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

Par : Catalina SEPULVEDA

Sujet : Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial.

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

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

Membre présents du jury de l'Institut des Actuaires

Membres présents du jury de la filière

Entreprise :

Nom : GENERALI

Signature :

Directeur de mémoire en entreprise :
Nom : Annabelle BONGO

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Secrétariat

Bibliothèque :

Résumé

Le présent mémoire porte sur la révision du zonage en santé. il a pour objectif de décrire et comparer deux méthodes qui visent à modèliser le risque géographique, grâce à un lissage effectué sur les résidus spatiaux d'un modèle linéaire généralisé de prime pure. Le risque total est d'abord décomposé par postes, pour constituer des périmètres de risque homogènes, afin d'optimiser les ajustements de prime pure de chaque poste aux lois théoriques choisies. Les postes traités sont l'hospitalisation, les soins courants, l'optique, la pharmacie et le dentaire. Chaque modélisation est initialisée avec les variables tarifaires classiques, zonier actuel excepté, et des variables externes significatives sont ajoutées afin d'expliquer provisoirement une partie de l'effet géographique, elles font ainsi office de zonier partiel provisoire. Les résidus de ce modèle sont ensuite projetés sur une carte puis lissés par deux méthodes, pour en extraire l'effet géographique manquant. Ce dernier est enfin combiné à l'effet des variables externes pour donner l'effet géographique total, qui une fois classé par la méthode de Ward constitue directement le zonier. L'idée de ce processus est d'avoir un effet tarifaire nettement distingué de l'effet géographique, c'est pourquoi le modèle est aussi adapté en neutralisant l'impact de la modalité Alsace-Moselle au niveau de la variable tarifaire régime, un assuré appartient à cette modalité si et seulement si il réside dans cette région-là, le critère de définition est donc purement géographique. Après l'estimation des effets externes dans ce modèle de prime pure, le traitement primordial est donc celui de l'application des deux méthodes de lissage des résidus du modèle ainsi ajusté.

La première méthode présentée est une méthode stochastique nommée, *Adjacency*, basée sur l'approche bayésienne, où le lissage est local et prend en compte le risque des communes immédiatement voisines, c'est-à-dire adjacentes. Nous avons comparé les résultats obtenus avec une méthode déterministe nommée *Distance*, qui utilise la théorie de la crédibilité. Cette deuxième méthode lisse de manière globale en tenant compte des risques de l'ensemble des communes de France, avec une influence décroissante en fonction de la distance à la commune voisine. Un niveau de lissage optimal est calculé pour chaque méthode, il correspond au point où les résidus lissés sont les plus prédictifs, plus proche de l'effet réel moyen recherchée. Avec ces paramètres optimaux, Adjacency offre des résultats plus probants à tous les niveaux. D'abord, au sein d'une région, le risque géographique apparaît plus homogène : les risques similaires sont mieux regroupés ensemble, et très rares sont les communes qui présentent un risque radicalement différent de celles autour d'elle. Les variances intra-zone sont en conséquence plus faibles, et la continuité entre les zones est aussi meilleure avec des paliers de risque (géographique) moyen inter-zones plus réguliers. Les résidus lissés avec Adjacency sont de surcroît plus prédictifs qu'avec la méthode Distance. L'effet géographique Adjacency est donc à la fois plus cohérent et plus prédictif.

On applique alors cette méthode optimale pour calculer l'effet géographique total de chaque poste, puis en combinant tous ces effets et en les classant par Ward, on obtient un zonier tous postes confondus. Ce zonier global est segmenté de manière cohérente avec les

zones propres aux postes, et il constitue ainsi une bonne synthèse du risque géographique santé. Toutefois, le besoin d'une structure tarifaire simple limite en pratique l'utilisation effective de plusieurs zoniers, un par poste, ce zonier global pourra donc être utilisé en remplacement. Pour finir, intégrer ce zonier tous postes dans la tarification apporte bien un gain significatif par rapport au zonier existant, car l'effet géographique réel, modélisé grâce à la méthode Adjacency, est mieux segmenté par les nouvelles zones. L'information géographique est ainsi mieux expliquée dans le modèle de tarification santé grâce à la zone globale, même lorsque l'environnement le permet on privilégiera toujours une tarification des postes combinés, où chaque prime pure de poste est expliquée par son zonier propre.

Mots-clés

Complémentaire Santé, Postes de Garantie, Modèles Linéaires Généralisés, Risque Géographique, Variables Externes, Lissage Spatial Bayésien, Lissage Spatial par la Théorie de la Crédibilité, Classification.

Geographical risk modelisation in health for the creation of a new zonage. Comparison of two methods of spatial smoothing

Abstract

The present work carries on updating spatial risk zoning ; it aims to both describe and compare two methods of geographical risk modelization, thanks to smoothing applied on the spatial residuals of a pure premium generalized linear model setted on that purpose. Firstly total risk is splitted by posts, so that pure premium can be estimated on an homogeneous risk distribution perimeter, in order to optimize the fitting to theoretical distributions. Developped posts are Hospice, Currents Cares, Optics, Pharmacy and Dentals. Each model starts by integrating classical pricing factors, present zoning excepted, then adds external significant factors to explain a part of geographical risk, as a temp partial zoning. Next step is to project residuals on a map where they are smoothed by the two methods, so that the missing spatial effect can be extracted. Finally this effect is combined to external effect, resulting in a total geographical effect, then grouped to form the zoning. Principals of this process are to cleary divide pricing on one hand and geographical effects on the other, that's why Alsace-Moselle effect is neutralized for Regime factor, this class being purely defined on geographical terms. Its effect then mainly spread on external factors as expected. After external impacts estimated by pure premium model, major step is applying the two smoothing methods to this present model.

The first method Adjacency is stochastic and simulates the residuals of a commune thanks to its neighbourough ones, and uses bayesian approach to make a prior hypothesis of attracting similar risks together. The second method is based on credibility theory and is estimating risk one shot by credibilising it with proximates risks, allowing a rising powerful influence to nearest communes. An optimal smoothing is calculated for each method and match with the more predictive residuals. Setting these parameters, Adjacency gives better results at all levels. Firstly, geographical risk areas are more homogeneous, similar risks being grouped together without extreme values. Variations in each zone are therefore reducted, and geographical effect is more harmonic in respect to increasing risk zones. Smoothed residuals are also more predictive than the Distance ones. By this way, Adjacency spatial effect is both the most consistent and most predictive of the two methods.

This optimal method is applied to calculate total spatial effect of each post, and then by combining and grouping them to obtain an all posts zoning. This global zoning is also consistent with the zonings attached at each post, synthetising as well global health spatial effect. In cases where dedicated zonings cannot be used, this global zoning can be used in replacement. Integrating this all posts zoning brings significant information to the model facing to the working zones, the spatial effect being better segmented by the new zones. Geographic informations are well explained in health pricing model through this new zoning, even if the approach combining posts pure premiums must always be choosen

as soon as it is possible.

Keywords

Health Insurance, Generalized Linear Model, Geographical Risk, External Variables, Spatial Bayesian Smoothing, Smoothing Distance Credibility Weighted, Clustering.

Synthèse

Le présent mémoire porte sur la révision du zonage en santé ; il a pour objectif de décrire et comparer deux méthodes qui visent à modèliser le risque géographique, grâce à un lissage effectué sur les résidus spatiaux d'un modèle linéaire généralisé de prime pure préparé dans cette optique.

En assurance santé, le risque d'un assuré est usuellement estimé à partir d'un nombre très restreint de variables. La plupart de ces variables, telles que l'âge, le régime et le nombre de bénéficiaires, sont fixes au sens où leurs modalités sont déjà déterminées à l'avance. Dans la définition des segments de risques pris en compte dans le tarif, le zonier qui entre dans cette structure tarifaire est une composante essentielle, sa définition oblige à un travail technique plus approfondi que les autres variables, du fait que leurs modalités sont fixes et déjà déterminées à l'avance. La variable zonier exprime le risque géographique, sa performance est entièrement déterminée par l'estimation de ce risque et des choix de classes effectués. À cet égard, ce mémoire s'intéresse à deux méthodes qui permettent d'identifier le risque géographique en s'appuyant sur des théories différentes. Son objectif est de comparer leur pouvoir prédictif et leur cohérence dans la segmentation. Ces méthodes s'appliquent sur les résidus d'un modèle de prime pure adapté à cet usage. Une partie de l'effet géographique est d'abord capturée par des variables externes ajoutées au modèle des variables tarifaires (zonier actuel exclu). Elles lissent les résidus projetés sur une carte des communes, pour en extraire l'effet géographique résiduel. Cet effet complète alors l'effet des variables externes pour constituer l'effet géographique réel recherché. Le zonier est finalement obtenu par un classement via la méthode de Ward, puis réintroduit à la place des variables externes dans la modélisation de prime pure.

La premiŕe partie de l'étude (composée des deux premiers chapitres du mémoire) traite le cadre assurantiel dans lequel le zonier s'inscrit. On y présente le fonctionnement de l'assurance santé dans ses aspects généraux et règlementaires ainsi que la structure des produits et du portefeuille socle de cette étude. Ensuite, on décrit le portefeuille étudié afin de mieux interpréter les effets estimés par la suite. Cette connaissance est également nécessaire pour la mise en place de la base de données apte à une modélisation de prime pure. La base a une structure « en changement de risque » : pour une année d'exposition au risque, si un bénéficiaire change de risque, alors une ligne est créée dans la base pour matérialiser ce nouveau profil de risque avec leur sinistralité correspondante. Pour arriver à une base où toutes les variables sont correctement formatées en vue de la modélisation, la création d'une variable formule synthétisant les multiples cas de formules [1] est par exemple nécessaire. C'est le travail de définition et de segmentation des variables, dont une partie se fait en amont de la modélisation. En supplément des variables « classiques »présentes

1. La formule représente le niveau de garantie ou bien de remboursement souscrit par l'assuré dans un contrat de complémentaire santé

ou fabriquées à partir du portefeuille, un travail exploratoire est fait pour aboutir à une liste de variables externes (source INSEE, DREES,...) utiles pour expliquer le risque santé. Au terme de ce travail préparatoire, les variables tarifaires éligibles à la modélisation sont l'année d'observation, l'âge de l'assuré, le régime, la formule de garanties et le nombre de bénéficiaires. Plusieurs variables externes ont par ailleurs été retenues : l'indice de vieillissement, taux de médecins généralistes et spécialites, l'espérance de vie à la naissance et après 65 ans par sexe, taux de pauvreté, entre autres. Une analyse de ces variables met en évidence diverses corrélations entre elles, nous avons tenté de synthétiser l'information et éliminer les corrélations existantes à partir d'une Analyse en composantes principales (ACP), mais les composantes n'ont finalement pas été retenues, car la quantité d'information perdue était trop importante. L'intérêt des variables externes est en effet d'avoir une information sur les facteurs globaux qui expliquent tel ou tel poste de garanties, c'est pour cela que l'on souhaite garder au maximum le sens rattaché à chaque variable pour s'enrichir d'une meilleure connaissance du risque. Enfin, des regroupements sont faits sur la plupart des variables externes en utilisant la méthode des K-means. La démarche est de modéliser le risque assuré en excluant le zonier actuel, calculer le nouveau zonier par commune, puis l'ajouter au modèle de base pour compléter l'explication du risque.

Il existe une multitude de méthodes pour la création d'un zonier, dans ce mémoire on a choisi d'isoler l'effet géographique contenu dans les résidus du modèle GLM et l'information donnée par les variables externes pour l'obtention du zonier final. On considère l'hypothèse selon laquelle malgré un ajout des variables externes il existera toujours des effets liés à la zone géographique du risque que le modèle GLM est incapable d'expliquer.

La deuxième partie de l'étude, correspondant au chapitre trois, s'attache à la modélisation de la prime pure, basée sur la théorie des modèles linéaires généralisés. Dans la base de modélisation, les caractéristiques constituant le profil de risque sont jugées responsables de sa sinistralité survenue. Le but de la modélisation souhaitée est de valoriser l'impact de chacune des caractéristiques à l'intérieur du profil. Elle utilise pour cela les différences de prime pure dans la base et isole l'effet marginal apporté en moyenne par la caractéristique. On considère une structure tarifaire classique, où les effets des modalités peuvent être isolés puis multipliés entre eux pour estimer un effet moyen approximant au mieux la sinistralité moyenne observée dans la base d'étude. Deux axes de modélisation sont décidés : d'une part, distinguer les modélisations par poste, de manière à avoir des périmètres homogènes, optimisés pour une meilleure adéquation aux lois théoriques que les distributions de risque suivent par hypothèse ; d'autre part, la corrélation entre la fréquence et le coût moyen se révèle presque toujours trop importante, et la prime pure est donc modélisée directement en l'ajustant à une loi Tweedie sur tous les postes sauf Soins courants. La sélection des variables explicatives se fait graphiquement en considérant la cohérence de l'effet marginal et l'effet moyen observé, et statistiquement avec les résultats du test de significativité du χ^2, ainsi que la déviance, les AIC et BIC générés par le modèle à chaque étape. En outre, la sélection s'opère en deux temps : sélection Backward sur les variables tarifaires, peu nombreuses ; puis sélection Forward sur les variables externes, plus nombreuses et amenant aussi plus de corrélation donc faisant potentiellement diverger l'algorithme d'estimation des coefficients. Les variables tarifaires significatives sont l'âge du bénéficiaire, le régime de prestations, la formule de garanties et le nombre de bénéficiaires. Au niveau des variables externes, elles sont différentes selon le poste modélisé, celles qui ressortent significatives dans le modèle hospitalisation sont le taux de retraite, l'espérance de vie à la naissance d'une femme, le taux de population entre 45 et 59 ans , l'indice de vieillissement, le taux de cohabitation des personnes de plus de 75 ans et la densité de pharmacies. En vue de l'esti-

mation de l'effet géographique, la modélisation a été ajustée dans l'intention de rendre le plus indépendantes possible les variables tarifaires (hors zonier actuel) qui représentent la partie « risque propre à l'assuré »des variables externes qui représentent la partie « risque géographique ». Pour ce faire, l'impact de la modalité Alsace-Moselle de la variable Régime a été neutralisé (au niveau de cette variable), par le biais d'un offset sur ses coefficients. Ce régime spécifique, qui autorise des remboursements sur des bases plus hautes que celles de la Sécurité Sociale, découle entièrement de critères géographiques. Son impact doit donc apparaître soit au niveau des variables externes, soit au niveau des résidus qui seront lissés par la suite. Lorsqu'on effectue le forçage, on observe bien un transfert de l'effet vers les variables externes, les coefficients des variables tarifaires ne variant quasiment pas. La modalité s'en retrouve alors en moyenne sous-estimée, les facteurs externes ont pris sur eux une partie de l'effet mais il manque globalement de l'information sur cette modalité, et cette différence d'estimation se retrouve alors dans les résidus.

La troisième partie, correspondant au quatrième chapitre, constitue le cœur du mémoire. Elle a comme objectif la création et l'insertion du zonier dans la structure du risque.

$$\textit{Risque} = \textit{Effets non géographiques} + \textit{nouveau Zonier} + \textit{epsilon.}$$

Les effets non géographiques correspondent aux variables tarifaires, zonier actuel exclu, et epsilon à l'erreur finale de la modélisation, sans tendance (le bruit blanc).

La méthodologie de construction du zonier se résume par les étapes suivantes.

Étape (a.)

Dans la deuxième partie, nous avons modélisé par la méthode de GLM les effets des variables connues, c'est-à-dire les variables tarifaires et les variables externes sélectionnées. La structure du risque à la fin de cette étape est,

$$\textit{Risque} = \textit{Effets connus} + \textit{résidus}_1,$$

où,

- Risque désigne soit la prime pure, soit la fréquence, soit le coût moyen.
- Les effets connus contiennent les effets non géographiques (via les variables tarifaires hors zonier actuel) et une partie de l'effet géographique externe (via les variables externes).
- Les résidus contiennent la partie inconnue de l'effet géographique, et l'erreur de modélisation.

Étape (b.)

L'objectif de cette étape est de lisser les $\textit{résidus}_1$ projetés sur la carte de France, de manière à ce que l'erreur finale du modèle se rapproche autant que possible d'un bruit blanc. Le résultat de ce lissage est l'information géographique systématique non expliquée par le modèle de prime pure à la fin de la partie précédente. L'effet géographique résiduel est ainsi créé tel que :

- Avant lissage :

$$\textit{résidus}_1 \;=\; \textit{effet géographique inconnu} + \textit{erreur de modélisation 1.}$$

- Après lissage :

$$\textit{résidus}_1 \;=\; \textit{effet géographique résiduel} + \textit{bruit blanc.}$$

Dans cette dernière égalité,

— L'*effet géographique résiduel* s'obtient en lissant les $résidus_1$.

— Le *bruit blanc*, erreur qui ne contient plus aucune tendance géographique.

Le principal apport de ce mémoire se situe à ce niveau, l'on va appliquer et confronter deux méthodes différentes de lissage, afin de nous orienter vers la méthodologie la plus adaptée pour la réussite de notre zonier en santé.

La première méthode est une méthode stochastique nommée, *Adjacency*, basée sur l'approche bayésienne, où le lissage est local et prend en compte le risque des communes immédiatement voisines, c'est-à-dire adjacentes. La force du lissage est donnée par le nombre de simulations effectuées. Nous comparerons les résultats obtenus avec une méthode déterministe nommée *Distance*, qui utilise la théorie de la crédibilité. Cette deuxième méthode lisse de manière globale en tenant compte des risques de l'ensemble des communes de France, avec une influence décroissante en fonction de la distance à la commune voisine. Le paramètre de lissage sur lequel il faut travailler correspond à celle de la distance, qui règle l'importance de l'influence accordée à la commune alentour.

Le niveau de lissage est par ailleurs déterminé pour être optimal, c'est-à-dire de manière à extraire la tendance systématique, l'effet qui se reproduit dans le temps à l'intérieur des résidus. En lissant trop peu, on garde une hétérogénéité dans zones qui est seulement due à du bruit d'échantillon. En lissant trop fort, on détruit la segmentation géographique en égalisant les résidus entre eux de manière abusive, on perd de la précision. Le point optimal est celui où les résidus lissés constituent le facteur géographique d'explication le plus prédictif. Il s'obtient en minimisant les SSE, indicateur qui mesure l'erreur quadratique entre les résidus lissés sur une partie de la base vs les résidus originaux de l'autre partie de la base. On retient ainsi le lissage renvoyant l'effet moyen qui arrive à reproduire le moins d'erreur sur l'échantillon test constitué par la deuxième partie de la base. La comparaison des deux méthodes se fait une fois leur paramètre optimal déterminé pour chacune, et le niveau de lissage correspondant appliqué sur les résidus. La méthode Adjacency offre des résultats plus probants à tous les niveaux. D'abord, au sein d'une région le risque géographique apparaît plus homogène : les risques similaires sont mieux regroupés ensemble, et très rares sont les communes qui présentent un risque radicalement différent des voisins. A contrario, à cause de ses sources déterministes, la méthode Distance n'arrive pas à être assez précise sur les communes qui divergent ponctuellement, elle conserve quelques extrêmes qu'elle n'arrive à supprimer qu'au prix d'un trop gros sacrifice d'information sur le risque moyen environnant. En conséquence les variances intra-zone sont plus faibles avec Adjacency qu'avec Distance, et la continuité entre les zones Adjacency est aussi meilleure, avec des paliers de risque (géographique) moyens inter-zones plus réguliers. Les résidus lissés avec Adjacency sont de surcroît plus prédictifs qu'avec la méthode Distance. L'avantage de la méthode Adjacency est par ailleurs encore visible directement dans l'effet de la zone, : l'effet réel en amont étant plus continu (les valeurs se prolongeant davantage les unes vers les autres car il y a moins de sauts de résidus), l'effet de la zone Adjacency est plus linéaire que celui de la zone Distance. L'effet géographique Adjacency se révèle à la fois plus cohérent et donnant de meilleures prédictions, il est donc considéré comme l'effet géographique réel recherché.

