times \sum_{a \in A, b \in B} d(a, b)$ où $|.|$ représente le cardinal d'un ensemble.
— Pour la **méthode de Ward** : $D(A \cup B) = \frac{|A| \times |B|}{|A| + |B|} \times d(\bar{x}_A, \bar{x}_B)^2$ où $\bar{x}_A$ (resp. $\bar{x}_B$) est le centre de gravité de A (resp. B), qui se définit comme suit :
  Pour un ensemble C, $\bar{x}_C = \frac{1}{|C|} \times \sum_{x \in C} x$ est le centre de gravité de C.

Annexe D

Tests d'adéquation de lois

D.1 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test statistique non-paramétrique visant à évaluer le degré de similitude entre une fonction de répartition empirique et une fonction de répartition théorique.

Considérons un vecteur aléatoire $(X_1, ..., X_n)$ de n variables aléatoires indépendantes et identiquement distribuées, et de fonction de répartition F_n. Soit F la fonction de répartition d'une loi de probabilité théorique.

La statistique T de ce test est l'écart maximal entre la fonction de répartition empirique donnée et la fonction de répartition théorique (voir équation D.1) :

$$T_n = max_{x \in \mathbb{R}} \{F_n(x) - F(x)\} \tag{D.1}$$

L'hypothèse nulle H_0 de ce test peut se formuler « La distribution de l'échantillon a pour fonction de répartition F ». Mathématiquement, sous l'hypothèse H_0, on a
$T_n \underset{n \to +\infty}{\longrightarrow} 0$.

D.2 Test de Cramer-Von Mises

Au même titre que le test de Kolmogorov-Smirnov, le test de Cramer-Von Mises mesure la similitude d'une distribution donnée à une distribution théorique en se basant sur l'écart entre les fonctions de répartition respectives F_n et F, sur l'ensemble des observations. Sa statistique est définie comme suit (voir équation D.2) :

$$T_n = \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 \, \mathrm{d}F^*(x) \tag{D.2}$$

L'hypothèse nulle H_0 est la même que pour le test de Kolmogorov-Smirnov. De plus, cette hypothèse équivaut à : $T_n \underset{n \to +\infty}{\longrightarrow} 0$.

Annexe E

Méthodes d'apprentissage supervisé

E.1 Processus de prédiction des forêts aléatoires

Soit $C = \{C_1, ..., C_n\}$ l'ensemble des classes auxquelles est susceptible d'appartenir une variable X, et Δ_n l'ensemble des probabilités de C, défini comme ci-dessous :

$$\Delta_n = \{p_1, ..., p_n | \sum_{i=1}^{n} p_i = 1,\ p_i \geq 0\}$$

Soit e_i un élément canonique de Δ_n avec 1 en i^{e} position. Si un arbre A prédit qu'une instance k appartient à C_i, on peut écrire : $\hat{Y}_{k,A} = e_i$. Cela fait le lien entre les prédictions d'un arbre et l'ensemble Δ_n des mesures de probabilité de C :

$$\hat{Y}_k = \frac{1}{N} \sum_{i=1}^{n} \hat{Y}_{k,A}$$

Avec N le nombre d'arbres dans la forêt. On a donc $\hat{Y}_k \in \Delta_n$ et la prédiction de la valeur finale pour l'instance k coïncide avec la classe C_i pour laquelle le i^{e} élément de Y_k est maximal.

E.2 Algorithme du Gradient Boosting

Soit L la fonction de perte, qui représente l'écart entre une valeur prédite $f(x_i) = \hat{y}_i$ et une valeur observée y_i (à titre d'exemple, on peut choisir $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$). Le Gradient Boosting vise à minimiser cette fonction L.

Enfin, pour tout i, on note la fonction f_i comme suit :

$$f_{i+1}(x) = f_i(x) + \eta \Delta L(y, f(x))$$

Avec ΔL le gradient de L.