La modélisation devient :

$$\begin{aligned} Risque = {} & Effets\ non\ géographiques \\ & + \ effets\ géographiques\ externes + effets\ géographiques\ résiduels \\ & + \ bruitblanc. \end{aligned} \tag{1}$$

Étape (c.)
À cette étape nous construisons le zonier à partir des effets uniquement géographiques. En combinant l'effet lissé des résidus à l'effet déjà estimé des facteurs externes, on obtient l'effet géographique total duquel on déduit le nouveau zonier, selon une méthode finale de classification choisie. L'effet géographique total est égal aux effets géographiques externes auxquels sont ajoutés les effets géographiques résiduels. En lui appliquant une classification hiérarchique par la méthode de *Ward*, on obtient le nouveau zonier,

$$Zonier = effet\ géographique\ regroupé\ en\ classes\ .$$

Le zonier est une synthèse des effets géographiques connus à l'étape (a.) (impact des variables externes) et dans les *résidus lissés*$_1$ à l'étape (b.).

Étape (d.)
Cette dernière étape à pour objectif d'intégrer le zonier dans le modèle. Le nouveau zonier est introduit dans le modèle de prime pure en échangeant les effets externes partiels contre ceux totaux synthétisés par l'effet de la zone, le zonier devient ainsi le facteur explicatif unique du risque géographique intégré dans la structure tarifaire. Remarquons que l'Alsace-Moselle est maintenant appartenant d'une zone cohérente avec son risque géographique réel, son coefficient dans le modèle de prime pure peut donc être calculé normalement (sans offset au niveau de la variable régime) pour améliorer l'estimation sur ce régime spécifique. Le risque s'explique finalement par,

$$Risque = Effets\ non\ géographiques + nouveau\ Zonier + epsilon'.$$

Quand les facteurs externes sont échangés contre le nouveau zonier, la différence d'information est seulement égale à l'effet géographique résiduel, et non à l'effet géographique total, ce qui aurait été le cas si les variables externes n'avaient pas été introduites en amont. Les coefficients des variables tarifaires sont alors moins perturbés par l'arrivée du zonier : les coefficients tarifaires à l'étape (d.) sont très proches de ceux à l'étape (b.), excepté pour la modalité Alsace-Moselle, dont l'offset à l'étape (a.) sert seulement à construire une zone correcte pour cette région. Néanmoins son impact décorrélé de l'effet géographique (maintenant présent dans la zone) doit tout de même être calculé correctement, c'est pourquoi on enlève l'offset dans l'estimation finale. Grâce aux variables externes, les paramètres tarifaires estimés sont ainsi plus stables sur l'ensemble du processus.

Ce travail est présenté plus en détail pour le poste Hospitalisation, mais il a été réalisé en parallèle sur tous les autres postes, permettant ainsi la construction d'un zonier représentant le risque géographique des postes Hospitalisation, Soins Courants, Pharmacie, Dentaire et Optique. La combinaison de ces zoniers a finalement permis d'obtenir un zonier global synthétisant la totalité de la consommation santé.

Voici la représentation de l'actuel zonier versus le nouveau zonier :

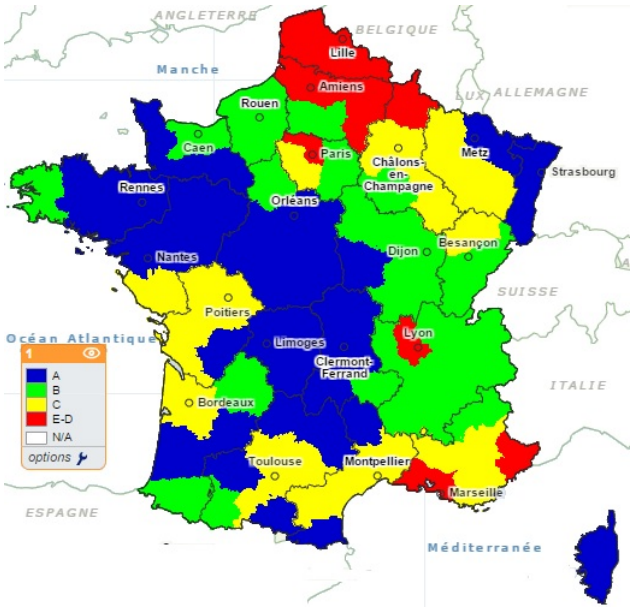


FIGURE 1 – Carte de l'Actuel Zonier

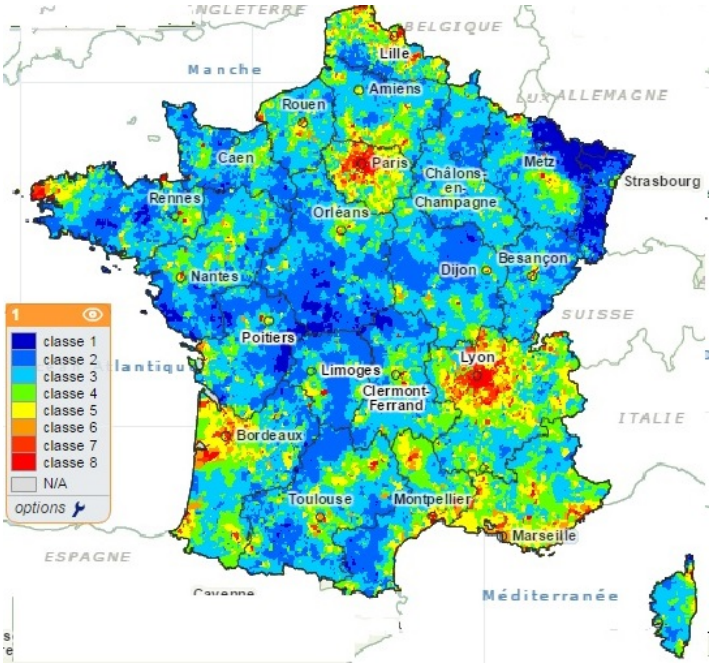


FIGURE 2 – Carte du nouveau Zonier

Grâce à son niveau de détail communal et à la méthode Adjacency appliquée sur une sinistralité plus récente, le nouveau zonier se révèle plus performant que le zonier actuel. Les risques sont ciblées de manière plus précise, avec une segmentation du risque géographique réel plus graduelle entre les zones et plus homogène à l'intérieur de ses zones. En insérant ce nouveau zonier dans le modèle de prime pure, cette meilleure estimation du risque géographique entraîne en définitive une meilleure performance du modèle de risque de santé global.

On a vérifié que ce zonier global est bien représentatif du risque géographique à l'intérieur de chaque poste, en comparant son impact à celui du zonier propre au poste. Toutefois, le besoin d'une structure tarifaire simple demandée dans l'entreprise limite en pratique l'utilisation effective d'un zonier par poste, et mène plutôt à une estimation directe de la consommation moyenne tous postes confondus.

Synthesis

The present dissertation carries on updating spatial risk zoning in health insurance ; it aims to both describe and compare two methods of geographical risk modelling, using a smoothing applied on the spatial residuals of a pure premium generalized linear model designed for this purpose.

In health insurance, risk for each policyholder is usually estimated by a limited number of variables. Most of these variables such as the age, plan and number of beneficiaries are fixed, in the way their modalities are determined in advance. When defining risk segments, the zoning which is taken into account in the pricing structure, is an essential component. Its definition needs a more thorough technical work than other variables because their modalities are fixed and already determined in advance. The variable "zoning" represents the geographical risk ; furthermore, its performance is fully determined by the estimation of this risk and the selection of appropriate classes. In this regard, this dissertation deals with two methods, relying upon two different theories, to determine the geographical risk. The purpose is to compare their predictive power and their consistency with segmentation. These methods are applied on the residuals of a pure premium model calibrated for this purpose. A part of the geographic effect is firstly captured by external factors added to the pricing model (excluding the initial zoning). Residuals from both methods are then projected on a map showing the municipalities, in order to extract residual geographic effect. The zoning is finally obtained by a classification within the Ward's method, for later being reintroduced in the pure premium model in replacement of external variables.

The first part of the study (Chapters 1-2) presents the main aspects of health insurance. We outline the health insurance operational principle in both general and regulatory aspects for later describe the products and portfolio, in order to give a global idea of the study's framework and better understand the estimated effects. This first step is necessary for implementing the database used in modelling the pure premium. The database has a "changing risk" structure : for a one-year exposure to risk, if a policyholder's risk profile changes, then a new line is created in order to take into account this risk profile change. To establish a base in which every variable is well formatted, the creation of a variable "package" which synthetize all possible coverage options is necessary. In addition to classic variables of the portfolio, exploratory works were carried out to determine a list of external variables (source INSEE, DREES, etc.) useful for explaining health risks. From this exploratory work, the pricing variables qualifying for the modelling are : policyholder's age, insurance regime, coverage options and number of beneficiaries. Amongst external variables, a few have been retained, such as : ageing rate, rate of general/specialist doctors, life expectancy at birth and after 65 by gender, poverty rate. An analysis of these variables shows several correlations between them, and a PCA (Principal Component Analysis) is used to synthetize the information and remove the correlations between variables. But this

approach is finally too costly in terms of qualitative information loss compared to correlation improvements. Indeed, external variables explain how global factors influence each risk category, that's why we prefer to keep the information attached to each variable. Finally, we carry out clustering on most of these external variables by using the K-means method. Thus the scheme is to model insured risk by excluding the initial zoning, then to calculate the new zoning by commune, and adding it to the original pure premium model to complete risk explanation. A zoning can be created by many different ways. In this report, to build the final zoning, we chose to isolate the geographic effect contained into the residuals of the pure premium GLM and the information given by external factors. We assume that even if external factors are added, effects linked to geographic zone unexplained by the GLM will still remain.

The second part of the study (Chapter 3) concerns the modelling of the pure premium using the theory of generalized linear model (GLM). In the data base, characteristics are assumed to be responsible for the occurred claims. The modelling aim is to isolate the individual impact of each characteristic on claim risk. The goal is to give a classic pricing structure, where characteristic effects can be isolated and multiplied together to approximate at best the observed average claim. Two modelling axis are decided : on one hand, split models by categories, allowing homogeneous perimeters where the fitting to theoretical distribution is better for each one. On the other hand, correlation between frequency and severity appears most of the time too large, guiding to use Tweedie models on every category except Current Cares. The explaining variables are selected graphically by considering the consistency between marginal and average effects, while statistics indicators such as χ^2 test, Deviance or AIC or BIC help to do the final cut. The selection process is done in two ordered steps : Backward selection on the few pricing variables ; then Forward selection on external variables, their high number can lead to correlation issues that may make diverge fitting algorithm. Major pricing variables are the policyholder's age, the insurance regime, the warranties packages and the number of beneficiaries. Significant external factors vary among the modelled categories, the selected ones for the hospitalization model are retirement rate, life expectancy at birth for women, population between 45 and 59 rate, ageing rate, cohabitation rate for over 75, and pharmacy rate. In the perspective of geographic effect estimation, model adjustments must be done in order to distinct more clearly two aspects : pricing variables that represent the « intrinsic policyholder risk », and external variables that represent the « geographic risk ». For that reason, Alsace-Moselle's impact is neutralized from the regime variable, by an offset on Alsace-Moselle modality. This specific regime, which provides higher reimbursements, is indeed only defined on geographic criteria. By forcing it, pricing variables effect remain the same, and the effect is transferred either to external factors or to residuals that will be smoothed later.

The third part (chapter 4) is the heart of the study, it presents the last steps to complete the zoning analysis, its objective is the creation of the zoning and its integration in the risk structure :

$$\textit{Risk} = \textit{non-geographical effects} + \textit{Zoning} + \textit{epsilon}.$$

Where non-geographical effects are the pricing variables, current zoning excluded, and epsilon is the model's error without spatial patterns. The creation of the zoning is divided into four steps.

Step a.)

At the second part of the study we have modelled the effect of external and pricing variables using the theory of generalized linear model (GLM). The risk structure at the end of this step can be summarized as :

$$Risk = Known\ effects + Residuals,$$

with,

- *Risk* : Pure premium or Frequency or Severity
- *Known effects* : pricing effect (policyholder's own risk) with offset on Alsace-Moselle + External effect (partial geographic risk)
- *Residuals* : Model's error, where researched systematic spatial effect remains.

Step b.)

At this step we smooth the residuals after projecting them on a map of France municipalities. The outcome of this smoothing corresponds to the spatial information unexplained by the previous pure premium model. The residual geographic effect is thus created :

Before smoothing we have,

$$Residuals = Unknown\ residual\ geographic\ effect + Model's\ error,$$

And after smoothing,

$$Residuals = Residual\ geographic\ effect + White\ noise,$$

with,

- *Residual geographic effect* : Obtained by smoothing Residuals with two different methods.
- *White noise* : Error without spatial patterns.

The first smoothing method called Adjacency is stochastic and simulates the residuals of a municipality thanks to its neighboring one, it uses a Bayesian approach to make a prior hypothesis of attracting similar risks together. Smoothing strength is given by the number of simulations. The second one called Distance is determinist, based on Credibility theory. It estimates the risk in one shot by credibilising it with proximate risks, allowing a rising influence to nearest municipalities. The smoothing parameter is the distance one, which fix influence allowed to surrounding municipalities.

Smoothing level is also set to be optimal, to extract the systematic trend, i.e. the spatial pattern contained by residuals. A too weak smoothing will keep heterogeneous effects that won't match with the average reality, whereas an abusive smoothing will give a lower precision. The optimal point is where smoothed residuals appear to be the most predictive geographical explanative factor. It is provided by the SSE minimization, this indicator being a measure of the quadratic error made between smoothed residuals of a large part of the database, compared to raw residuals issued from a smaller part of the data base. Therefore, the effect that keeps coming back through random data is retained. Then, the two methods are compared with parameters set to be optimal and optimal smoothing applied on residuals. Under these conditions, Adjacency gives more satisfying results. Firstly, geographical risk is more homogeneous within a zone : similar risks are better grouped together, and are less sensitive to extreme variations. In contrast, as a result of its determinist characteristic, Distance method is not precise enough and still

contains zones with local variations not consistent with the dominant risk. And taking off these zones causes the removal of a large amount of information. The adjacency zones have therefore a lower variance, and they present more continuous risk variations. Moreover, Adjacency smoothed residuals are both more consistent and more predictive, it is also considered as the method giving the spatial risk closest to reality.

The model becomes :

$$\begin{aligned} \textit{Risk} = \textit{Pricing effect (non-geographic)} \\ + \textit{External effect (geographic)} + \textit{Residuals geographic effect} \\ + \textit{White noise.} \end{aligned}$$

Step c.)

Combining the smoothed residual effect to external effects we obtain a total geographical effect from which zoning is created by the Ward classification method.

Thus, the new zoning becomes :

$$\textit{Zoning} = \textit{Ward class of Total geographic effect.}$$

Step d.)

In this last step, new zones are introduced in the pure premium models by replacing the partial external effects with the total external effects, synthetized by the zoning effect. Alsace-Moselle now being affected to a well-adapted zone, its pure premium coefficient can be reset as normal (without offset on Regime's modality) to improve the precision of the estimation of this specific regime.

Thus, final pure premium model is :

$$\textit{Risk} = \textit{Final pricing effect (non-geographic)} + \textit{New Zoning} + \textit{Final white noise.}$$

When External factors are replaced with new zoning, the difference of information is only equal to residual geographic effect, and not to Total geographic effect. This would be the case if there were no external factors at the beginning. Pricing effects are therefore less disturbed by introducing the zoning into the pure premium model : Final pricing effect at step d.) is very close to pricing effect at step b.), except for the Alsace-Moselle modality, offset at step a.) only to build its correct geographic risk in the zoning. Once the geographical risk is integrated in its zone, its real impact on pure premium still have to be calculated normally, that's why its offset is cancelled in the final pure premium estimation. Thanks to external factors, estimated parameters are thus more stable during the whole process.

Detailed results are only presented for the Hospitalization category, but the same work has been done on every other category, in order to build a final zoning that takes all the categories into account.

Representation of the actual versus new zoning :

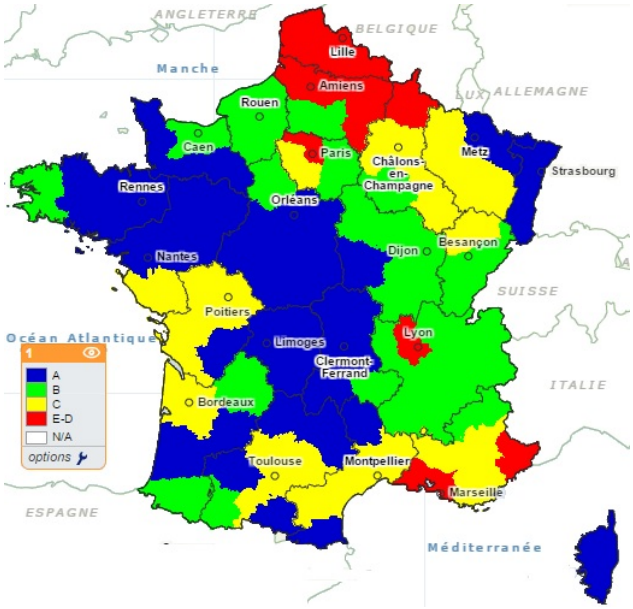


FIGURE 3 – Old Zoning's map

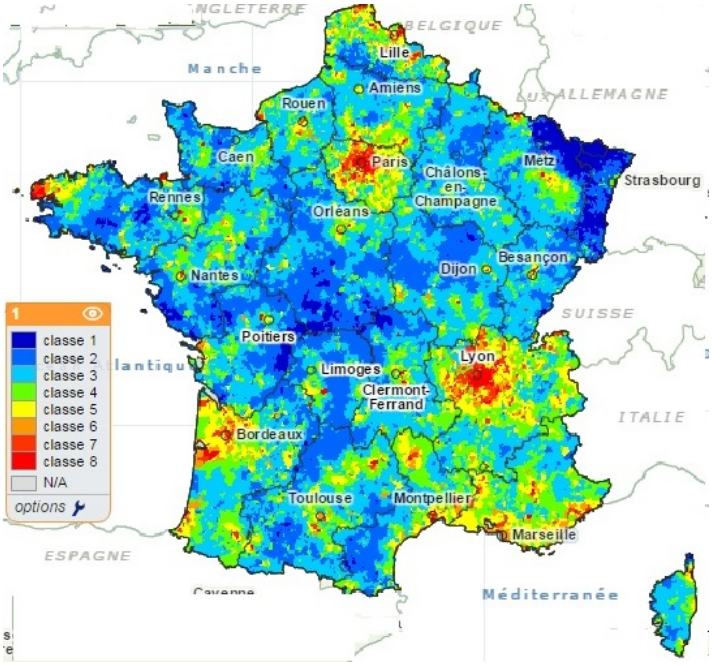


FIGURE 4 – New Zoning's map

Due to a better level of details the new zoning appears to be more efficient that the original one. In addition, the better estimation of the geographical risk, thanks to the insertion of the zoning, results in a better performance of our health risk model. This global zoning segmentation is consistent with each category's specific zoning. It is therefore a good synthesis of the health geographical risk. When dedicated zonings cannot be used, this global zoning is a good replacement.

Remerciements

Je tiens en premier lieu à remercier toute l'équipe de la Direction Technique du Marché de Particuliers de Generali pour son accueil tout au long de mon apprentissage. Je remercier chaleureusement ma tutrice en entreprise, Mlle Annabelle Bongo, pour ses nombreux conseils, sa disponiblité et pour m'avoir permis de bénéficier de son expérrience.

Je souhaiterais exprimer ma reconnaissance à M. Christophe Mouren, Mme Florence Peyres et Frabrice Staad pour m'avoir fait confiance en me permettant d'effectuer mon alternance au sein de l'équipe Prévoyance Santé.

Je tiens également à remercier Olivier Saldana, Ghislaine Guibert et Fatemeh Abdollahi pour le temps, l'écoute et leurs précieux conseils. En particulier, je remercie Remi Michel, son savoir-faire et technicité a été d'une aide fondamentale à l'écriture de ce mémoire.

J'adresse mes sincères remerciements à M. Mathieu Rosenbaum, mon tuteur pédagogique, pour son encadrement et la pertinence de ses conseils au cours de mon travail de mémoire. Je remercie également l'ensemble du corps éducatif de lÍSUP pour m'avoir accueillie tout au long de ma formation, ainsi pour leurs connaissances transmises.

Je souhaite finalement exprimer ma reconnaissance à ma famille et mes amies qui m'ont toujours encouragé. Merci profondément à ma mère, mon père et ma soeur d'être, malgrè la distance, toujours là. Et un grand merci à Nicolas Arancibia, qui m'a soutenue inconditionnellement tout au long de cet aventure.

Table des matières

Introduction

En assurance santé, le risque d'un assuré est usuellement estimé à partir d'un nombre très restreint de variables. La plupart de ces variables, telles que l'âge, le régime et ou le nombre de bénéficiaires, sont fixes au sens où leurs modalités sont déjà déterminées à l'avance. Du point de vue de la définition des segments de risque pris en compte dans la tarification via ces variables, l'intervention et la plus-value de l'actuaire se borneront à effectuer des regroupements pertinents de modalités. Un travail plus approfondi peut en revanche être développé au niveau de la variable zonier qui exprime le risque géographique, et dont la performance est en très grande partie expliquée par le choix de classes effectués. À cet égard, ce mémoire s'intéresse à deux méthodes qui permettent d'identifier le risque géographique en s'appuyant sur des théories différentes. Son objectif est de comparer leur pouvoir prédictif et leur cohérence dans la segmentation. Ces méthodes s'appliquent sur les résidus d'un modèle de prime pure adapté à cet usage.

Les différentes étapes menant à la construction du zonier ont été réparties en quatre chapitres :

Le premier chapitre traite du cadre assurantiel dans lequel le zonier s'inscrit. On y présent le fonctionnement de l'assurance santé dans ses aspects généraux et règlementaires ainsi que la structure des produits et du portefeuille socle de cette étude. Le lecteur pourra ainsi acquérir une vision macro de l'environnement de travail.

Le deuxième chapitre porte sur l'analyse préparatoire des données en vue du modèle. On détaille la constitution de la base, ainsi que la sélection a priori des variables initialement présentes en portefeuille, variables tarifaires , puis complétées par des informations environnementales ou géographiques, variables externes. Une analyse du zonier existant sera également mise en œuvre, pour examiner sa pertinence pour faire face au risque actuel.

Dans un troisième chapitre nous introduisons la théorie des modèles linéaires généralisés (GLM), utilisée pour estimer le risque de l'assuré, sa prime pure. On y détaille le processus, des le type de modélisation choisie (prime pure directe ou décomposition en fréquence×coût moyen) jusqu'à la validation du modèle, en passant par les étapes majeures de sélection des variables explicatives et l'estimation des coefficients ajustés par des regroupements éventuels assurant la robustesse du modèle. Ce modèle sera également adapté en vue de son utilisation pour la construction du zonier, avec la mise en place de forçage sur la modalité Alsace-Moselle de la variable régime.

Le quatrième et dernier chapitre décrit les deux méthodes de modélisation du risque géographique. L'information non expliquée par le modèle de prime pure est reflétée par les résidus du modèle, et chaque méthode va s'attacher à capturer l'information géographique qui y subsiste. On présente d'abord les théories propres à chaque méthode, l'une nom-

mée Adjacency et se basant sur une approche bayésienne impliquant des simulations de variables aléatoires, l'autre nommée Distance, méthode déterministe basée sur la théorie de la crédibilité. On y détaille les outils indispensables au choix du lissage optimal, puis les méthodes seront comparées grâce à des indicateurs statistiques et graphiques permettant de juger de leur performance. La méthode optimale est ainsi choisie, on teste ensuite les hypothèses en amont d'ajout des variables externes et de forçage sur l'Alsace-Moselle. A terme, on utilisera cette méthode sur l'ensemble des postes principaux en santé, pour aboutir à un zonier propre à chaque poste, puis en combinant leurs effets, à un zonier représentatif du risque global. On s'assure que ce dernier reste significativement corrélé et représentatif du risque géographique, y compris à l'intérieur de chaque poste. Enfin, ce zonier global est comparé au zonier existant pour pouvoir mesurer l'apport du travail effectué.

Cette étude a été menée sur l'ensemble des postes principaux de santé ; Hospitalisation, Soins Courants, Pharmacie, Dentaire et Optique. Afin de fluidifier la lecture du mémoire et d'éviter des analyses redondantes, ainsi que la répétition des calculs, les résultats de l'étude sont principalement présentés sur le poste Hospitalisation, ceux des autres postes s'obtenant de manière analogue.

Chapitre 1

Présentation de l'assurance complémentaire santé en France

1.1 Assurance Complémentaire Santé

Trois types d'opérateurs se partagent le marché de la complémentaire santé en France : les mutuelles, les sociétés d'assurance et les institutions de prévoyance. Les mutuelles couvrent un peu plus de la moitié des personnes bénéficiant d'une couverture complémentaire santé.

1.1.1 La sécurité sociale

La Sécurité Sociale, est un système d'organismes destinés à protéger les individus résidant sur le territoire français des risques sociaux (maladie, maternité, incapacité de travail ou invalidité, vieillesse, décès, charges de famille, chômage, entre autres).

Cette protection s'exerce par l'affiliation des assurés sociaux et de leurs ayants-droits à l'un des régimes de la Sécurité Sociale. Cette affiliation est obligatoire et le régime de rattachement dépend de la situation professionnelle de l'assuré social.

La Sécurité Sociale regroupe trois régimes principaux : le régime général, le régime social des indépendants et le régime agricole ainsi que d'autres régimes spéciaux comme le régime Alsace-Moselle.

1.1.2 L'assurance complémentaire santé

Une assurance complémentaire santé est un contrat qui a pour objet de prendre en charge tout ou partie des dépenses de santé non couvertes par l'assurance maladie obligatoire. C'est en cela que le risque santé est également très sensible aux désengagements de la Sécurité sociale, et plus généralement au risque réglementaire.

La complémentaire santé se distingue de la prévoyance. En effet, la prévoyance couvre les conséquences financières d'un risque grave et peu fréquent, en l'occurrence le décès, alors que les frais médicaux correspondent généralement à des risques fréquents, avec coût moyen faible, mise à part une éventuelle hospitalisation.

1.1.3 Mécanisme de remboursement

L'assurance complémentaire santé intervient, tout d'abord, en complément des prestations du régime obligatoire, pour les frais de soins qui font l'objet d'une prise en charge par ce dernier. Ensuite, elle peut également proposer des prestations supplémentaires, pour des

actes de soins ou de prévention non pris en charge par le régime obligatoire : par example, forfaits médecine douce, cures thermales. Les prestations de la Sécurité Sociale sont assises sur une base de remboursement (tarif conventionnel, sans dépassement d'honoraires) à laquelle s'applique le taux de remboursement. Ces deux données dépendent de l'acte de soin considéré. La différence entre la base de remboursement et la somme remboursée par la Sécurité Sociale est appelée ticket modérateur. L'assurance complémentaire rembourse (tout ou partie) du ticket modérateur et des dépassements éventuels (différence entre tarif de l'acte de soin ou du bien médical avec la base de remboursement)

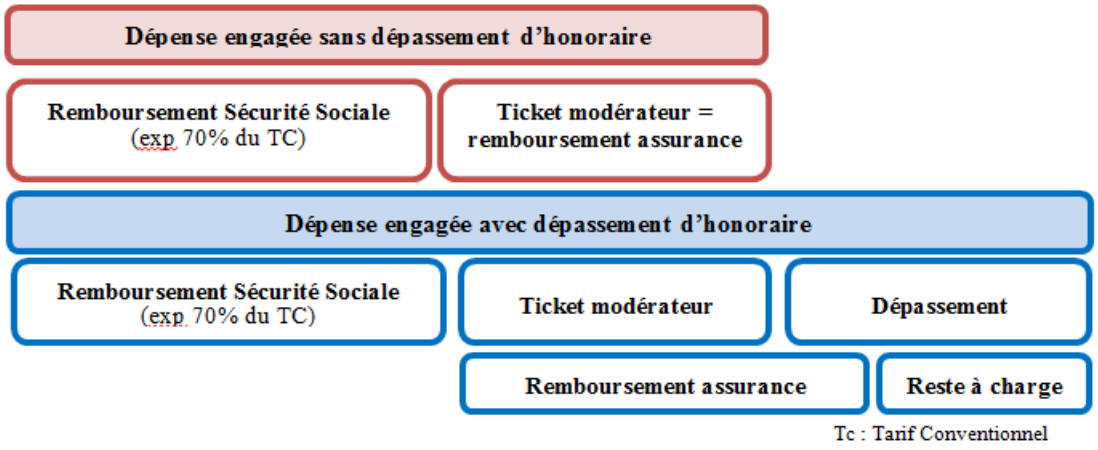


FIGURE 1.1 – Base de Remboursement

1.2 Actualités et réglementation en assurance santé

1.2.1 Comptes nationaux de la santé

En 2014, la consommation de soins et de biens médicaux (CSBM), c'est-à-dire la valeur totale des biens et services qui concourent au traitement d'une perturbation provisoire de l'état de santé, est évaluée à 190,6 milliards d'euros, soit près de 2 900 euros par habitant. La CSBM représente ainsi 8,9 pour cent du PIB en 2014 ; ce pourcentage croît légèrement depuis 2012. Sur la période 2006-2014, le rythme de croissance en valeur de la CSBM a connu un ralentissement sensible : il était de +4,0 pour cent en 2007, il s'est établi à +3,1 pour cent en 2008 et 2009, et reste inférieur à +3 pour cent depuis cinq ans. Il est néanmoins légèrement plus élevé en 2014 (+2,5 pour cent) qu'en 2013 et 2012. Cette dépense est financée à 76,6 pour cent par l'Assurance maladie. Cette part a crû de 0,7 point entre 2011 et 2014, en raison du dynamisme des dépenses qu'elle prend le plus en charge (soins hospitaliers, soins d'infirmiers, transports sanitaires, médicaments coûteux, en particulier rétrocession hospitalière) et de la progression continue du nombre d'assurés exonérés du ticket modérateur. À l'inverse, phénomène nouveau, la part des organismes complémentaires diminue légèrement, et représente 13,5 pour cent de la dépense en 2014. Le reste à charge des ménages s'établit en 2014 à 8,5 pour cent. Pour la troisième année consécutive, il baisse de 0,2 point, par contrecoup de l'évolution constatée sur les autres financeurs. Quant à la d 'epense courante de santé (DCS), c'est-à-dire la somme de toutes les dépenses « courantes »engagées par les financeurs publics et privés pour la fonction

santé, elle s'élève à 256,9 milliards d'euros en 2014, soit 12,0 pour cent du PIB. Cette dépense représente environ 3 900 euros par habitant.

1.2.2 L'Accord National Interprofessionnel (ANI)

L'Accord National Interprofessionnel (ANI) du 11 janvier 2013 s'insère dans le cadre des mesures prises pour un nouveau modèle économique et social au service de la compétitivité des entreprises et de la sécurisation de l'emploi et des parcours professionnels des salariés. Son objectif est de permettre aux salariés qui ne bénéficient pas encore d'une couverture collective à adhésion obligatoire en matière de remboursements complémentaires de frais de santé au niveau de leur entreprise d'accéder à une telle couverture.

Les effets attendus de l'ANI sont principalement :

- Un faible impact sur le taux de pénétration de la complémentaire santé : la part des non couverts devrait se réduire à 2,8 pour cent en 2016 contre 4 pour cent en 2013.
- Un transfert massif de l'individuel vers le collectif : l'assurance santé individuelle ne couvrirait plus que 43,3 pour cent des personnes contre 57,5 pour cent en 2013 soit une perte de 9 millions d'assurés représentant un manque à gagner de 4 à 5 milliards d'euros.

Ces chiffrages étant d'autant plus délicats qu'un transfert peut en cacher un autre. Les salariés ne quitteront pas tout seul l'univers de l'individuel : ils pourraient bien emmener avec eux le conjoint et les enfants.

En plus, elle concerne le maintien des garanties santé et prévoyance d'entreprise pour les salariés qui perdent leur emploi. La portabilité des droits est portée de neuf à douze mois à compter du 1er juin 2014 pour la complémentaire santé et au 1er juin 2015 pour la prévoyance (invalidité, incapacité, décès). Les secteurs non couverts par l'accord de généralisation de la couverture santé (agriculture, économie sociale et TNS) devront également mettre en place ce maintien selon le même calendrier.

Le panier de soins minimum à respecter

Les garanties minimales à mettre en place au 1er janvier 2016, à défaut de tout accord aux niveaux de la branche ou de l'entreprise devront être égales à :

— 100 pour cent de la base de remboursement des consultations, actes techniques et pharmacie.
— le forfait journalier hospitalier.
— 125 pour cent de la base de remboursement des prothèses dentaires.

Les conséquences sur le portefeuille actuel

Il existe 3 régimes principaux : Salarié, Agricole et TNS, donc il y aura une migration des contrats du régime salarié de la couverture individuelle vers la couverture collective. Or, l'accord doit s'appliquer obligatoirement à partir le 1er Janvier 2016, néanmoins le transfert des contrats salariés est progressif depuis 2013.

1.2.3 Les contrats solidaires et responsables

Une complémentaire santé est « solidaire »lorsque l'Assureur ne fixe pas les cotisations en fonction de l'état de santé des individus couverts et ne recueille aucune information médicale à l'adhésion. Depuis la loi n° 2004-810 du 13 août 2004 relative à l'assurance maladie, les contrats de complémentaire santé sont qualifiés de « responsables »dès lors qu'ils respectent les règles des contrats responsables définies par le code de la Sécurité sociale. Les organismes de complémentaire santé signalés comme responsables bénéficient d'avantages fiscaux et sociaux : Depuis le 1er janvier 2014, TSCA (Taxes spéciale sur les conventions d'assurances) de 7 pour cent appliquée aux contrats responsable contre 14 pour cent pour les contrats non responsables.

L'objectif du contrat responsable est donc d'inciter les assurés au respect du parcours de soins, via des règles prévoyant un ensemble d'interdictions et d'obligations de prise en charge.

Le contrat responsable rembourse à minima le ticket modérateur	Le contrat responsable ne rembourse pas
• Pour les consultations du médecin traitant dans le cadre du parcours de soins. • Pour les médicaments remboursés à 65 pour cent par la Sécurité sociale. • Pour les examens de biologie prescrits par le médecin traitant.	• Les dépassements et majorations liés au non respect du parcours de soins. • La participation forfaitaire de 1 euro applicable aux consultations et à certains examens médicaux. • Pour au moins 2 prestations de prévention fixées par la réglementation. • Les franchises appliquées aux médicaments, aux actes paramédicaux et aux frais de transport.

Table 1.1 – Contrat responsable

Synthèse des impacts du nouveau contrat responsable par poste

Tous postes cofondus

- Obligation de prise en charge du ticket modérateur pour l'ensemble des actes remboursés par l'Assurance maladie (sauf cures thermales et pharmacie prise en charge à 30 pour cent et à 15 pour cent)

Soins courants

Instauration d'une double limite sur la prise en charge des dépassements d'honoraires des médecins :

- 100 pour cent de la base de remboursement pour les médecins non adhérents au Contrat d'Accès aux Soins (CAS).
- Différentiel d'au moins 20 pour cent entre la prise en charge des dépassements d'honoraires en faveur des médecins adhérents au CAS vs. les médecins non adhérents au dispositif.

Optique

- Instauration de planchers et de plafonds de prise en charge sur les équipements optiques selon les types de verres.
- Maximum de remboursement fixé à 150 euros pour la monture.
- Remboursement d'un équipement par période de 2 ans, sauf pour les mineurs et en cas de changement de correction.

Hospitalisation

- Prise en charge du forfait journalier sans limitation de durée.

1.3 Les produits santé

1.3.1 Les produits et leurs réseaux de commercialisation

Les produits Santé à Generali sont commercialisés par trois réseaux :

i. Les Agents, qui ont des points de ventes fixes et commercialisent principalement des produits Generali.

ii. Les Courtiers, qui possèdent également des points de ventes fixes, mais commercialisent des produits de différents assureurs et placent leurs assurés en fonction du rapport qualité/prix.

iii. Les Salariés qui font des démarches à domicile pour la recherche des nouveaux assurés.

Les réseaux Agents et Courtiers ont dans leur portefeuille 11 produits santé en commun dont deux seulement sont encore en commercialisation, pour le reste, il s'agit des produits en «run-off». Le réseau Salariés a quant à lui 6 produits en portefeuille mais un seul est commercialisé, le reste est en « run-off ».

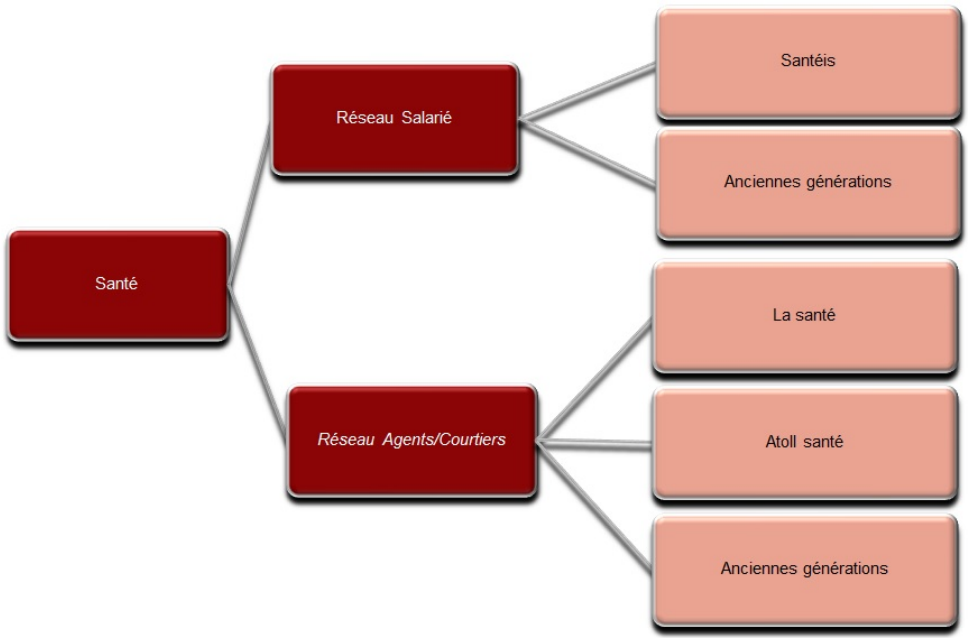


FIGURE 1.2 – Reseaux et Produits

Ces différents produits consistent au versement de prestations complémentaires à celles du régime obligatoire pour les frais médicaux ; frais liés à la l'hospitalisation ou à la chirurgie occasionnés par un accident, une maladie ou une maternité, par exemple.

Ces produits s'adressent à 4 types d'assurés, ceux qui sont affiliés a un des régimes suivants :

- Régime général (appelé ici le régime des salariés).
- Régime social des indépendants (régime des TNS - Travailleurs Non-Salariés).
- Régime agricole.
- Régime de l'Alsace et la Moselle : ces derniers bénéficient d'une couverture santé plus importante par leur régime obligatoire, la couverture complémentaire sera par conséquent plus faible, ils bénéficient de tarifs plus avantageux.

Les niveaux de remboursement proposés par les produits sont déterminés en fonction de la formule et des options choisies. Sur l'annexe 1 le lecteur pourra trouver les tableaux des formules de chaque produit.

1.3.2 Présentation des postes de garantie

En santé, il y a différents postes de garantie proposées par les produits. Seuls les postes que nous estimons, a priori, comme étant les plus discriminants en terme de niveau de garantie ont été sélectionnés, ce sont les suivantes :

Remboursement des soins courants :
Consultations et visites au médecins généralistes et spécialistes, les indemnités de déplacement des praticiens, les frais liés à l'imagerie et radiologie, les frais liés aux analyses et prélèvements ainsi que les frais pour les auxiliaires médicaux.