A partir de ces éléments, l'algorithme du Gradient Boosting, utilisé à des fins régressives, procède comme suit :

1. Initialisation : $f_0(x) = argmin_\theta \sum_{i=1}^{n} L(y_i, \theta)$
2. Pour chaque k, calculer les pseudo-résidus :

$$r_{i,k} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right] \Bigg|_{f=f_k} \quad \forall i \in \{1, ..., n\}$$

3. Ajuster d'un arbre de décision $h_k(x)$ entraîné sur l'échantillon (x, r_k) avec $x = (x_1, ..., x_n)$ les variables explicatives et $r_k = (r_{1,k}, ..., r_{n,k})$ les variables réponses.
4. Actualiser $f_k(x) = f_{k-1}(x) + \theta_k h_k(x)$, avec $\theta_k = argmin_\theta \sum_{i=1}^{n} (L(y_i, f_{k-1}(x_i)) + \theta h_m(x_i))$
5. Répéter les étapes 2 à 4 jusqu'à ce que $L(y, f_k(x)) < \varepsilon$, avec $\varepsilon > 0$ fixé en amont.

E.3 Algorithme des k plus proches voisins

L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé ayant pour objectif de classer des points un point selon un vote majoritaire des k points les plus proches.

Notons $X = \{X_1, ..., X_n\}$ la base d'apprentissage où pour chaque i, X_i est un point d'apprentissage et $C = \{C_1, ..., C_n\}$ représente les classes respectives correspondantes. On pose X^* un point dont la classe est inconnue. Sa classe est déterminée comme suit :

1. Mesurer la similarité entre X^* et chacun des points de X. Cette similarité peut se mesurer via la distance euclidienne par exemple.
2. Trouver les k voisins les plus proches par la distance euclidienne dans la base d'apprentissage, et attribuer à X^* la classe à laquelle appartient la majorité de ces voisins.

L'étape 2 peut varier selon que l'algorithme soit utilisé à des fins de classification ou de régression. Dans le second cas, le point X^* prendra la valeur moyenne des k points les plus proches. Formellement, on peut l'écrire comme suit :

Soient les couples $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ où, pour chaque i, X_i est un point de la base d'apprentissage, et Y_i sa variable réponse, et soit X^* un point dont on ignore la variable réponse. Notons alors $\{X_{i_1}, X_{i_2}, ..., X_{i_k}\}$ ses k plus proches voisins où, pour chaque i_j, $X_{i_j} \in \{X_1, ..., X_n\}$, et choisis tels que :

$$\|X_{i_1} - X^*\| \leq \|X_{i_2} - X^*\| \leq ... \leq \|X_{i_k} - X^*\| \leq \|X_i - X^*\| \ \ \forall i \notin \{i_1, ..., i_k\}$$

La variable réponse de X^* sera alors modélisée comme suit :

$$Y^*(X^*) = \frac{1}{k} \times \sum_{i \in \{i_1, ..., i_k\}} Y_i$$

Pour que la méthode soit optimale, il convient de choisir la bonne valeur de k. Pour cela, il existe une méthode itérative, la validation croisée.

E.4 Méthode des k-fold (validation croisée)

La validation croisée est fréquemment utilisée pour comparer les performances de modèles de Machine Learning sur un échantillon de données limité. L'objectif de cet algorithme est de trouver le nombre de groupes k dans lesquels un échantillon donné doit être réparti.

Un échantillon donné est divisé en k sous-échantillons de tailles égales. Parmi ces k échantillons, un seul est retenu pour être la base de test, et les $k-1$ autres servent de base d'apprentissage. On récupère ensuite l'erreur de prédiction de ce modèle. Le processus est alors répété k fois, de sorte que chaque sous-échantillon soit considéré tour à tour comme base de test tandis que les $k-1$ autres servent de base d'apprentissage, en récupérant à chaque itération l'erreur de

prédiction. Les k résultats obtenus sont alors moyennés pour obtenir une estimation de l'erreur global. Ainsi, le k optimal est celui pour lequel l'erreur globale est minimisée.

L'algorithme procède comme suit :

1. Tirer aléatoirement et sans remise l'ensemble des données de l'échantillon ;
2. Diviser l'ensemble du tirage en k groupes ;
3. Pour chaque groupe :
   — Considérer le groupe comme une base de test ;
   — Prendre les $k-1$ groupes restant comme une base d'apprentissage ;
   — Construire un modèle d'apprentissage et le tester sur l'ensemble des $k-1$ groupes ;
   — Conserver l'erreur de prédiction, et passer au groupe suivant ;
4. Moyenner l'ensemble des erreurs retenues.