Pharmacie :
Médicaments et homéopathie prescrits par un médecin.

Hospitalisation :
Frais d'hospitalisation dans les établissements conventionnés et non conventionnés : honoraires des praticiens et auxiliaires médicaux, frais des salles d'opération, frais de séjour y compris le forfait hospitalier. Frais de transport du malade ou de l'accidenté, séjour en maison de convalescence ou de repos, hospitalisation à domicile.

Dentaire :
Frais de soins, prothèses dentaires, frais d'orthodontie, implantologie et prothèses dentaires non remboursées par le Régime obligatoire et figurant à la nomenclature générale des actes professionnels (NGAP).

Optique :
Forfait optique en complément du remboursement des verres correcteurs, des montures et des lentilles correctrices prises en charge par le Régime obligatoire et des lentilles correctrices prescrites médicalement mais non prises en charge par le Régime obligatoire, forfait complémentaire par œil pour la chirurgie réfractive sur prescription médicale.

1.3.3 Analyse préliminaires des données

Pour l'étude, la période d'observation des mouvements des assurés a été fixée entre début janvier 2011 et fin décembre 2014.

Ces analyses sont établies à partir de l'ensemble des prestations versées par l'assureur pendant le temps que le bénéficiaire a été exposé au risque. Pour avoir un maximum d'information, on a choisi de prendre en compte tous les sinistres survenus entre 2011 et 2014, qui ont été réglés jusqu'au 31/05/2015. Ce recul est suffisant car l'assurance santé complémentaire est un risque dont la charge finale d'un exercice est connue rapidement. En effet suite à une étude, on a démontré que chaque année à la fin mai, l'assureur connaît 95 pour cent du total des prestations correspondant à l'exercice passé.

Les études préliminaires sont importantes pour bien maîtriser les différentes caractéristiques de chaque variable.

Description des données :

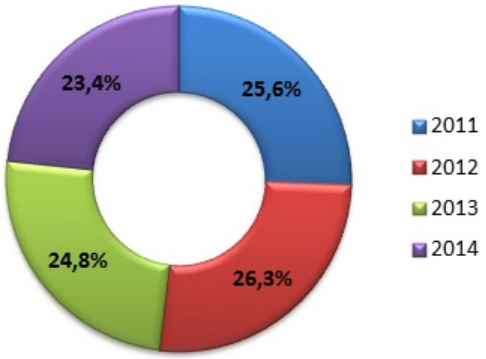


FIGURE 1.3 – Répartition des effectifs par année d'exercice

Les proportions pour chacune des années sont proches de 25 pour cent, il y a donc une stabilité dans le temps et un apport d'informations équilibré entre les quatre années.

Le tableau suivant montre la répartition des montants des prestations et du nombre de sinistres survenus par année. Pour analyser les impacts de la sinistralité, on a considéré l'année 2012 comme l'année de référence, car il porte la plus grande population. Pour la confidentialité des données, les chiffres sont présentées en base 1.

Exercise	Montants	Nb. Sinistres	Nb. Bénéficiaires
2011	0.96	0.98	0.97
2012	1	1	1
2013	0.91	0.92	0.94
2014	0.88	0.88	0.89

TABLE 1.2 – Comparaison des montants, nombre des sinistres et nombre des bénéficiaires par année

Pour chacune des années, les prestations ainsi que le nombre de sinistres sont fortement

corrélées à la taille de la population sous risque. La sinistralité moyenne est donc stable dans le temps.

Répartition des effectifs par âge et sexe

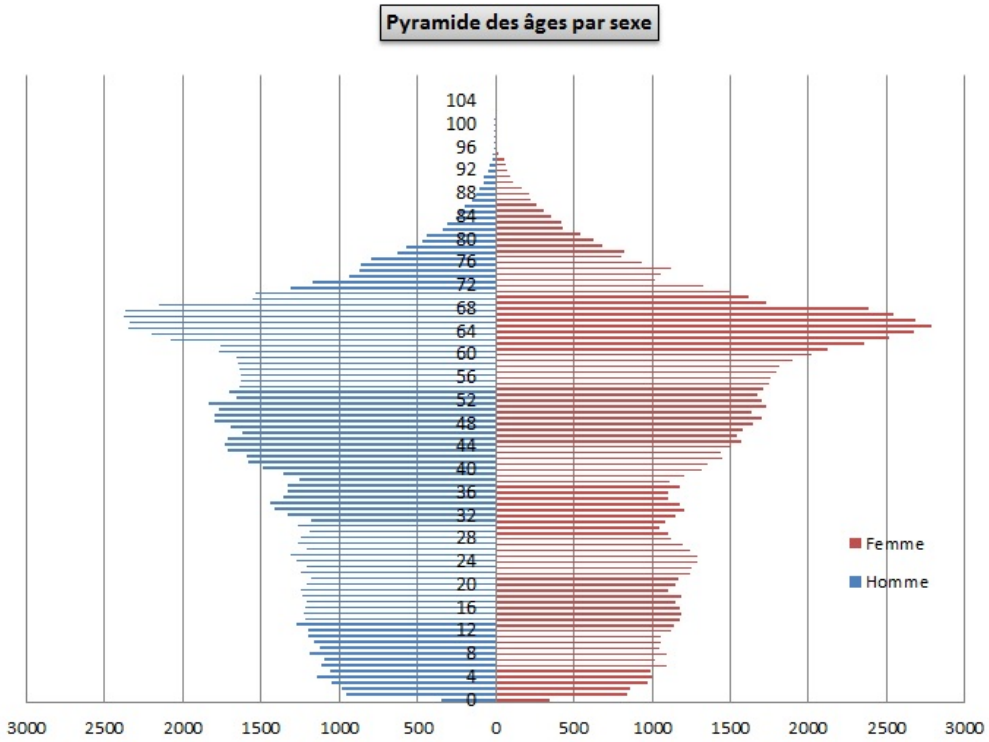


FIGURE 1.4 – Pyramides des âges

Les distribution des effectifs hommes et femmes sont très proches, bien qu'on observe que le nombre de femmes est légèrement supérieur. Les pics se trouvent entre les âges 58 et 70 ans. Il est important de remarquer pour la suite de notre travaux, que le nombre des assurés ayant 80 ans ou plus, est faible, ces âges sont donc peu significatifs pour notre étude.

Performance des réseaux

Réseaux	2011	2012	2013	2014
Agents	48.7%	49.3%	50.0%	49.5%
Courtiers	4.9%	6.3%	8.3%	10.6%
Salarié	46.3%	44.4%	41.8%	39.9%

TABLE 1.3 – Répartition des expositions par réseau

Le réseau le plus stable dans le temps est celui des Agents, 50 pour cent des assurés sont issus de ce réseau. Le réseau des courtiers est le moins important en exposition, mais la part des assurés issus de ce réseau progresse dans le temps. En parallèle, la part des assurés issus du réseau des Salariés est décroissante en fonction du temps, sa taille s'est

réduite de plus de 6 pour cent entre 2011 et 2014.

Rapport des prestations des sinistres sur les primes encaissées par réseau

Un indicateur important de la sinistralité est le ratio S/P, rapport entre les prestations des sinistres sur le total des primes encaissées par la compagnie d'assurance. Il s'agit d'un indicateur de référence utilisé par les assureurs pour trouver le meilleur équilibre tarifaire compte tenu de la sinistralité observée.

- Les Agents ont eu un S/P de 63 pour cent en 2014, (+0.4 pour cent par rapport à 2013).
- Les Courtiers ont eu un S/P de 69.9 pour cent en 2014, (+1.3 pour cent par rapport à 2013).
- Les Salaries ont eu un S/P de 49.2 pour cent en 2014, (-4.4 pour cent par rapport à 2013).

Si le S/P est inférieur à 100 pour cent, les prestations sont inférieures aux primes et l'activité est rentable ou au moins équilibrée. Si c'est le contraire, les prestations sont supérieures aux primes et l'activité est déficitaire.

Les autres indicateurs qui mesurent la performance des réseaux sont les suivants :

Taux de résiliation des affaire nouvelles

Défini par la formule,

$$\frac{\text{Nombre de résiliation des AN}}{\text{Nombre de An}}.$$

Une affaire nouvelle (AN) correspond à la souscription d'un nouveau contrat. Ce taux permet de connaitre la capacité d'un réseau à garder ses nouveaux clients.

Exercise	Réseau Agents/Courtiers	Réseau Salarié
2011	6%	18%
2012	8%	19%
2013	11%	18%
2014	15%	18%

TABLE 1.4 – Totaux de résiliation des affaires nouvelles

Les affaires nouvelles du réseau Salarié montrent un comportement moins fiable pendant leur première année que ceux du réseau Agents/Courtiers. Bien que le taux de résiliation des affaires nouvelles pour le réseau Salariés soit important, il est stable dans le temps. Au contraire, pour les réseaux agents et courtiers, le taux était relativement faible en 2011, il tend à augmenter tous les ans.

Taux de couverture

Définie par la formule,

$$\frac{\text{Nombre des affaires nouvelles}}{\text{Nombre total de résiliation}}.$$

Le taux de couverture est défini comme étant le rapport entre les contrats entrants et les contrats sortants pendant un mème année. Cet indicateur permet de bien mesurer la performance du réseau d'une année sur une autre, il est représenté sur le graphique suivant :

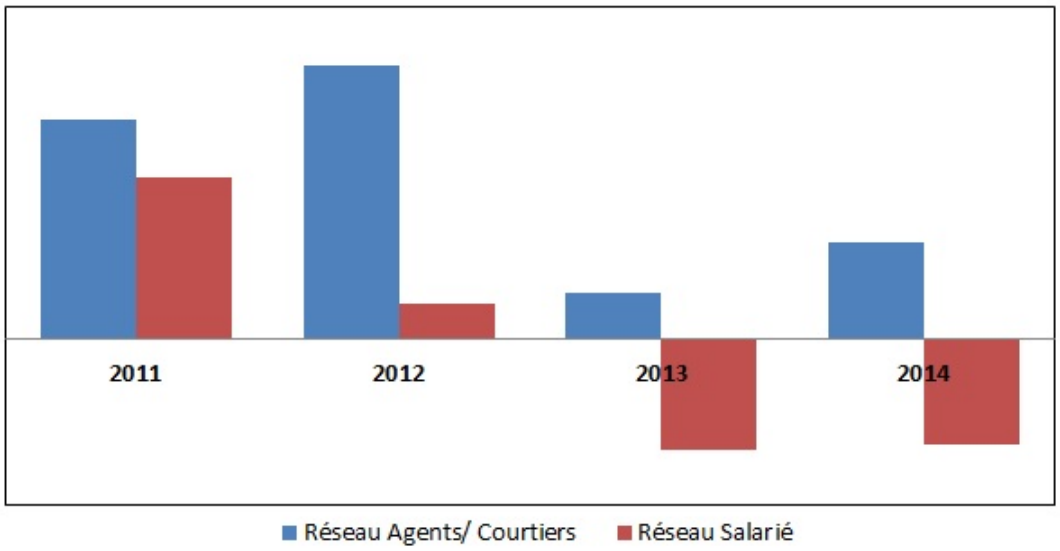


FIGURE 1.5 – Evolution du taux de couverture par réseaux

Pendant l'année 2013 tous les réseaux ont eu une forte baisse de production, cependant le réseau Agents/courtiers a toujours conservé une production nette positive, ce qui signifie que son portefeuille continue d'être en croissance. Au contraire le portefeuille du réseau Salarié est devenu déficitaire avec les années, même s'il béneficie d'une légére amélioration en 2014.

Répartition des prestations par année et par garantie

Exercise	Hospitalisation	Soins courants	Pharmacie	Dentaire	Optique
2011	27,9%	27,1%	20,2%	10,3%	7,8%
2012	27,5%	26,8%	20,8%	10,6%	10,0%
2013	27,0%	26,5%	21,6%	10,7%	10,0%
2014	25,5%	26,8%	23,2%	10,3%	10,0%

TABLE 1.5 – Tableau de prestations totaux des garanties par année

Les garanties proposées par la complémentaire Santé ont des impacts différents pour l'assureur : les postes de garantie hospitalisation, Soins courants et Pharmacie représentent plus de 75 pour cent des prestations totales versées, tandis que les garanties Dentaire et Optique sont à un niveau plus faible.

Il est également intéressant de regarder la fréquence de consommation de ces différentes garanties :

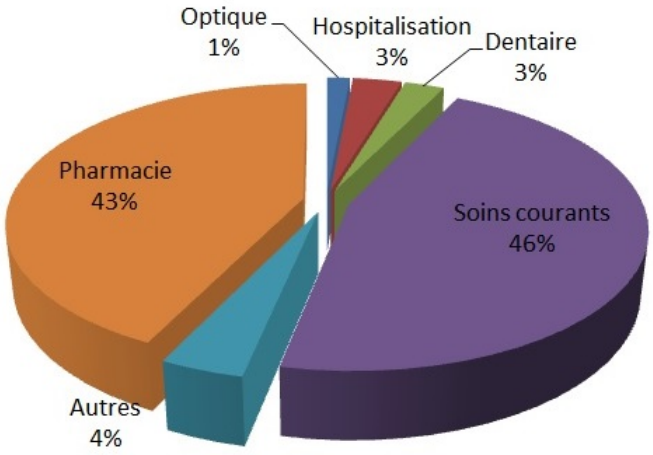


FIGURE 1.6 – Répartition de la fréquence par poste

D'après le tableau et le graphique précédents, on constate que selon le postes, la fréquence de consommation élevée n'explique pas toujours l'importance des montants des prestations. Dans les postes soins courants et pharmacie la sévérité des montants est expliquée par le grand nombre de sinistres, il est donc possible que la survenance d'un seul sinistre ne génère pas un fort impact pour l'assureur. Cependant la somme importante des prestations sur le poste Hospitalisation n'est pas corrélée avec le nombre de sinistres, le faible nombre de survenances de sinistre donne à comprendre qu'un seul sinistre pour ce poste sera très probablement accompagné par des coûts plus importants.

Chapitre 2

Préparation des données en vue de la modélisation

Introduction

En santé, les produits proposent plusieurs formules qui contiennent des garanties différentes. Pour chacune de ces formules, la prime pure est en général calculée comme étant la moyenne des prestations de toutes les garanties confondues, il n'y a donc pas de primes pures propres à chaque poste de garantie. La prime pure étant une mesure du risque, si elle est calculée d'une manière globale, l'assureur n'aura pas la connaissance en détail des risques. Il est donc intéressant de modéliser séparément une prime pure propre à chaque poste de garantie, cela pourra lui permettre d'avoir un recensement des risques individuels mieux ajusté.

L'étude, dans tout ce qui suit, sera segmentée par poste de garantie, il y aura, par exemple, une base de données propre à chaque poste.

2.1 Construction de la base de donnés et des hypothèses

2.1.1 Le périmètre de l'étude

Le pilotage d'un produit nécessite la création d'une base de données fiable afin de déterminer une nouvelle modélisation des primes pures permettant de mesurer les risques. La connaissance des données est donc fondamentale pour arriver à comprendre et bien interpréter les résultats obtenus et éviter ainsi la mise en place de modèles inadaptés. Les données utilisées pour cette étude sont issues des bases brutes extraites mensuellement des systèmes de gestion et informatique. On a créé une historisation des données à partir de ces bases mensuelles pour recenser toutes les informations relatives aux assurés.

Dans un premier temps, la démarche consiste à prendre connaissance des différentes bases, séparées par réseaux. Le traitement des données a été fait en plusieurs étapes : vérification des données, reconnaissance des différentes caractéristiques et sélection des variables significatives. Cette dernière étape est importante puisque la pertinence des résultats dépend directement de celle-ci.

Il y a un grand nombre de variables dans les bases, il est donc nécessaire de faire une analyse de chacune d'entre elles pour arriver à sélectionner celles qui peuvent apporter le plus d'informations.

2.1.2 Hypothèses et création de nouvelles variables

L'information brute des bases de données n'est pas adaptée pour réaliser l'étude, certaines variables ne sont pas utiles dans la modélisation et il a fallu en retraiter d'autres pour mieux comprendre l'information qu'elles apportent. De plus, il est nécessaire de définir de nouvelles variables pour la modélisation.

- Création de la variable *Nombre de bénéficiaires*

Elle représente le nombre de bénéficiaires existants par contrat et par année d'exercice. Dans l'assurance santé complémentaire, elle est classiquement considérée comme une variable tarifaire.

- *Exposition* (Nombre inferieur ou égal à 1)

Temps d'exposition aux risques de l'assuré par année d'exercice, il s'agit de la fraction d'année correspondant à la durée pendant laquelle l'assuré est resté dans le portefeuille.

- Retraitement de la variable *Age*

Dans la base des données brutes sont renseignés la date de naissance et l'âge exact des assurés. Cela peut créer des doublons de lignées dûs à la survenance de l'anniversaire du bénéficiaire avant la date d'anniversaire du contrat, générant alors deux âges différents du bénéficiaire pendant une même année d'exercice. L'âge a donc été retraité pour remédier à ce problème, il est maintenant calculé par différence de millésime, c'est-à-dire comme la différence entre l'année d'exercice et l'année de naissance de l'assuré.

- Retraitement de la variable *nombre de sinistres*

Cette variable, à l'origine représente le nombre d'actes consommés pour un sinistre donné. Par exemple, si un assuré a une ordonnance médicale avec 3 médicaments, dans la base brute le nombre de sinistres vaudra 3 pour la garantie Pharmacie, mais il s'agit d'un seul sinistre donc la valeur que l'on s'attend à voir pour cette variable doit être plutôt égal à 1. Sans retraitement, cette variable conduit à une surestimation de la fréquence des sinistres. Il a donc été nécessaire de créer une nouvelle variable «Nombre de sinistres 2» qui donne le nombre réel de sinistres pour une année donnée, grâce à l'identification des sinistres par le numéro de dossier.

-Renseignement de la commune des assurés, pour le rattachement du code INSEE

Dans le cadre de ce mémoire, une des avancées majeures de la création du nouveau zonier est, entre autres, d'avoir un niveau de détail communal plus fin que celui utilisé actuellement. En effet le zonage actuel est un indicateur tarifaire segmenté par département, la seule variable de l'identification géographique des assurés bien renseignée est celle du département de l'assuré. La commune de l'assuré n'était pas une information priorisée dans les bases.

La variable *Commune*, qui désigne le nom de chaque ville de la France métropolitaine, n'est donc pas renseignée entièrement dans les bases de données, malgré l'importance de cette variable pour la reconnaissance des risques selon la localisation de l'assuré. Nous

sommes donc obligés de faire les démarches nécessaires pour arriver à récupérer le maximum d'information.

Le renseignement historisé de toutes les communes et de tous les changements d'adresse des assurés, pendant leur période d'exposition, a été trouvé facilement pour le réseau Agents/Courtiers. Par contre pour le réseau Salarié, il a fallu réaliser un travail considérable de récupération des adresses des assurés, où finalement la seule information récupérée, concernant la commune des assurés, a été leur dernière adresse en date. Il existe donc une seule adresse connue par assuré pendant toute leur période d'étude, ce qui ajoute une petite contrainte : la possibilité que l'assuré ait déménagé pendant sa période en portefeuille.

On remarque que la variable département est bien renseignée pour chaque assuré pendant tout le période d'exposition.

Problématique

Pour le réseau Salarié, l'information concernant la commune est constante pendant toute la période d'exposition au risque de l'assuré. Pour pouvoir garder un maximum d'information mais avec le minimum d'erreur possible, il va donc falloir réaliser une étude sur la fréquence de changement de la commune des assurés du réseau Agents/Courtiers. Et de cette manière pouvoir construire des hypothèses sur les données du réseau Salarié. Nous étudions donc la fréquence de changement de commune des assurés par année d'exercice, en conditionnant par le fait que l'assuré n'a pas changé de département pendant la même année d'exercice.

Dans une première étape, un calcul de la fréquence du déménagement de département des assurés, sur le réseau salarié, a été fait pour mesurer l'impact du conditionnement imposé.

Les résultats obtenus sont que 97.9 pour cent des assurés n'ont jamais déménagé de département pendant toute leur période en portefeuille.

Il est obligatoire de supprimer tout assuré qui a changé de département sur le réseau Salarié, car le fait qu'ils n'habitent plus dans le même département, implique qu'ils ont forcément changé de commune. Il correspond à l'élimination de 5121 contrats, soit 2.1 pour cent de la base de données du réseau Salarié.

Dans la deuxième étape, nous calculons la fréquence de changement de commune sur le réseau Agents/Courtiers, nous avons pris en compte seulement les données d'assurés qui n'ont jamais déménagé de département.

Avant de faire ce calcul, commençons par la définition suivante :

Definition 2.1.1. *Soit la variable,* ***Changement****, une variable binaire définie par la formule,*

$$\boldsymbol{Changement}(i) := \begin{cases} 1 & \text{Si l'assuré } i \text{ a changé de commune pendant son exposition totale,} \\ 0 & \text{Sinon.} \end{cases}$$

Ainsi la fréquence estimée de changement de commune est donnée par la formule suivante :

$$\textit{Taux de changement}_{commune(j)} = \sum_{i=1}^{N} \frac{\mathbf{Changement}(i)}{N}.$$

Les conclusions obtenues sont qu'il y a 97.22 pour cent des assurés qui ne changent pas de commune pendant leur période en portefeuille.

Le changement de commune en moyenne pour un assuré du réseau Agents/Courtiers est assez faible, donc on obtient une faible marge d'erreur de 2.8 pour cent pour l'acceptation de l'hypothèse qui suit,

<< Si l'assuré n'a pas changé de département alors il n'a pas changé de commune >> .

Les noms des communes étant renseignés pour tous les assurés, il est donc possible de leur associer le code INSEE qui caractérise chaque commune.

2.1.3 Traitement des niveaux de garanties

Le niveau de garantie, c'est-à- dire formule de garantie d'un produit, détermine le montant qui sera engagé par l'assureur en cas de sinistre, c'est pourquoi il intervient en tant que critère de tarification.

Au vu de la diversité du contenu des formules, il nous a paru nécessaire de trouver un indicateur synthétique permettant d'effectuer un classement significatif entre des niveaux de garanties de différentes natures.

Problématique

Pour la modélisation il faut garder des variables qui apportent de l'information mais qui sont entre elles le moins corrélé possible. Bien que la modélisation sera effectuée séparément par poste de garantie et par réseau, il reste encore des variables qui présentent des liaisons entre elles. Les bases contiennent les variables produit et formule, où pour un même produit il y a plusieurs catégories de formules. En observant un même poste de garantie, on constate qu'il y a des formules qui reviennent avec des mêmes niveaux de remboursement, leur distinction n'apporte alors que très peu d'information.

La création d'une nouvelle variable formule à 6 niveaux différents de remboursement pour chaque poste de garantie a été développée. De cette manière les produits seront homogènes entre eux et ce nouvel indicateur viendra remplacer l'ancienne variable formule. Il réduira ainsi les abondantes catégories des formules existantes, qui pour la modélisation des postes de garantie n'ont plus de sens.

La construction de cet indicateur s'est basée sur des niveaux originaux de remboursements proposés aux assurés au moment de la souscription du contrat. Ces niveaux de remboursements théoriques sont présentés dans les tableaux des garanties qu'on retrouve dans l'annexe 1.

La nouvelle variable formule aura les niveaux suivants :

- Formule 1 : elle représente un niveau de remboursement du ticket modérateur ;
- Formule 2 : elle correspond à un niveau d'entrée de gamme ;
- Formule 3 et 4 : elles ont des niveaux de remboursement de milieu de gamme, le niveau 3 est inférieur au niveau 4.
- Formules 5 et 6 : elles correspondent à des niveaux de remboursement haut gamme.

Voici un schéma qui illustre la création de la nouvelle variable formule :

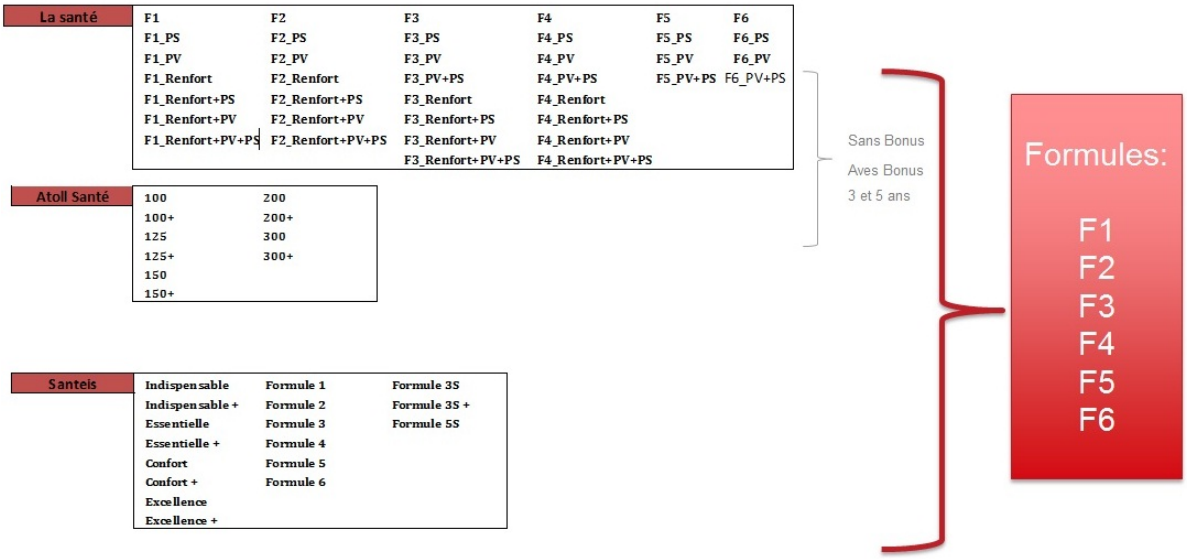


FIGURE 2.1 – Liste de formules commercialisées par produit

On remarque que les produits « La Santé »et « Atoll Santé », commercialisés par le réseau Agents/Courtiers, ont en plus la caractéristique suivante : le bonus à 3 et 5 ans. Ce bonus de fidélité fait bénéficier l'assuré d'un couverture plus importante que celle souscrite au moment de la signature de contrat, l'assuré passe ainsi au niveau supérieur de remboursement. Un traitement plus complexe a donc été réalisé pour ces cas là.

L'illustration suivante donne l'exemple du traitement d'une formule du Produit «La Santé» :

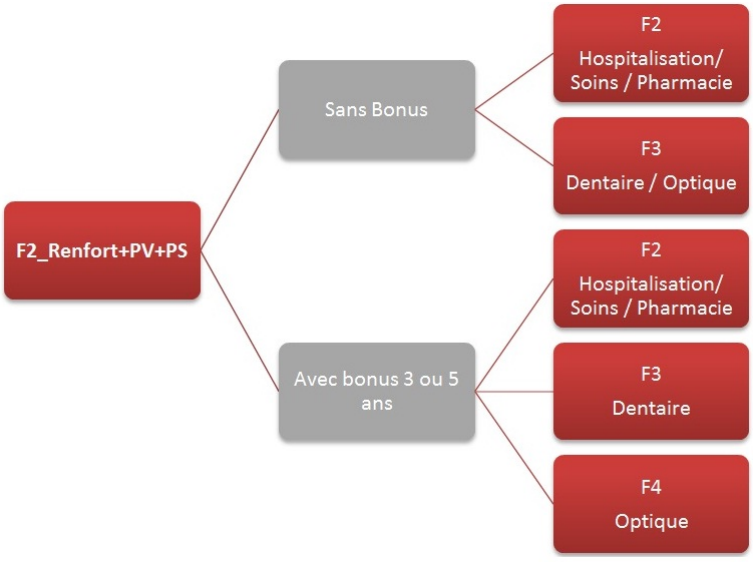


FIGURE 2.2 – Illustration du passage vers la nouvelle variable formule

Cette étape du travail a été longue à cause de la grande quantité de formules initiales et de l'étude détaillée nécessaire pour chaque tableau. Le rassemblement des formules apporte donc une approche plus juste pour la modélisation souhaitée pour chaque poste de garantie.

2.1.4 Liste des variables tarifaires potentiellement éligibles au modèle

Par réseau, les bases finales des risques portent le même nombre de lignes. Les bases Hospitalisation, Soins Courants, Pharmacie, Dentaire et Optique possèdent toutes les mêmes dimensions, avec les mêmes caractéristiques par bénéficiaire. Les seules variables qui changent entre elles sont le nombre de sinistre et le montant de prestations, car elle sont issues du poste de garantie respectif.

Chaque ligne de la base représente un bénéficiaire qui n'a pas changé sa situation de risque pendant une même année d'exercice. Par conséquent suite à un changement de situation de risque on crée une nouvelle ligne, par exemple, si un bénéficiaire a changé de régime de TNS vers salarié, sa situation de risque change et on aura donc une nouvelle ligne correspondant à sa nouvelle situation. Les situations de risque sont caractérisée par les variables portefeuille suivantes :

Variable	Descriptif
Police	Numéro propre à chaque contrat souscrit
Rang	Chiffre attribué à chaque bénéficiaire dans un contrat
Sexe	Indicateur du sexe de chaque assuré
Age∗	Age millesimal de l'assuré
Régime	TNS, Salarié, Agricole, Alsace-Moselle ou autre
Zone	Zone détermine par l'ancienne zonier
Département	Département de la France métropolitaine
Commune∗	Nom de la commune
Code INSEE	Code officiel géographique des communes
Nb.benef∗	Nombre des béneficiaires par contrat
Exposition	Durée d'exposition au risque par exercice
Réseau	Agents/courties ou Salarié
Formule∗	Code de la formule souscrite dans le contrat
Annee.surv	Anné de survenance du sinistre
MT.Remb.poste	Prestation versées par l'assureur
Nb.sinistre.poste	Nombre de sinistres survenus

TABLE 2.1 – Liste des variables du portefeuille

On remarque que les variables accompagnées d'astérisques correspondent à toutes les variables que nous avons dû retravailler ou créer.

Liste des variables tarifaires

Les variables ci-dessus ne sont pas tous nécessaires pour effectuer la modélisation, voici la liste finale des variables tarifaires potentiellement éligibles dans le modèle :

Variable	Descriptif
Age	Age millesimal de l'assuré
Régime	TNS, Salarié, Agricole, Alsace-Moselle ou autre
Nb.benef∗	Nombre des béneficiaires par contrat
Formule	Code de la formule souscrite dans le contrat
Annee.surv	Anné de survenance du sinistre

TABLE 2.2 – Variables Tarifaires

Les variables « Zone », « Dpartement », « Commune »et « Code INSEE »ne sont pas des variables à utiliser directement dans le modèle de prime pure, elles sont nécessaires à la sélection des variables externes pour la modélisation du risque géographique. Et les variables « Exposition », « Montant de rembousement », et « Nombre de sinistres »correspondent celles qui constituent la ou les variables à expliquer.

2.2 Zoom sur le coefficient de risque géographique

2.2.1 Etude sur la conformité de l'actuel zonier

Jusqu'à présent, les tarifs des produits Santé ont été déterminés par les coefficients géographiques donnés par l'actuel zonier. Ce zonier est unique quel que soit le niveau de garantie. Il est donc important qu'il représente bien le niveau de risque d'une manière claire et cohérente. Ce zonier a été fait en 2010, grâce à la modélisation du risque des données du réseau « agents/Courtiers »par la méthode du GLM, suivant les variables de risques propres au portefeuille à l'époque.

Les zones ont été créées à partir de la connaissance du risque moyen en santé, c'est à dire une prise en compte de tous les postes de garanties confondus. Il est donc important de réaliser une analyse descriptive des résultats empiriques pour déduire si le zonier est donc encore adapté ou pas.

Dans la section précédente on avait déjà constaté que les postes de garanties n'ont pas les mêmes impacts sur le portefeuille, la sévérité d'un sinistre survenue dans un poste n'est pas comparable à celle produite dans un autre. Si l'actuel zonier est encore bien adapté au portefeuille, il faut qu'il continue à conserver un effet décroissant par zone.

Les graphiques suivants représentent la consommation moyenne observée par poste :

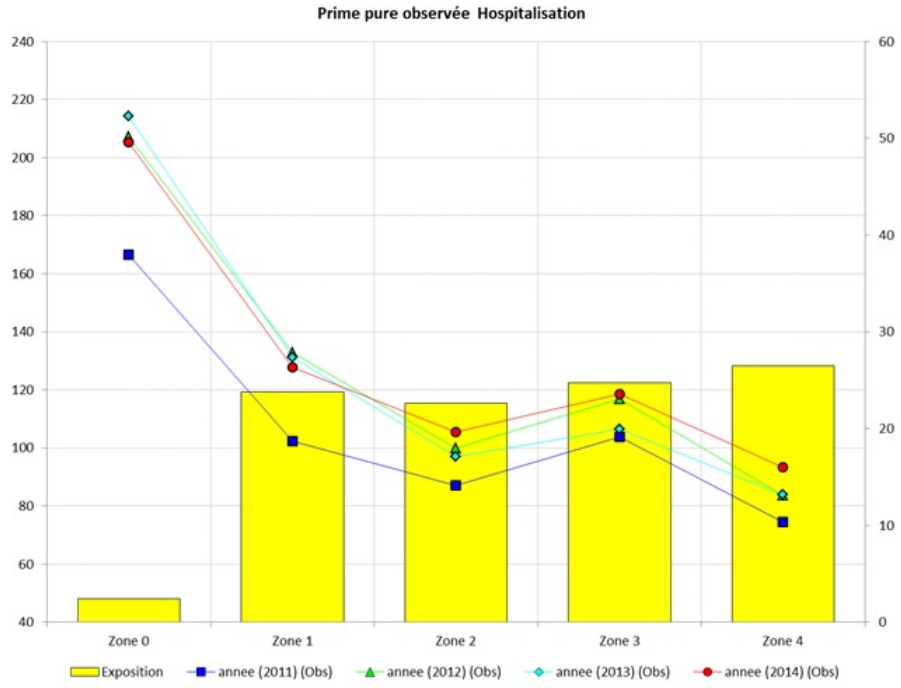


FIGURE 2.3 – Répartition par zone de la prime pure observée du poste Hospitalisation

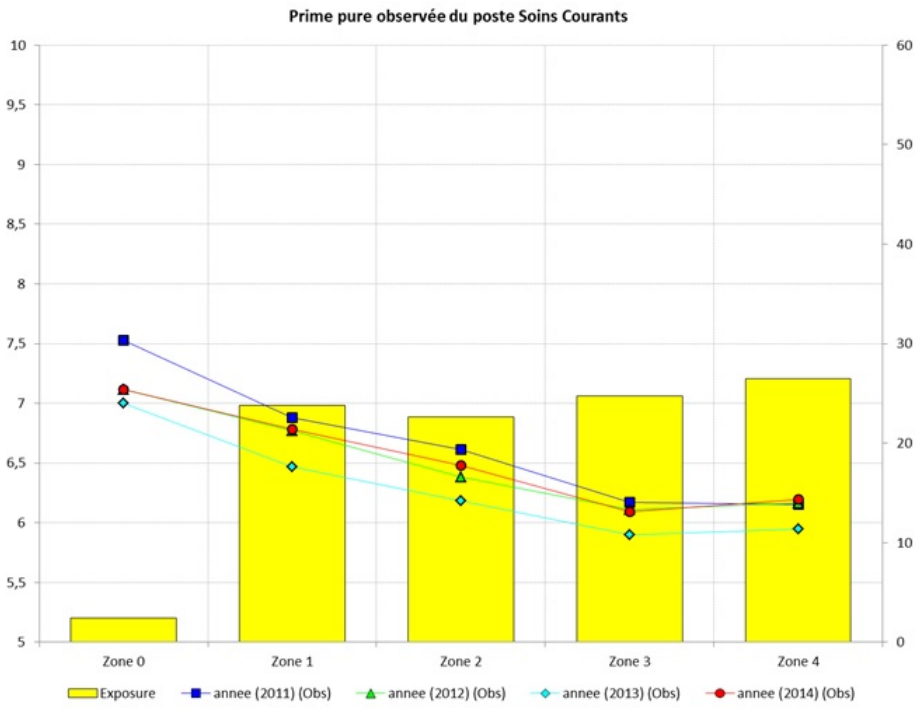


FIGURE 2.4 – Répartition par zone de la prime pure observée du poste Soins Courants