La figure E.1 illustre cet algorithme.

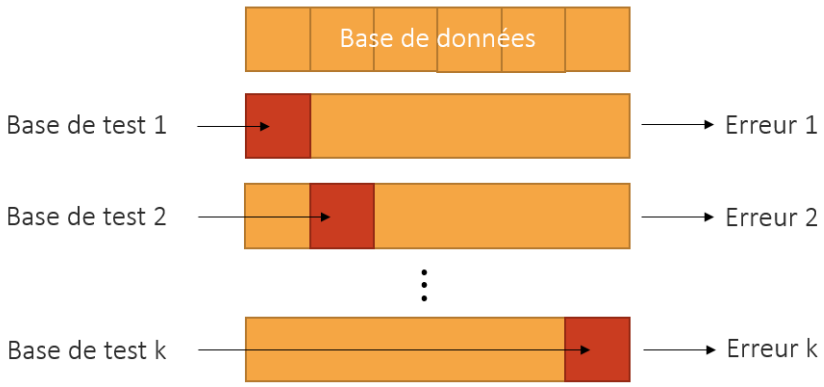

FIGURE E.1 – Algorithme des k-fold

E.5 Les indicateurs d'écart

Les indicateurs de performance basés sur l'écart ont pour objectif de comparer la qualité de prédiction de plusieurs modèles. Le modèle le plus qualitatif est celui pour lequel l'indicateur est le plus faible.

E.5.1 La Mean square error

La Mean Square Error (ou MSE), traduit « le carré moyen des erreurs » est la moyenne arithmétique du carré des écarts entre les prévisions d'un modèle et les observations.

Soit $y = (y_1, ..., y_n)$ un échantillon observé, et soit $\hat{y} = (\hat{y_1}, ..., \hat{y_n})$ la prédiction de y. On écrit la MSE comme suit (voir équation E.1) :

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{E.1}$$

Cet indicateur de performance a pour avantage de pénaliser les erreurs importantes via le passage au carré de l'erreur résiduelle.

E.5.2 La Root mean square error

La Root mean square error (ou RMSE), traduite « la racine carrée du carré moyen des erreurs » est la racine carrée de la MSE. Elle s'écrit donc comme suit :

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

Comme la MSE, cet indicateur pénalise les grosses erreurs.

E.5.3 La Mean absolute error

La Mean absolute error (ou MAE) est la moyenne de la valeur absolu des écarts entre une prédiction et une observation. Elle s'écrit comme suit :

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{E.3}$$

Annexe F

Modélisation de la sinistralité

F.1 Indépendance entre la fréquence et le coût

Le niveau d'indépendance entre deux variables peut se mesurer à travers le test de corrélation de Pearson.

Le calcul du coefficient de corrélation de Pearson a été fait pour l'acte choisi, et sur une base de prestations par assuré regroupant le nombre et la charge de sinistres de chaque assuré. De plus, ce calcul est effectué sur les observations non-nulles, c'est-à-dire sur les personnes ayant consommé au moins une fois l'acte étudié et donc ayant une charge de sinistres non-nulle. En effet, conserver les observations nulles risque d'engendrer un biais conséquent quant à la dépendance entre la fréquence et le coût puisqu'une quantité nulle implique nécessairement un montant nul.

Ainsi, en effectuant le test de corrélation de Pearson entre la fréquence et le coût moyen on s'attend à obtenir un coefficient proche de 1 en cas de dépendance croissante, et proche de -1 en cas de dépendance décroissante.

Le test de corrélation de Pearson effectué pour l'acte *Consultations spécialistes* donne un coefficient de corrélation de 0.003, avec une p-value de 0.59. Le test ne nous permet pas de valider avec quasi-certitude l'hypothèse d'indépendance. La figure F.1 ci-dessous présente la dispersion de la fréquence en fonction du coût pour cet acte.