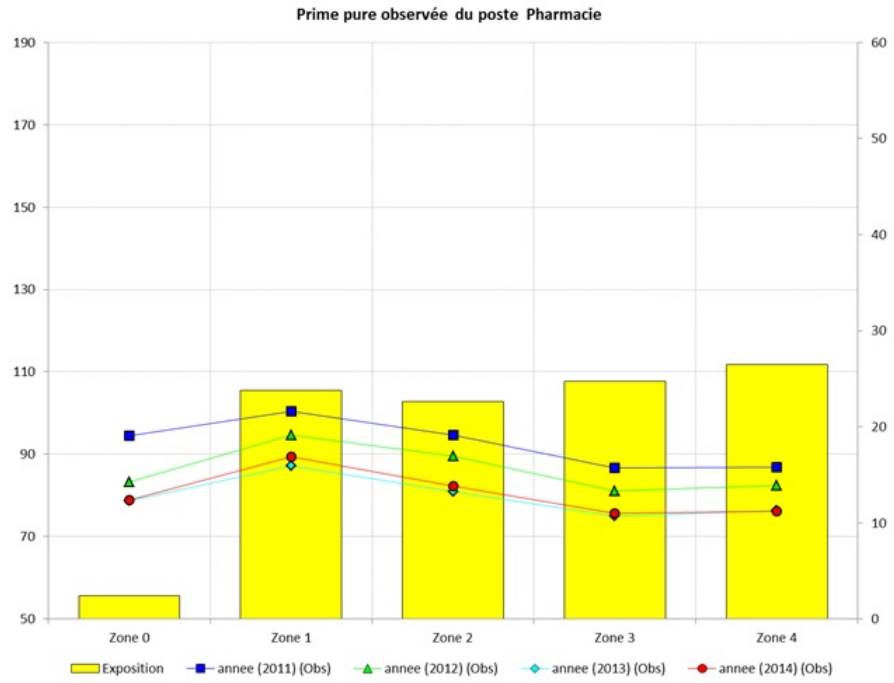


FIGURE 2.5 – Répartition par zone de la prime pure observée du poste Pharmacie

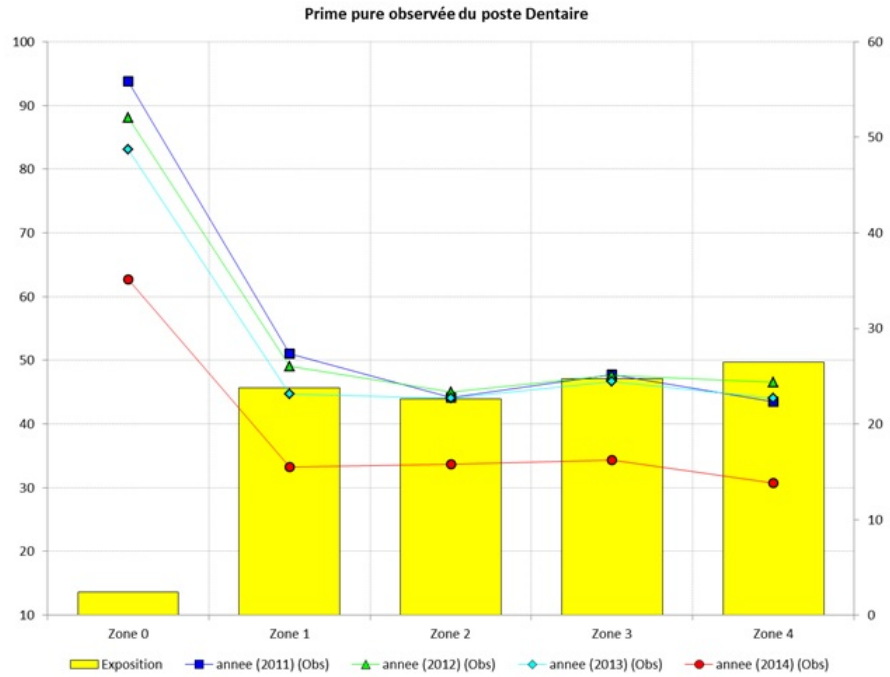


FIGURE 2.6 – Répartition par zone de la prime pure observée du poste Dentaire

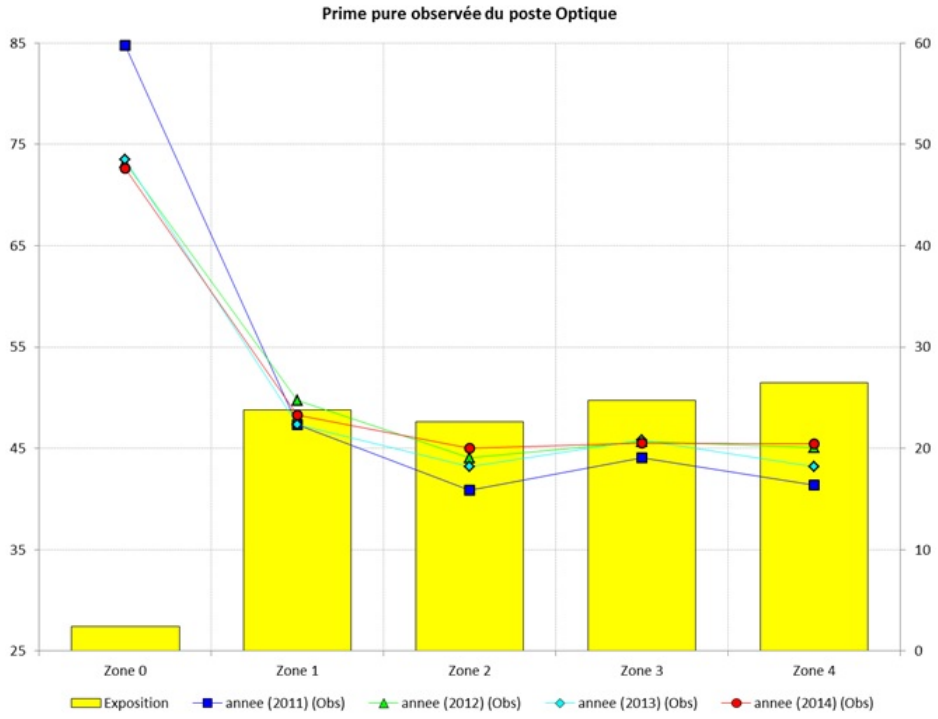


FIGURE 2.7 – Répartition par zone de la prime pure observée du poste Optique

Dans un premier temps, on constate que le nombre des effectifs est très faible dans la zone zéro, elle correspond à celle qui possède le coefficient tarifaire le plus fort. Globalement cette zone dispose d'une consommation moyenne observée très élevée avec des écarts significatifs à l'égard des zones qui la suivent.

On remarque que pour les postes hospitalisation, dentaire et optique, la zone 3, une des plus représentées, rompt la tendance et présente un niveau de consommation plus important que celle de la zone 2. En général la consommation moyenne empirique ne respecte pas toujours le caractère décroissant imposé par les zones dans le tarif, en conséquence l'actuel zonier n'est plus adapté à la sinistralité observée dans ces derniers années et confirme donc la nécessité de son ajustement.

2.2.2 Représentation de la consommation moyenne par département

L'actuel zonier présente la répartition suivante sur la carte des départements :

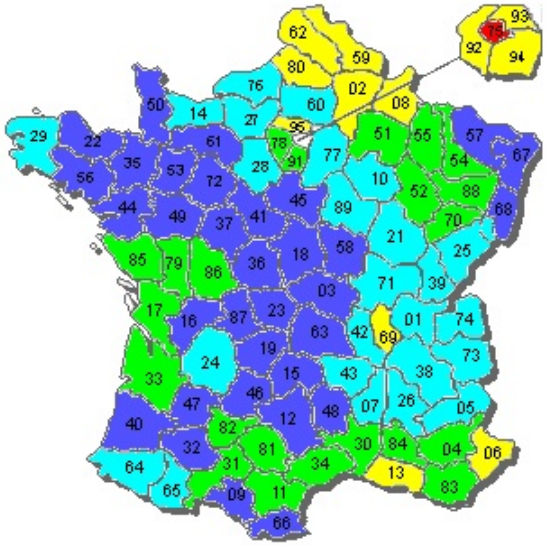


FIGURE 2.8 – Répresentation de l'actuel zonier sur la carte de la France

La zone rouge est la zone 0, elle correspond à la zone où le tarif est le plus élevé. la zone jaune correspond à la zone 1 ; la zone verte est la zone 2,elle correspond à la zone où le tarif a un niveau moyen ;la zone bleu clair correspond à la zone 3 et la zone bleu foncé est la zone 4, elle correspond à la zone où le tarif est le moins élevé.

La nouvelle proposition de modélisation de la prime pure repose sur une vision du risque mieux segmentée, on a donc réalisé une étude séparée pour chaque poste de garantie.
Les classes ont été créées en fonction de la distribution des quantiles des dépenses pondérée par le poids du département dans le portefeuille, nous avons choisi 5 niveaux différents :

Zone 1 : niveau de consommation moyenne compris dans les 20% plus bas.
Zone 2 : niveau de consommation moyenne compris entre les quantiles de 20 et 40%
Zone 3 : niveau de consommation moyenne compris entre les quantiles de 40 et 60%
Zone 4 : niveau de consommation moyenne compris entre les quantiles de 60 et 80%
Zone 5 : niveau de consommation moyenne supérieure au quantile de 80%.

	Hospitalisation	Soins Courants	Pharmacie	Dentaire	Optique
Zone 1	[0, 78.62]	[0, 86.22]	[0, 73.17]	[0, 29.84]	[0, 34.47]
Zone 2	] 78.62, 87.59]	] 86.22, 92.51]	] 73.17, 79.77]	] 29.84, 32.91]	] 34.47, 37.99]
Zone 3	] 87.59, 101.39]	] 92.51, 97.67]	] 79.77, 86.31]	] 32.91, 37.21]	] 37.99, 40.90]
Zone 4	] 101.39, 120.70]	] 97.67, 106.37]	] 86.31, 96.00]	] 37.21, 42.93]	] 40.90, 45.56]
Zone 5	> 120.70	> 106.37	> 96.00	> 42.93	> 45.56

TABLE 2.3 – Intervalles de prime pure observée par poste

Les cartes suivantes indiquent la répartition des classes de la consommation moyenne annuelle par département.

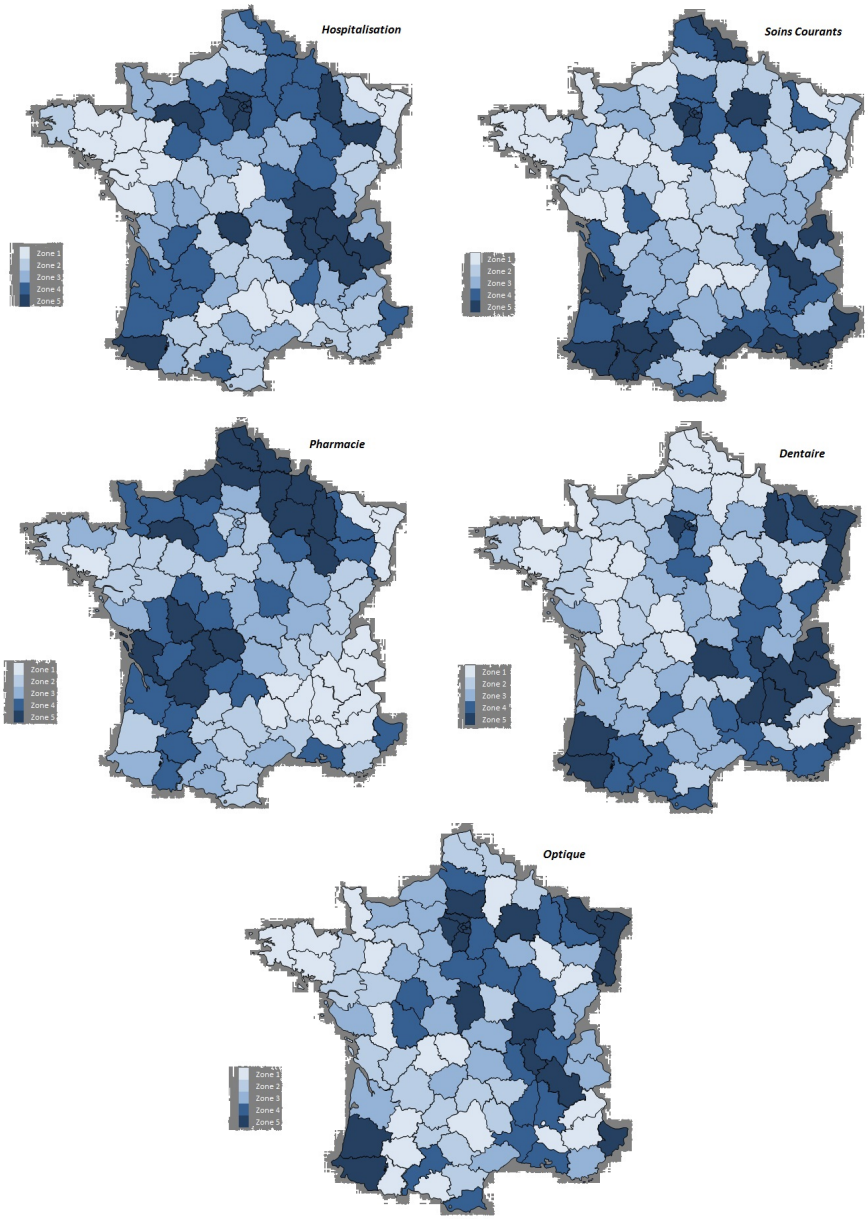


FIGURE 2.9 – Répartition des zones de prime pure observée par poste

En général la répartition par département n'est pas la même entre les poste de garanties.

On constate que les zones définies en 5 niveaux de consommation moyenne annuelle ne sont pas homogènes entre les postes.

Pour avoir une bonne maitrise du risque, il est donc fondamental que l'analyse puis la modélisation du risque géographique en santé soient réalisées séparément pour chacun des postes. De cette façon, on obtiendra une segmentation appropriée aux différents niveaux de consommation, et l'estimation du risque sera ainsi plus représentative du vrai comportement des assurés en fonction de son lieu de résidence.

2.3 Les variables externes

2.3.1 Recherche des variables externes

L'objectif principal de ce mémoire est la modélisation du risque géographique en santé. On part de l'hypothèse que les niveaux de dépenses ne sont pas définis uniquement par les caractéristiques propres de chaque assuré, c'est-à-dire que les composantes du risque d'un assuré ne dépendent pas exclusivement de son âge, du régime et du niveau de garantie souscrite dans son contrat d'assurance complémentaire santé (Habiter dans une zone ou dans une autre interfère directement dans les niveaux de dépenses en santé, soit à cause des tarifs imposés par l'offre et la demande (note : bien que l'état intervienne pour régulariser certains niveaux de dépassements, il n'est pas obligatoire pour les acteurs de soins en santé d'y adhérer), soit par la qualité de vie de la zone où l'assuré habite).

Pour identifier le vrai effet de risque lié à la zone géographique, il est fondamental de disposer d'une connaissance, la plus large possible, de chaque région. Le portefeuille utilisé pour l'étude ne contient pas une information exhaustive sur les caractéristiques de chaque département ou commune de la France, c'est pourquoi on doit incorporer des variables externes dans l'étude. Les variables externes constituent un complément d'informations qui contribueront à la création du nouveau zonier.

La recherche des variables externes se base sur des conclusions obtenues auprès de différentes publications des études gouvernementales sur les dépenses en santé. Ces études ont pour objectif de prévoir l'évolution des dépenses totales de santé en fonction de l'évolution de différents indicateurs socioéconomiques reflétant les perspectives démographiques d'un pays et/ou de la richesse nationale.

« L'évolution de l'état sanitaire de la population façonne les besoins et donc la demande en matière de soins. À cet égard, si le vieillissement est bien de nature à susciter une hausse des besoins, son effet en propre ne fait pas consensus dans la littérature. Par ailleurs, il faut noter que la liaison entre dépense de santé et vieillissement est complexe car à double sens (si l'espérance de vie augmente, c'est notamment du fait d'une dépense de santé plus élevée) », Les dépenses de santé en France : déterminants et impact du vieillissement à l'horizon 2050, Ministère de Santé, **[8]**.

« En un sens étroit, le mécanisme examiné ici est l'élévation de la demande de soins entraînée par une hausse du revenu. Bien que certaines estimations agrégées concluent à une forte liaison entre ces deux variables (élasticité proche ou supérieure à l'unité), il apparaît douteux de donner à ces résultats une interprétation directement causale. Plus vraisemblablement, une hausse générale du niveau de vie s'accompagne d'un ensemble d'évolutions de nature à stimuler la dépense de santé, tant du côté de l'offre (hausse des ressources collectives permettant de mettre plus facilement en place de grands programmes publics) que de la demande (élévation du niveau d'éducation et plus généralement des mœurs suscitant de

nouvelles attentes de la population) », Projection des dépenses de santé à l'horizon 2060, le modèle PROMEDE, Minitère de Santé, **[16]**.

Dans un premier temps, on a du vérifier l'existence d'un lien entre le portefeuille et la population de la France, pour cela on a étudié la corrélation entre la répartition de la population et le nombre des effectifs du portefeuille par département.

Population de la France vs effectifs du portefeuille par département	
Coefficients de Corrélation de Pearson	0.84292
Coefficients de Corrélation de Spearman	0.89439

TABLE 2.4 – Corrélation entre la répartition des effectifs du portefeuille et la population de la France

Il existe une forte corrélation entre le portefeuille et la population de la France. En conséquence, la base de données représente bien la structure de la France, la répartition des départements coïncidant grandement avec elle.

Suite à la vérification de la concordance entre l'information de la France et le portefeuille, il est donc possible de partir à la recherche des indicateurs externes qui viendront enrichir les caractéristiques des assurés en fonction de la zone géographique où ils habitent.

2.3.2 Indicateurs externes représentant la sinistralité à modéliser

La sélection à priori des variables externes, potentiellement explicatives dans le modèle, a été réalisée en fonction d'une étude de liaisons entre ces variables et un indicateur externe particulier. Ce dernier a été choisi de manière à ce qu'il représente au mieux le risque à modéliser.

Les différentes sources mettent à disposition plusieurs données ainsi qu'un grand nombre d'informations. Il est donc important d'utiliser un indicateur externe largement corrélé avec la consommation moyenne observée du portefeuille, pour pouvoir effectuer une présélection des facteurs externes sur la base des effectifs de la population française.

Les indicateurs externes ont été choisis de manière à ce qu'ils expliquent au mieux les risques liés à chaque poste. A titre d'exemple, ci-dessous, la vérification du choix de l'indicateur externe pour le poste hospitalisation :

Poste Hospitalisation

- Variable externe représentant la variable à modéliser :

$$\text{Fréquence des Séjours hospitalisation } = \frac{\sum \text{Nombre de séjours en hospitalisation}}{\text{population totale du département}}$$

- Variable interne à modéliser, données entre 2011-2014 :

$$\text{Fréquence Hospitalisation} = \frac{\sum \text{Sinistres du poste hospitalisation}}{\sum \text{exposition du département}}$$

Statistiques simples						
Variable	N	Moyenne	Ecart-type	Médiane	Minimum	Maximum
Sejour.hab	95	0.26888	0.02131	0.26674	0.21441	0.32760
Freq.Portefeuille	95	0.48646	0.07346	0.48546	0.22722	0.67223

TABLE 2.5 – Statistiques simples

fréquence des séjours hospitaliers par habitant et du portefeuille par département	
Coefficients de Corrélation de Pearson	0.4255
Coefficients de Corrélation de Spearman	0.3985

TABLE 2.6 – Corrélation entre la fréquence national des séjours hospitaliers et fréquence moyenne du portefeuille

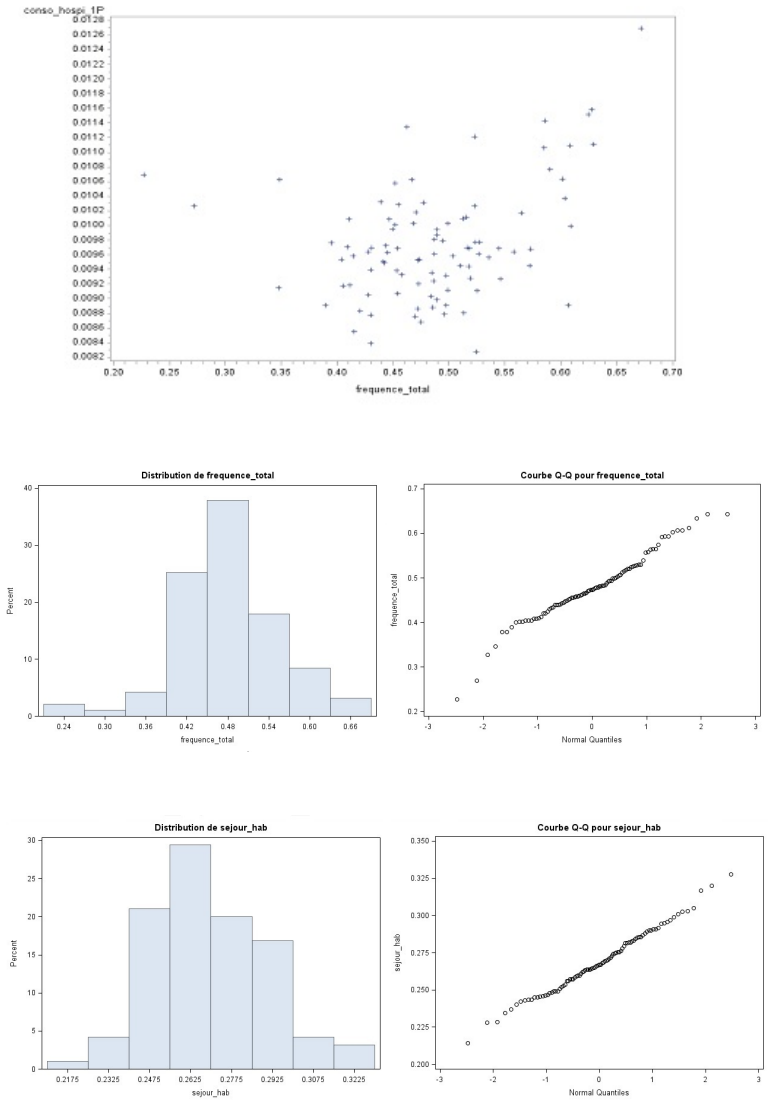


FIGURE 2.10 – Étude sur la variable Hospitalisation

On constate que la fréquence des séjours à l'hôpital par département de la France est un bon indicateur de la sinistralité qu'on cherche à modéliser pour le poste hospitalisation. Les deux distributions sont similaires et il existe une corrélation entre ces deux variables. On peut donc sélectionner cet indicateur pour la sélection des différentes variables externes qui s'ajouteront à la modélisation du risque géographique en hospitalisation.

On a ensuite obtenu un indicateur externe pour chaque poste de garantie. Ils seront utilisés dans la sélection de variables externes qui sont listées dans la section suivante.

2.3.3 Liste des variables externes potentiellement explicatives

Liste des variables externes au niveau départemental

Indicateurs démographiques

- Structure par âge de la population.
- Taux de natalité en 2012, Naissances domiciliées pour 1000 habitants.
- Taux de mortalité en 2012, Décès domiciliés pour 1000 habitants.
- Indice de vieillissement de la population en 2013, Nombre de personnes de 65 ans et plus pour 100 personnes de moins de 20 ans.
- Espérance de vie à la naissance en 2012 Hommes et Femmes.
- Espérance de vie à 65 ans en 2012.

Causes médicales de décès

- Maladies infectieuses et parasitaires, dont sida et V.I.H.
- Tumeurs.
- Maladies endocriniennes, nutritionnelles et métaboliques.
- Troubles mentaux et du comportement, dont abus d'alcool (y compris psychose alcoolique).
- Maladies du système nerveux et des organes des sens.
- Maladies de l'appareil circulatoire.
- Maladies de l'appareil respiratoire.
- Maladies de l'appareil digestif.
- Maladies de l'appareil génito-urinaire.
- Causes externes de blessure et empoisonnements, dont suicides et accidents de transport.
- Symptômes, états morbides mal définis et autres causes.

État de santé

- Maladies à déclaration obligatoire : nombre des cas Hépatite A et B, Infection par le VIH, Infections invasives à méningocoques, Tuberculose et Toxi-infections alimentaires collectives.
- Séjours en hospitalisation en Médecine, Chirurgie, Gynécologie-obstétrique, secteur public et secteur privé.
- Mode de cohabitation des personnes de 75 ans et plus en 2011 :
  — Vivant en couple, en couple ou seules avec leur(s) enfant(s) (%).
  — Vivant en institution (%).
  — Ne vivant pas seules : autres cas (%).
  — Vivant seules (%).

- Pauvreté et lutte contre les exclusions : Taux de pauvreté monétaire 2011 et Bénéficiaires de la couverture maladie complémentaire (CMUC) au 31.12.2012.

Liste des variables externes au niveau communal

- Superficie.
- Densité en Km^2.

Contexte sociodémographique

- Structure de la population.
- Catégories socioprofessionnelles.
- Population active.
- Niveau de revenus moyen des ménages.
- Niveau de revenus par activité.
- Nombre des chômeurs.
- Nombre des retraités.

Nombre de fonctions médicales et paramédicales

- Médecin omnipraticien (généraliste).
- Médecin Spécialiste (cardiologie, dermatologie, gynécologie, gastro-entérologie, psychiatrie, entre autres).
- Spécialiste en ophtalmologie.
- Chirurgien-dentiste.
- Sage-femme, Infirmier.
- Masseur kinésithérapeute, Orthophoniste, Orthoptiste, Pédicure-podologue, Audio prothésiste, Ergothérapeute et Psychomotricien.

Nombre d'équipements et de services de santé

- Etablissement santé (Hopitaux).
- Etablissement psychiatrique.
- Centre lutte cancer.
- Urgences.
- Maternité.
- Centre de santé.
- Structures psychiatriques en ambulatoire.
- Centre médecine préventive.
- Dialyse.
- Hospitalisation à domicile.
- Maison de santé pluridisciplinaires.
- Pharmacie.
- Laboratoire d'analyses médicales
- Ambulance.
- Transfusion sanguine.
- Etablissement thermal.
- Etablissement lutte contre l'alcoolisme.

Les données externes ont été collectées à partir de différentes sources : Données INSEE **[3]**, Données DREES **[2]**, IRDES **[4]**, Recueil d'indicateurs régionaux ; Offre de soins et état de santé édition 2014, **[11]** et STATISS, statistiques et Indicateurs de la Santé et du Social 2014, **[27]**.

2.3.4 Traitement des variables externes

La liste des variables externes doit être traitée afin de rendre l'information exploitable et adaptée à notre étude. Voici quelques exemples :

- Traitement de la variable externe revenu médian par commune,

$$taux_{revenu} = \frac{Revenu\ médian\ par\ comunne}{Revenu\ National\ par\ menage\ médian}$$

où *Revenu National par menage médian* en 2011 est égal à 29590 euros.
Ce traitement est nécessaire pour pouvoir comparer les niveaux de revenus entre les différentes communes. Cette nouvelle variable permet d'indiquer si une commune possède ou pas un niveau de revenu au-dessus de la mèdiane nationale.

- Nombre de chômeurs entre 15 et 64 ans dans la commune i,

$$taux_{chomage_i} = \frac{Nb\ de\ chômeurs\ en\ i}{Population\ totale\ de\ i}.$$

- Traitement de la variable « Nombre de pharmacies »dans la commune i,

$$densi_{pharmacie_i} = \frac{Nb\ de\ pharmacies\ dans\ i}{Superficie\ de\ i}.$$

- Traitement de la variable « Nombre de décès avant 65 ans »,

$$Taux\ DC_{avant65ans_i} = \text{Nb de décès avant 65 ans}_d times \frac{Popultion_i}{Population_d},$$

où i correspond à la commune et d au département auquel la commune i appartient.
Ce dernier traitement est indispensable pour passer des variables départementales à des variables communales dans la modélisation du risque au niveau du code INSEE.

2.3.5 Etude d'analyse des composantes principales (ACP)

Les variables externes propres à l'état de santé des habitants sont assez nombreuses. Bien qu'un maximum d'information aide toujours à mieux expliquer le risque à modéliser, il faut rester prudent et attentif au fait que l'utilisation d'un grand nombre de variables rend plus difficile la sélection des variables explicatives. Lorsqu'on étudie simultanément un nombre important de variables quantitatives, l'analyse des composantes principales vient à réduire la dimension et ainsi éviter la redondance d'information, en la remplaçant par une plus juste et objective.

L'objectif de l'ACP est de construire de nouvelles variables, appelées composantes principales, non corrélées et qui permettent de synthétiser l'information. Elles sont construites comme des combinaisons linéaires des variables initiales. Pour visualiser la liaison entre la composante principale et les variables initiales, on représente en ACP normée, les variables dans les plans factoriels. Ces variables sont réduites, c'est-à-dire telles que la variance vaut 1. Cela est nécessaire pour attribuer à chacune d'entre elles une même importance dans l'analyse, leur contribution étant proportionnelle à leur variance. Pour définir le nombre de composantes principales, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projetée sur un axe donné. On caractérise ainsi chaque axe par le pourcentage d'inertie qu'il permet d'expliquer.

Voici les résultats obtenus de l'analyse des composantes principales sur les variables externes propres à la santé.

Valeurs propres de la matrice de covariance			
Valeur propre	Différence	Proportion	Cumulé
16.5709279	10.1041705	0.5021	0.5021
6.4667573	3.8512661	0.1960	0.6981
2.6154913	0.6244485	0.0793	0.7774
1.9910428	0.9940331	0.0603	0.8377
0.9970097	0.2672306	0.0302	0.8679
0.7297790	0.0392278	0.0221	0.8900

TABLE 2.7 – Valeurs propres de la matrice de covariance

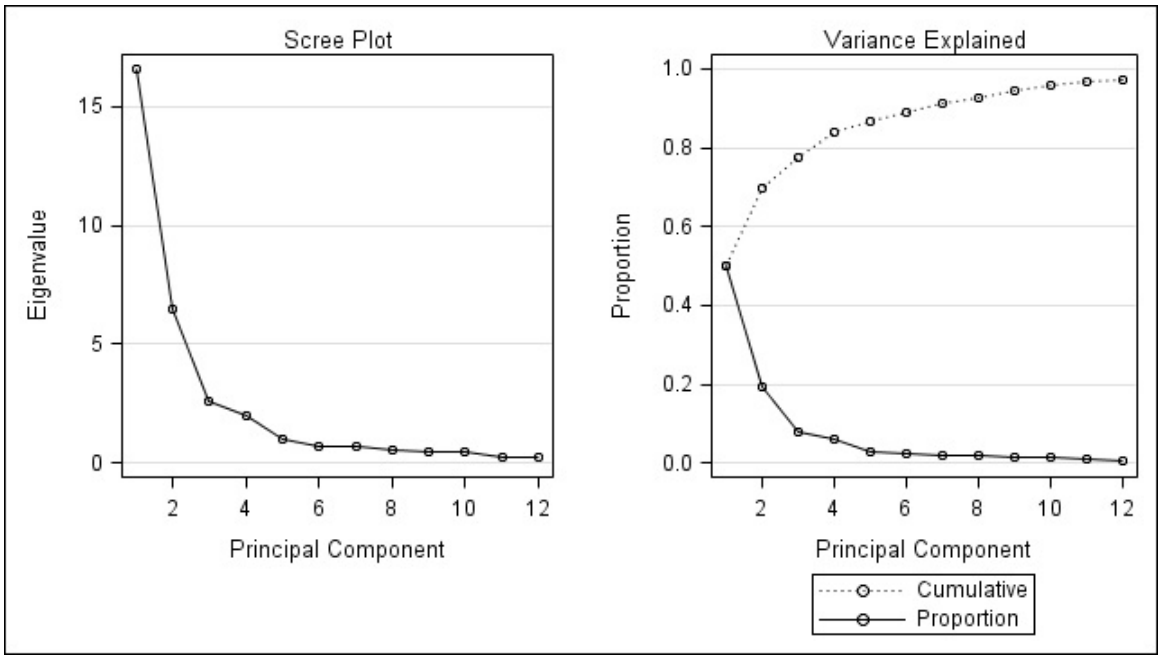


FIGURE 2.11 – Valeurs propres

Sélection des nombres d'axes à retenir :

— **Critère du coude de Cattell** : si on observe un décrochement (coude) sur les valeurs propres suivi d'une décroissance régulière, on doit alors retenir uniquement les axes avant le décrochement.
— **Critère de Kaiser** : on ne retient que les axes dont l'inertie est supérieure à l'inertie moyenne I/p, où p correspond au nombre de variables. Dans le cas d'une ACP normée, on ne retiendra que les axes associés à des valeurs propres supérieures à 1.

À partir des résultats obtenus, on constate seuls 4 valeurs propres répondent aux critères cités ci-dessus.

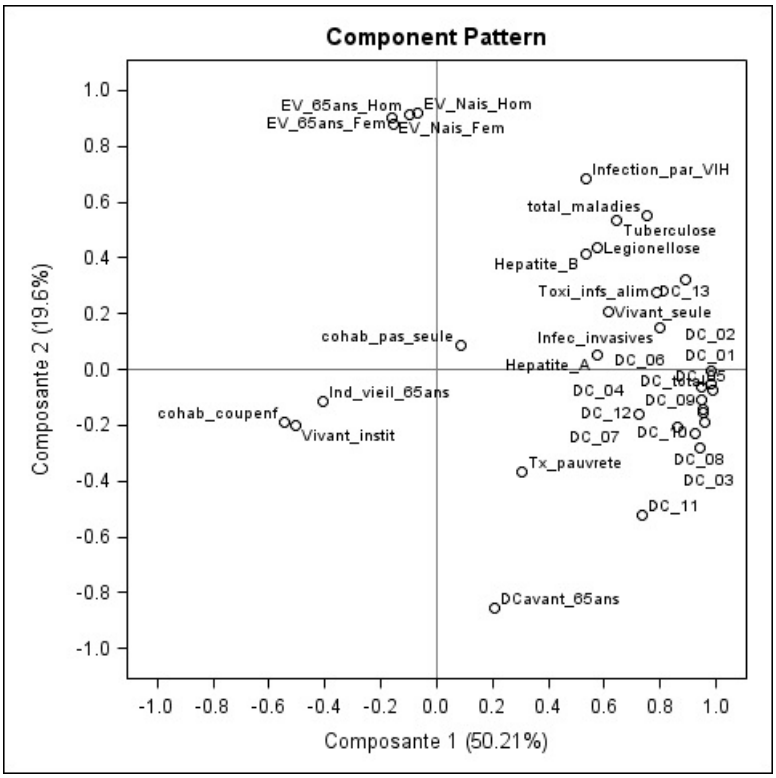


FIGURE 2.12 – Plan factoriel

La plupart des variables santé sont assez liées entre elles. D'un côté, on peut voir que les variables qui appartiennent à la section « Maladies à déclaration obligatoire »(nombre de cas Hépatite A et B, Infection par le VIH, ...) et celles des « Causes médicales de décès »(décès par tumeurs, par Maladies de l'appareil circulatoire, ...) sont corrélées positivement par rapport à l'axe 1. D'un autre, on constate une forte liaison entre les variables concernant l'espérance de vie des habitants, elles sont bien expliquées par l'axe 2 et on identifie une indépendance avec les variables qui sont bien expliquées par l'axe 1. Cependant, pour l'étude il n'est pas recommandé d'effectuer un regroupement des variables santé par les composantes principales car on perd beaucoup d'information qualitative.

La modélisation du risque par poste pourra nécessiter certaines variables santé sur un poste et très probablement d'autres sur un autre poste. Néanmoins cette analyse est importante pour identifier et mieux comprendre les relations existant entre ces variables, et ainsi anticiper les difficultés qui pourront se présenter au moment de la modélisation.

Pour une explication plus détaillée sur la dépendance entre variables explicatives, voir la section 3.2.2.

2.3.6 Classifications des variables par la méthode de « k-means »

Les variables externes sont des variables quantitatives, pour pouvoir les intégrer dans un modèle GLM il est donc nécessaire de les transformer en variables qualitatives, à partir de regroupement en classes les plus homogènes possibles. La mèthode des k-means reste actuellement la méthode la plus utilisée, elle est particulièrement adaptée quand le nombre de données est important. La classification se base sur le critère des voisins les plus proches, ce qui veut dire que chaque individu est affecté à une classe donnée s'il est très proche du centre de gravité de la classe.

L'algorithme

Etape 1 : Initialement, le choix des centres s'effectue sur la base d'un tirage aléatoire sans remise de k observations à partir de la population à classifier, pour former les premiers k centres provisoires des classes. Ensuite chaque observation est affectée à la classe le plus proche, après avoir affecté toutes les observations, on obtient k classes avec ses centres respectifs.

Etape 2 : En considérant les k classes obtenues, on calcule leurs nouveaux centres de gravité et on obtient donc k nouveaux centres. On utilise la même règle d'affectation de l'étape 1 pour obtenir les k nouvelles classes. L'algorithme s'arrête lorsque deux itérations successives conduisent à une même partition ou bien lorsqu'on fixe un critère d'arrêt tel que le nombre maximal d'itérations.

Remarques sur la méthode :
- La classification utilisant cette méthode dépend du choix des centres initiaux. Deux tirages aléatoires des centres peuvent engendrer deux typologies différentes.
- La méthode des k-means est fortement liée au nombre k de classes fixées à priori. Cependant la classification en l classes avec $l > k$ peut être largement différente de la classification en k classes.

Notons finalement que le choix du nombre de classes k est basé sur la règle de Sturges

$$k = 1 + \frac{33}{10} \cdot \log(n), \quad \text{avec } n = 36610.$$

D'où $k = 16$.

Voici un exemple du classement obtenu pour la variable « taux de revenu »,

N° classe	Centre de classe	Répartitions des communes
1	0,372	3,7%
4	0,462	6,8%
15	0,515	9,8%
10	0,555	14,0%
9	0,589	13,7%
6	0,624	13,2%
14	0,659	11,6%
5	0,695	9,0%
13	0,736	7,0%
2	0,783	5,0%
8	0,838	2,8%
7	0,904	1,6%
12	0,979	0,9%
11	1,076	0,5%
16	1,195	0,3%
3	1,387	0,1%

TABLE 2.8 – Classement obtenu pour la variable taux de revenu

Chapitre 3

Modélisation des postes par la méthode du GLM

Introduction

Ce chapitre présente les bases aussi bien théoriques que pratiques qui donnent les connaissances nécessaires à l'identification des profils de risque, en expliquant les différents niveaux de consommation. L'objectif est d'estimer la consommation moyenne en santé à partir des variables explicatives, non seulement des variables tarifaires propres au portefeuille, mais aussi en intégrant des variables externes qui vont apporter l'information supplémentaire, en vue de la modélisation du risque géographique pour la construction d'un nouveau zonier.

L'estimation revient à prédire l'espérance d'une variable aléatoire et la méthode choisie dans cette optique est celle des Modèles linéaires généralisés, lesquels sont couramment utilisés pour la modélisation non vie. Les modèles linéaires généralisés ont remplacés progressivement les régressions linéaires simples car ils ont l'avantage d'être adaptés à la modélisation de variables qui prennent leurs valeurs dans un sous-ensemble de $\mathbb{R}$ comme $\mathbb{R}^+$, par conséquent le caractère normal de la variable à expliquer Y n'est plus imposé, maintenant il suffit que Y appartienne à la famille des distributions exponentielles.

3.1 Les modèles linéaires géneralisés

3.1.1 Définition des Modèles linéaires généralisés

Les modèles linéaires généralisés ou GLM (pour Generalized Linear Models) ont comme objectif l'estimation de l'espérance d'une variable aléatoire réponse (variable à expliquer), et consiste à la représenter comme la combinaison linéaire d'un ensemble de prédicteurs (variables explicatives).

Dans ce mémoire la variable à expliquer Y sera, soit

i. Discrète : estimation du nombre de réalisations d'un événement.

ii. Continue : modélisation du coût moyen d'un sinistre ou de la prime pure.

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

1. $(Y_1, \cdots, Y_n)$ définit une famille de variables aléatoires indépendantes qui suivent une distribution appartenant à la famille exponentielle.
2. Les prédicteurs $(X_1, \cdots, X_p)$ correspondent aux composants déterministes du modèle sous la forme de combinaison linéaire.
3. Pour tout $i \in \{1, \cdots, n\}$ la loi de Y_i est supposée appartenir à une famille de distributions dont les paramètres dépendent des variables explicatives à travers une fonction de lien, g, strictement monotone.

La détermination des coefficients d'un modèle GLM à p variables explicatives consiste à rechercher les coefficients $(\beta_0, \cdots, \beta_p)$ tels que pour tout $i \in \{1, \cdots, n\}$,

$$g(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \tag{3.1}$$

avec $x_{i,j}$ la valeur pour l'individu i de la variable explicative j. Le premier membre de la formule est la composante aléatoire. Le second membre, composé de la combinaison linéaire des variables explicatives, est la composante déterministe.

Fonctions de liens

Les résultats rendus par l'ajustement d'un modèle GLM dépendent de la fonction de lien employée. Lors de cette étude, la fonction de lien utilisée est définit par,

$$\begin{aligned} g :]0, 1] &\to \mathbb{R} \\ g(x) &= \ln(x). \end{aligned}$$

Ainsi le modèle sera multiplicatif et s'utilisera de la façon suivante,

$$\ln(E[Y_i]) = \beta_0 + \sum_{i=1}^{p} \beta_j x_{i,j} \iff E[Y_i] = \exp(\beta_0 + \sum_{i=1}^{p} \beta_j x_{i,j}) \tag{3.2}$$

$$\tag{3.3}$$

On a alors,

$$E[Y_i] = \exp(\beta_0) \cdot \exp(\beta_1 \cdot x_{i,1}) \cdots \exp(\beta_p \cdot x_{i,p}). \tag{3.4}$$

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

Dans ce qui suit on présente la forme générale des distributions exponentielles ainsi que des distributions particulières de cette famille et les distributions de tweedie.

Famille exponentielle

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

$$f_{\theta,\phi} = \exp\left\{\frac{y\theta - b\theta}{a(\phi)} + c(y, \phi)\right\}, \quad y \in \mathbb{R}, \tag{3.5}$$

avec,

1. θ le paramètre de la moyenne,
2. ϕ le paramètre de dispersion lié à la variance,
3. a une fonction définie sur $\mathbb{R}$ non nulle,
4. b une fonction définie sur $\mathbb{R}$ au moins deux fois dérivable, avec une dérivée seconde positive,
5. c une fonction définie sur $\mathbb{R}^2$.

La moyenne et la variance d'une variable aléatoire dont la densité est de la forme exponentielle sont définies de la façon suivante,

$$E[Y] = b'(\theta) \quad \text{et} \quad \text{Var}(Y) = b''(\theta) \cdot a(\phi).$$

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

$$\begin{aligned} \theta &= g_c(\mu) \\ &= g_c(E[Y]) \\ &= g_c(g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)) \end{aligned}$$

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

$$\theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

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

3.1.2 Estimation des coefficients par la méthode du maximum de vraisemblance

Notons $(y_1, \cdots, y_n)$ un échantillon aléatoire de taille n inépendantes et identiquement distribuées, où leur loi appartient à la famille exponentielle.