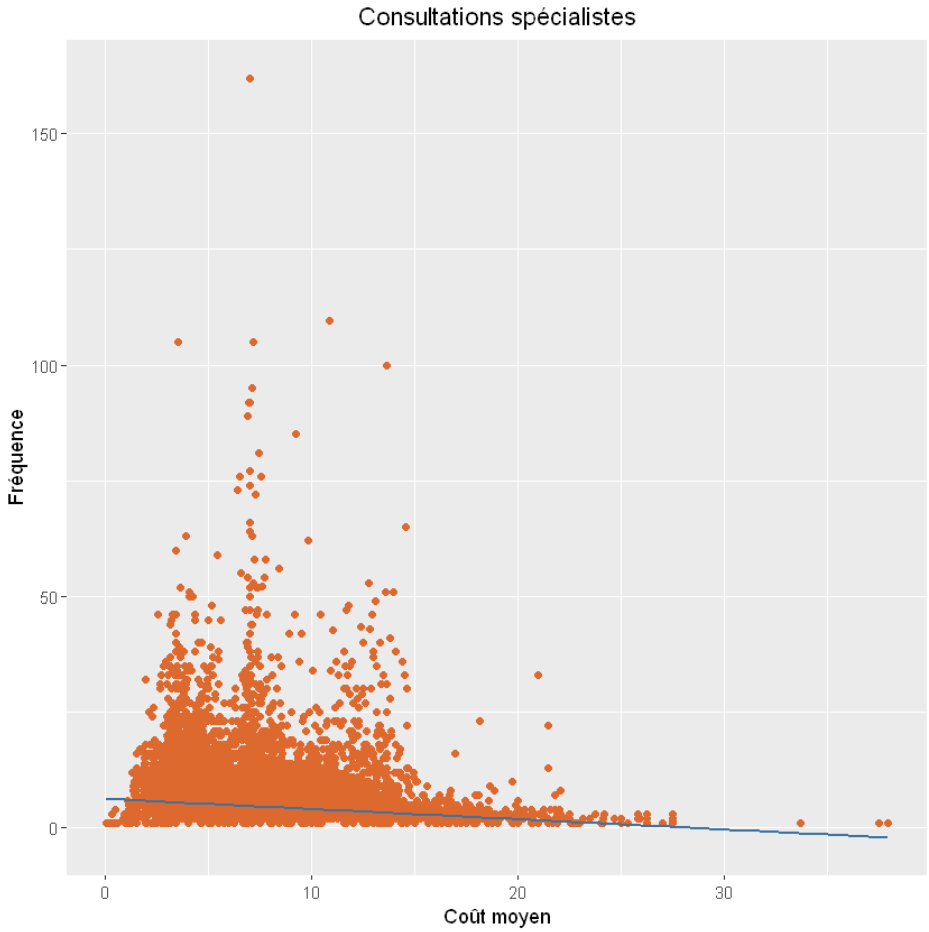

FIGURE F.1 – Diagramme de dispersion pour l'acte *Consultations spécialistes*

La dispersion semble uniforme, et on peut constater une décroissance de la droite de régression, qui peut s'expliquer par le fait que les montants élevés soient moins consommés en termes de quantité.

A noter qu'on peut apercevoir une forte représentation du coût de 7,50€, qui correspond, dans les garanties de ce contrat, au remboursement complémentaire d'une consultation spécialiste d'un médecin conventionné secteur 1.

Les résultats du test ne nous permettent pas d'affirmer une dépendance entre le coût et la fréquence pour l'acte *Consultations spécialistes*. Ainsi, l'hypothèse d'indépendance sera par défaut retenue.

Annexe G

Étude des résidus

G.1 Importance des variables pour la modélisation de l'effet spatial

G.1.1 Effet spatial fréquence

L'effet spatial fréquence a été modélisé à partir des résidus grâce aux v variables géographiques.

Le niveau d'importance de chaque variable pour le Random Forest est donné dans le tableau G.1.1.

Variable	Importance
Médecins généraliste	46.38
Chirurgiens dentiste	45.06
Infirmier	71.23
Radiologue	19.43
kinésithérapeute	65.70
Psychologue	35.02
Sage femme	23.44
Niveau de vie	146.43
Nombre de ménages	123.08
Superficie	118.30
Naissance	133.59
Décès	96.88
Logements	91.28
Travailleurs	100.69
Population totale	90.91
Chômeurs	102.89
Présence de site	96.69
Ménages fiscaux	281.97
Taux de pauvreté	168.85

TABLE G.1 – Importance des variables géographiques pour la modélisation de l'effet spatial par Random Forest

Le tableau G.1.1 donne l'importance des variables géographiques pour la modélisation de l'effet spatial par GBM.

Variable	Importance
Ménages Fiscaux	36.61
Présence de site	17.17
Taux de pauvreté	16.66
Naissance	7.70
Niveau de vie	3.46
Superficie	2.87
Nombre de ménages	2.80
kinésithérapeute	2.06
Décès	1.73
travailleurs	1.39
Logement	1.36
Population totale	1.23
Infirmier	0.48
Radiologie	0.39
Chirurgien dentiste	0.13
Sage femme	0.10
Médecin généraliste	0.08
Psychologue	0.07

TABLE G.2 – Importance des variables géographiques pour la modélisation de l'effet spatial par GBM

G.1.2 Effet spatial coût

L'effet spatial coût a été modélisé à partir des résidus grâce aux variables géographiques.

Le niveau d'importance de chaque variable pour le Random Forest est donné dans le tableau G.3.

Variable	Importance
Médecin généraliste	304.62
Infirmier	437.90
Radiologue	84.04
kinésithérapeute	491.60
Niveau de vie	1078.07
Nombre de ménages	715.96
Superficie	1393.13
Naissance	801.00
Deces	495.62
Logement	619.81
travailleurs	581.93
chômeurs	628.39
Psychologue	386.78
Sage femme	156.31
Menages fiscaux	6233.60
Taux de pauvreté	2864.90

TABLE G.3 – Importance des variables géographiques pour la modélisation de l'effet spatial coût par Random Forest

Le niveau d'importance de chaque variable pour le GBM est donné dans le tableau G.4

variable	Importance
Ménages Fiscaux	59.79
Taux de pauvreté	23.50
Population totale	3.03
Superficie	2.84
Naissance	2.07
Logement	0.80
Nombre de ménage	0.74
Niveau de vie	0.67
Psychologue	0.66
Deces	0.49
Radiologues	0.33
Masseur.kinésithérapeute	0.31
Chômeurs	0.20
Travailleurs	0.19
Sage femme	0.14
Médecin généraliste	0.13
Infirmier	0.00

TABLE G.4 – Contribution des variables géographiques pour la modélisation de l'effet spatial coût par GBM

G.2 Choix du nombre de classes k pour l'effet spatial coût

Comme pour l'effet spatial fréquence, pour choisir notre k, on observe l'évolution de la variance intra-classe en fonction du nombre de classes, et choisit le seuil à partir duquel cette variance se stabilise.

Maille départementale

La figure G.1 montre l'évolution de la variance intra-classe de l'effet spatial pour la maille départementale.

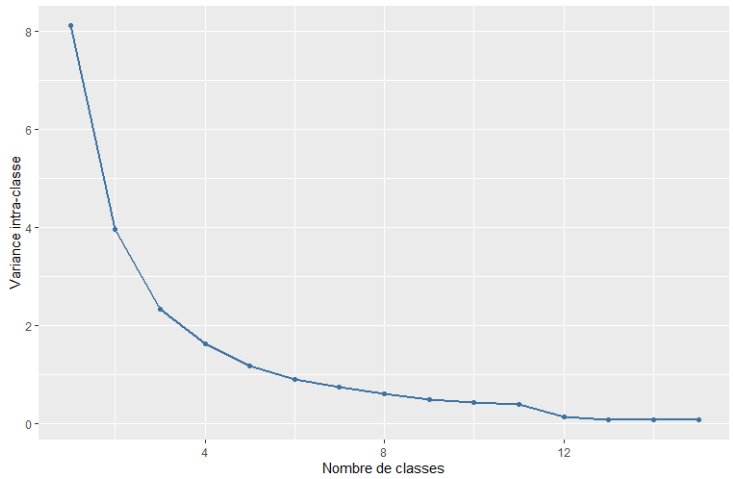

FIGURE G.1 – Variance intra-classes par nombre de classes k

La variance semble converger à partir de 12 classes, ce qui peut sembler conséquent pour regrouper des départements, d'autant plus que, pour $k > 11$, certaines classes ne regroupent

qu'un département.

Observons l'AIC du modèle complet évoluer en fonction du nombre de classes k. Voir tableau G.2.

Nombre de classes	AIC
2	16 828
3	16 674
4	16 617
5	16 526
6	16 516
7	16 517
8	16 519
9	16 507
10	16 507
11	16 508
12	16 504
13	16 503
14	16 501
15	16 503

TABLE G.5 – Évolution de l'AIC du modèle complet en fonction du nombre de classes

Au vu de l'évolution de l'AIC, on peut conclure qu'un nombre de classes $k = 9$ soit suffisant. C'est le nombre de classes qui sera utilisé pour le zonier à la maille départementale.

Maille code postal

La figure G.2 montre l'évolution de la variance intra-classe de l'effet spatial pour la maille code postal en faisant varier k de 1 à 20.

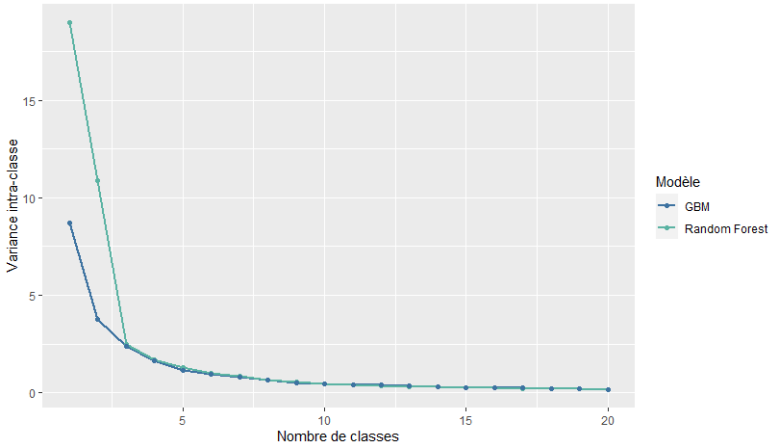

FIGURE G.2 – Variance intra-classes par nombre de classes k

La variance semble converger à partir de 6 classes, pour nos deux modèles.

Observons l'AIC du modèle complet évoluer en fonction du nombre de classes k. Voir tableau G.2.

Nombre de classes	Random Forest	GBM
2	16 600	17 447
3	16 435	17 156
4	16 319	17 121
5	16 234	17 121
6	16 228	17 121
7	16 203	17 122
8	16 186	17 096
9	16 188	17 096
10	16 179	17 098
11	16 141	17 090
12	16 143	17 091
13	16 137	17 091
14	16 135	17 093
15	16 137	17 092
16	16 138	17 091
17	16 132	17 092
18	16 131	17 094
19	16 129	17 076
20	16 129	17 078

TABLE G.6 – Évolution de l'AIC du modèle complet en fonction du nombre de classes

L'AIC s'améliore en fonction du nombre de classes, mais cette amélioration est modérée à partir de $k = 11$. C'est le nombre de classes qui sera retenu pour le zonier à la maille code postal.

Bibliographie

[Ailliot, 2020] Ailliot, P. (2020). Théorie de la crédibilité. Notes de cours `http://pagesperso.univ-brest.fr/~ailliot/doc_cours/M1EURIA/credib/main.pdf`.

[Ameli,] Ameli. Le site officiel de l'assurance maladie. `http://ameli.fr/` Consulté le 13 Avril 2021.

[Arbogast et Mahuzier, 2014] Arbogast, E. et Mahuzier, A. (2014). Performance de la crédibilité. Journées d'études IARD de l'Institut des Actuaires.

[Bonnifait, 2019] Bonnifait, C. (2019). *Optimisation d'un outil de tarification santé destiné au pilotage des grands comptes et Branches professionnelles.* Mémoire, EURIA.

[Brownlee, 2020] Brownlee, J. (2020). *A Gentle Introduction to k-fold Cross-Validation.* Machine Learning Mastery. `https://machinelearningmastery.com/k-fold-cross-validation/`. Page consultée le 23 Août 2021.

[Buzyn, 2017] Buzyn, A. (2017). Renforcer l'accès territorial aux soins. `https://solidarites-sante.gouv.fr/IMG/pdf/acces_aux_soins_dp_vdef_131017.pdf`.

[Bühlmann, 1967] Bühlmann, H. (1967). Introduction report experience rating and credibility. *ASTIN Bulletin*, 4(3):199–207.

[Cook, 1982] Cook, Dennis Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York, 2 édition.

[Delcaillau, 2019] Delcaillau, D. (2019). *Contrôle et Transparence des modèles complexes en actuariat.* Mémoire, EURIA.

[Delignette-Muller et Dutang, 2021] Delignette-Muller, M.-L. et Dutang, C. (2021). *Help to Fit of a Parametric Distribution to Non-Censored or Censored Data.* CRAN. `https://cran.r-project.org/web/packages/fitdistrplus/fitdistrplus.pdf`. Consultée le 10 Juillet 2021.

[Hastie *et al.*, 2001] Hastie, T., Tibshirani, R. et Friedman, J. (2001). The elements of statistical learning. chapitre 13. Springer.

[Holtz, 2018] Holtz, Y. (2018). *The R Graph Gallery.* `https://www.r-graph-gallery.com/`. Consultée le 18 Août 2021.

[Jr., 1951] Jr., F. J. M. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46:68–78.

[Lambert, 1992] Lambert, D. (1992). *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing.*

[Laval, 2020] Laval, E. (2020). *Impact de la réforme 100% Santé sur les contrats de frais de santé par un modèle de crédibilité.* Mémoire, Université Paris-Dauphine.

[Leslie Jones, 2019] Leslie Jones, P. M. (2019). *L'application de la théorie de la crédibilité dans l'industrie canadienne de l'assurance-vie* . Rapport technique 219120, Institut canadien des actuaires (ICA), Society of Actuaries (SOA).

[Mingfeng Lin, 2013] Mingfeng Lin, Henry C. Lucas Jr, G. S. (2013). Research commentary—too big to fail : Large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill, New York.

[Molnar, 2021] Molnar, C. (2021). *Interpretable Machine Learning.* A Guide for Making Black Box Models Explainable `https://christophm.github.io/interpretable-ml-book/index.html`. Consultée le 5 Juillet 2021.

[Palczewska *et al.*, 2014] Palczewska, A., Palczewski, J., , Robinson, R. M., et Neagu, D. (2014). Interpreting random forest classification models using a feature contribution method. *In Integration of Reusable Systems*, chapitre 26, pages 193 – 218. Springer.

[Pesneaud, 2019] Pesneaud, A. (2019). *Création de zoniers en assurance habitation à l'aide de variables externes et de méthodes de Data Science.* Mémoire, ISUP.

[Refaeilzadeh *et al.*, 2009] Refaeilzadeh, P., Tang, L. et Liu, H. (2009). *Cross-Validation.* Encyclopedia of Database Systems. Springer.

[Saporta, 2006] Saporta, G. (2006). *Probabilités, analyses des données et statistique.* Éditions Technip, 2e édition.

[Sepulvada, 2015] Sepulvada, C. (2015). *Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial.* Mémoire, ISFA.

[Sharman *et al.*, 2007] Sharman, R., Ramesh, R. et Kishore, R. (2007). Regression analysis in ncss. *In Ontologies : a handbook of principles, conceptsand applications in information systems*, volume 14 de *Integrated series in information systems*, chapitre 2, pages 21–47. Springer, New York.

[Stephan et Lazic, 2020] Stephan, A. et Lazic, S. (2020). Tarification et suivi d'un régime de frais de soins. Caritat Recherche & Formation.

[Thwe, 2019] Thwe, Y. M. (2019). *Applying Clustering Techniques for Refining Large Data Set (Case Study on Malware).* Graduate School of Advanced Science and Technology.

[Van Der Laan *et al.*, 2007] Van Der Laan, M. J., Polley, E. C. et Hubbard, A. E. (2007). *Super learner, Statistical Applications in Genetics and Molecular Biology.*

[Vermet, 2020] Vermet, F. (2020). Apprentissage statistique. Notes de cours.

[Wasserman, 2013] Wasserman, L. (2013). *All of Statistics : A Concise Course in Statistical Inference.* Springer Science Business Media.