Alors la fonction de vraisemblance peut s'écrire de la forme suivant,

$$L(\theta, \phi, y_1, \cdots, y_n)) = \prod_{i=1}^{n} \left[\frac{y_i \theta i - b(\theta i)}{a(\phi)} + c(y_i, \phi) \right]$$

Avec le paramètre canonique θ inconnu, le paramètre de dispersion ϕ supposé connu et,

$$\begin{aligned} E[Y] = b'(\theta) &= \frac{\partial b(\theta)}{\partial \theta} \\ \text{Var}(Y) = b''(\theta) \cdot a(\phi), \quad &\text{où} \quad b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}. \end{aligned}$$

Il faut se souvenir que le paramètre θ est une fonction des coefficients $\beta = (\beta_1, \cdots, \beta_p)$, ainsi l'estimation des coefficients du GLM se fait en cherchant les $\hat{\beta}$ qui maximisent la vraisemblance, c'est-à-dire qu'ils vérifient les conditions suivantes :

$$\frac{\partial L(\theta, \phi, y_1, \cdots, y_n)}{\partial \beta} = 0 \quad \text{et} \quad \frac{\partial^2 L(\theta, \phi, y_1, \cdots, y_n)}{\partial \beta^2} < 0.$$

Pour la plupart des modèles linéaires généralisés, les équations qui déterminent les paramètres au sens du maximum de vraisemblance sont non linéaires et les estimateurs n'ont pas d'autres expressions formulables que comme solutions de ces équations.

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

3.1.3 Distributions

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

La loi Normale

La loi Normale est une loi de probabilité absolument continue qui dépend de deux paramètres : son espérance, un nombre réel noté μ, et son écart type, un nombre réel positif noté σ. Fixons donc un couple (μ, σ) avec $\mu \in \mathbb{R}$, $\sigma > 0$, et considérons une variable aléatoire Y suivant une loi Normale $N(\mu, \sigma^2)$. La densité de Y est alors,

$$\varphi_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad y \in \mathbb{R}.$$

Cette densité peut être écrite sous la forme générale de la densité d'une loi de la famille des distributions exponentielles **(3.5)**. Soit $\mu \in \mathbb{R}$ et $\sigma > 0$, on peut écrire,

$$\begin{aligned}\varphi_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) &\iff \varphi_{\mu,\sigma^2}(y) = \exp\left(-\frac{\ln(2\pi\sigma^2)}{2} - \frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right) \\ &\iff \varphi_{\mu,\sigma^2}(y) = \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)}{2}\right)\end{aligned}$$

Notons $\theta = \mu$ et $\phi = \sigma$, alors par identification avec la formule **(3.5)** on détermine les fonctions suivantes,

$$a(\phi) = \phi^2 \quad b(\theta) = \frac{\theta^2}{2} \quad c(y, \phi) = -\frac{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)}{2}$$

La moyenne et la variance d'une variable Y suivant une loi Normale sont,

$$E[Y] = \mu \quad \text{et} \quad \text{Var} = \sigma^2$$

La loi de Poisson

La loi de Poisson est une loi discrète dépendant d'un paramètre d'intensité, un nombre réel noté λ. Soit Y une variable aléatoire suivant une loi de poisson de pararmètre réel positif λ, $Y \sim P(\lambda)$. La loi de Y est alors :

$$\forall k \in \mathbb{N}, \qquad P(Y = k) = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

Cette loi peut être écrite de manière différente pour identifier les fonctions présentes dans la formule **(3.5)**. Soit $k \in \mathbb{N}$,

$$\begin{aligned} P(Y = k) = \exp(-\lambda)\frac{\lambda^k}{k!} &\iff P(Y = k) = \exp\left(-\lambda + \ln\left(\frac{\lambda^k}{k!}\right)\right) \\ &\iff P(Y = k) = \exp(k\ln(\lambda) - \lambda - \ln(k!)). \end{aligned}$$

Notons $\theta = \ln(\lambda)$ et $\phi = 1$, alors par identification avec la formule **(3.5)** on détermine les fonctions suivantes,

$$a(\phi) = 1 \quad b(\theta) = \exp(\theta) = \lambda \quad c(k, \phi) = -\ln(k!).$$

La loi de Poisson a la particularité d'avoir une moyenne et une variance qui sont égales,

$$E[Y] = \text{Var}(Y) = \lambda.$$

La loi Gamma

Pour pouvoir définir la loi Gamma ils nous faut d'abord rappeler que la fonction gamma est une fonction définie sur l'ensemble des réels positifs, dont l'expression est la suivante :

$$\forall \alpha > 0, \qquad \Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-1}dx.$$

La loi Gamma est une loi de probabilité continue dépendant de deux paramètres, deux nombres réels strictement positifs α et β qui vont respectivement affecter la forme et l'échelle de sa représentation graphique. Soit Y une variable aléatoire suivant une loi gamma de paramètres α et β, strictement positifs, ce que l'on note $Y \sim \Gamma(\alpha, \beta)$. La densité de Y est alors,

$$\gamma_{\alpha,\beta}(y) = \frac{\beta^\alpha}{\gamma(\alpha)}y^{\alpha-1}\exp(-\beta y), \quad y > 0.$$

Il est important de noter qu'une variable aléatoire Y suivant une loi Gamma ne prend que des valeurs positives.

La fonction de densité de la loi Gamma peut être présentée sous la forme générale d'une distribution de la famille des lois exponentielles **(3.5)**. Soit $\alpha > 0$ et $\beta > 0$,

$$\begin{aligned} \gamma_{\alpha,\beta}(y) &= \frac{\beta^\alpha}{\gamma(\alpha)}y^{\alpha-1}\exp(-\beta y) \\ \iff \gamma_{\alpha,\beta}(y) &= \exp\left(\frac{y(-\frac{\beta}{\alpha}) - (-\ln(\beta))}{\frac{1}{\alpha}} + (\alpha - 1)\ln(y) - \ln(\Gamma(\alpha))\right) \end{aligned}$$

Notons $\theta = -\frac{\beta}{\alpha}$ et $\phi = \frac{1}{\alpha}$, alors par identification avec la formule **(3.5)** on détermine les fonctions suivantes,

$$a(\phi) = \phi \quad b(\theta) = -\ln(-\alpha\theta) \quad c(y,\phi) = (\alpha - 1)\ln(y) - \ln(\Gamma(\alpha))$$

La moyenne et la variance d'une variable aléatoire Y suivant une loi gamma $\Gamma(\alpha, \beta)$ sont les suivantes :

$$E[Y] = \frac{\alpha}{\beta} \quad \text{et} \quad \text{Var}(Y) = \frac{\alpha}{\beta^2}$$

Les distributions de Tweedie

La modélisation de la consommation par assuré d'un contrat en assurance nécessite la prise en compte des contrats ne présentant pas de sinistres pour la période considérée par l'étude. Cependant les lois que nous avons vues précédemment ne permettent pas de le faire car elles ne possèdent pas de masse de probabilité en zéro.

Un moyen pour estimer la prime pure d'un contrat d'assurance de façon directe est d'utiliser une distribution de Tweedie. En effet, cette distribution présente la particularité d'avoir une masse de probabilité (mesure de Dirac) en zéro.

Prenons une variable aléatoire Y suivant une distribution de Tweedie. Notons $\mu \in \mathbb{R}$ l'espérance de Y, $p \geq 0$ le paramètre de forme de la distribution de Tweedie et ϕ le paramètre de dispersion. La densité d'une distribution de Tweedie peut alors être exprimée sous la forme,

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

avec,

$$\begin{aligned} \theta &= \frac{\mu^{1-p}}{1-p} \\ a(\phi) &= \phi \\ b(\theta) &= \frac{\mu^{2-p}}{2-p} \\ c(y,\phi) &= \ln\left(\sum_{n=1}^{+\infty} \frac{\beta^{n\alpha}\lambda^n y^{n\alpha} - 1}{\gamma(n\alpha)n!}\right) \end{aligned}$$

où,

$$\beta = \frac{1}{\phi(1-p)} \quad \text{et} \quad \lambda = \frac{1}{\phi(2-p)}.$$

Soit Y une variable aléatoire suivant une distribution de Tweedie. Notons μ l'espérance de cette loi et soit $p \geq 0$ le paramètre de forme de la distribution. La variance de Y est,

$$\text{Var}(Y) = a(\phi) \cdot \mu^p.$$

Ainsi pour des valeurs particulières de p on retrouve les loi classiques utilisées pour l'ajustement de modèles GLM.

Dans le cas où $p = 0$, on retrouve la loi Normale. En effet soit $Y \sim N(\mu, \sigma^2)$, alors,

$$\text{Var}(Y) = a(\phi) \cdot \mu^p = a(\phi) = \sigma^2.$$

Dans le cas où $p = 1$ et $a(\phi) = 1$, on retrouve la loi de Poisson. En effet soit $Y \sim P(\lambda)$, alors,

$$\text{Var}(Y) = a(\phi) \cdot \mu^p = \mu = \lambda.$$

Dans le cas où $p = 2$, on retrouve la loi Gamma. En effet soit $Y \sim \Gamma(\alpha, \beta)$, alors,

$$\text{Var}(Y) = a(\phi) \cdot \mu^p = \frac{1}{\alpha} \times \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

La valeur du paramètre de forme p va donc déterminer quelle sera la forme de la distribution. Lorsque la valeur de p tend vers 1, le comportement de la distribution de Tweedie se rapproche de celui de la loi de Poisson. Lorsque la valeur de p tend vers 2, le comportement de la distribution de Tweedie se rapproche de celui de la loi Gamma. La valeur limite entre les deux comportements est $p = 1,5$.

3.2 Modélisation de la prime pure

En assurance, il faut être en mesure de quantifier le risque porté par l'assureur au titre des garanties, donc, pour une bonne compréhension des coûts d'un portefeuille il est essentiel de pouvoir évaluer le comportement d'une somme de variables aléatoires. L'espérance des coûts moyens des contrats d'assurance correspond à la prime pure. dans cette situation, on souhaite estimer la consommation moyenne des frais médicaux en santé,

$$\Pi = E\left(\frac{S}{n}\right).$$

Soit S la somme des coûts totaux de n assurés, c'est bien une variable aléatoire.

$$S = \sum_{i=1}^{n} X_i$$

Avec X_i la charge annuelle de l'assuré i. Le vecteur $X = (X_1, \cdots, X_n)$ est considéré comme une suite de variables indépendantes et identiquement distribuées.

Estimer cette consommation revient à minimiser l'écart quadratique moyen entre la charge totale des sinistres et la prime pure hypothétique payée en moyenne par chaque assuré. Ce résultat doit permettre à l'assureur de rembourser l'intégralité des frais engagés lors de la période de couverture.

$$\text{Min}\left\{E\left[\left(\frac{S}{n} - \Pi\right)^2\right]\right\} \longrightarrow \Pi = E\left(\frac{S}{n}\right).$$

Aujourd'hui la prime pure de référence qui est utilisée pour le calcul du tarif d'un produit proposé pour une affaire nouvelle a l'aspect général suivant :

$$\begin{aligned}\Pi_{\text{pure}} = \Pi_{\text{reference}} \times \text{Coeff}_{\text{age}} \times \text{Coeff}_{\text{regime}} \times \text{Coeff}_{\text{formule}} \\ \times \text{Coeff}_{\text{Zone}} \times \text{Coeff}_{\text{autres}}.\end{aligned}$$

Cette prime pure actuelle a été estimée grâce au méthode du GLM, mais avec comme objectif la modélisation de la consommation générale pour un produit santé. La sinistralité de tous les postes de garantie confondus a été utilisée pour ce calcul, donnant en conséquence une connaissance du risque assez large.

Une nouvelle approche par poste de garantie sera proposée pour mieux concevoir le risque santé. Grâce à cette nouvelle connaissance, il sera envisageable de proposer dans le futur un produit adaptable au réel besoin de l'assuré.

Idée : L'objectif principal est la modélisation de la consommation moyenne des assurés, donc,

$$\Pi_{\text{Global}} = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right).$$

Pour arriver à cet objectif, la décomposition suivante est proposée,

$$\Pi_{\text{Global}} = \Pi_{\text{Hospitalisation}} + \Pi_{\text{Soins Courants}} + \Pi_{\text{Pharmacie}} + \Pi_{\text{Dentaire}} + \Pi_{\text{Optique}} + \Pi_{\text{autres}}.$$

Cette décomposition est possible grâce à la linéarité de l'espérance mathématique, une de ses propriétés élémentaire.

3.2.1 Type de modélisation selon les caractéristiques du poste

Suite à l'analyse du portefeuille, il est évident que le risque porté en santé ne dépend pas seulement du niveau de garantie proposé. En effet, à l'intérieur d'une même formule les principaux postes de santé : Hospitalisation, Soins courant, Pharmacie, Dentaire et optique présentent des couvertures différents et, en concéquence, des niveaux de consommation non homogènes. Il est donc important de réaliser l'étude séparément par poste pour une meilleur modélisation.

L'estimation de la prime pure est possible par différentes méthodes. Selon la nature de la consommation de chaque poste, on peut réaliser soit une modélisation directe de la consommation moyenne soit une modélisation en décomposant la fréquence et le coût moyen des sinistres.

Modèle de fréquence coût moyen

Pour un assuré i, sa consommation annuelle totale X_i est engendrée par la somme des coûts des sinistres survenus de manière indépendante pendant une période fixée. Il est donc possible de décomposer la charge de sinistres totale S de la façon suivante,

$$S = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \sum_{j=1}^{K_i} Y_{i,j} = \sum_{l=1}^{N_S} Y_l.$$

Avec N_S une variable aléatoire définie comme le nombre total de sinistres, tous assurés confondus, alors N_S correspond à la somme du nombre de sinistres totaux de chaque assuré :

$$N_S = \sum_{i=1}^{n} K_i.$$

Et Y_l est la variable aléatoire correspondant au coût moyen lié à chaque sinistre.

On suppose que les variables sont indépendantes entre elles pour tout i et tout j en plus d'être identiquement distribués sachant K_i et que les variables K_i sont identiquement distribuées et indépendantes entre elles pour tout i. Calculons l'espérance de $S = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i\right) &= E\left(\sum_{l=1}^{N_S} Y_l\right) \\ &= E\left(E\left(\sum_{l=1}^{N_S} Y_l | N_S\right)\right) \\ &= E\left(N_S E(Y|N_S)\right) \\ &= E(N_S E(Y)), \end{aligned}$$

et par l'hypothèse d'indépendance on obtient,

$$E\left(\sum_{i=1}^{n} X_i\right) = E(N_S)E(Y).$$

La prime pure est donc le produit de la fréquence probable par le coût probable d'une consommation.

$$\Pi = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = E(N_S/n)E(Y).$$

Liaison entre la fréquence et le coût moyen par poste

Calcul de corrélation de Pearson.
Relation et dépendance : Soit (X, Y) un couple de caractères quantitatifs, décrivant un même ensemble. On dit qu'il existe une *relation* entre X et Y si l'attribution des modalités sur l'une génère un effet sur l'autre. De cette façon le fait d'avoir connaissance des valeurs de X permet de prédire, dans une certaine mesure, les valeurs de Y.

Les coefficients de corrélation constituent une mesure de l'intensité de liaison entre deux variables, voici la formule qui représente le coefficient de corrélation de Pearson,

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Le calcul du coefficient de corrélation entre la fréquence et le coût moyen a été fait conditionnellement aux variables tarifaires (âge, département, régime, formule).

Postes	Corrélation de Pearson
Hospitalisation	0,153
Soins courants	0,048
Pharmacie	0,314
Dentaire	0,250
Optique	0,463

TABLE 3.1 – Coefficient de corrélation par poste

Notons que les postes Pharmacie, Dentaire et Optique ont une grande force de liaison entre la fréquence d'un sinistre et son coût moyen : sur ces trois postes plus la survenance d'un sinistre est fréquente plus son coût moyen est importante. Les postes Hospitalisation

et Soins courants présentent un coefficient assez proches de 0 donc une force de liaison presque nulle.

L'analyse graphique est une bonne façon de mieux identifier les différents caractéristiques de relation entre deux variables est l'analyse graphique. Sur les graphiques suivants, on observe l'existence d'une relation ou non entre la fréquence de survenance d'une sinistre et de son coût moyen associé. Pour illustrer cette situation, ci-dessous sont présentés les graphiques concernant les données du poste Hospitalisation et du poste Optique, ces deux postes ont été choisis pour comparer et identifier des différences à l'intérieur du poste et aussi entre les postes.

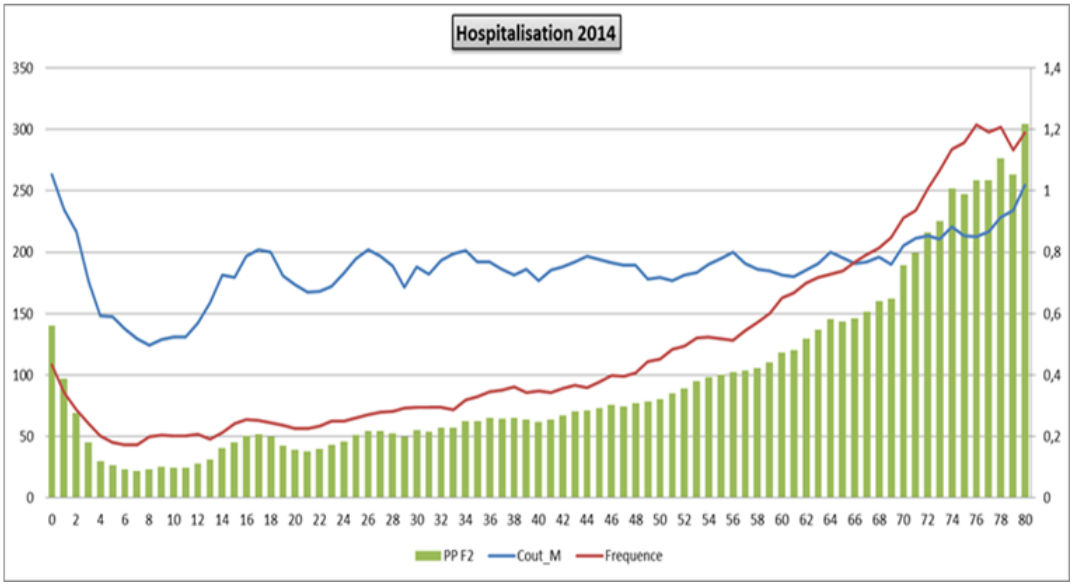


FIGURE 3.1 – Répartition de la Prime Pure, de la Fréquenc et du Coût moyen, en fonction de l'âge

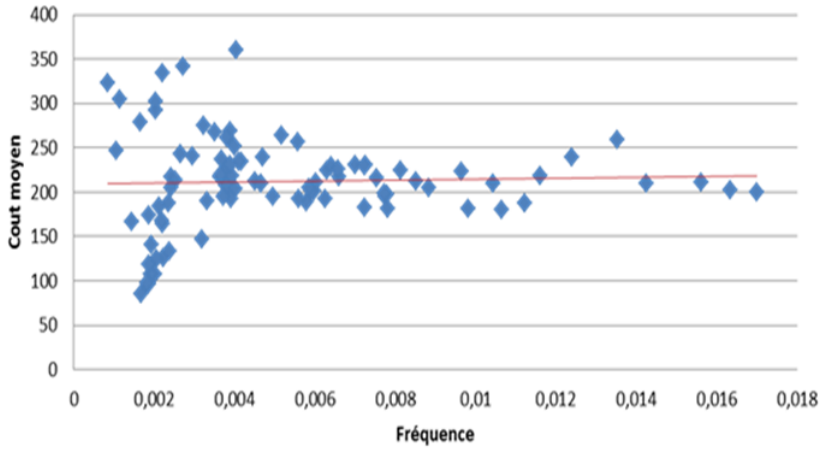


FIGURE 3.2 – Nuage de points Coût moyen Vs Fréquenc du poste Hospitalisation

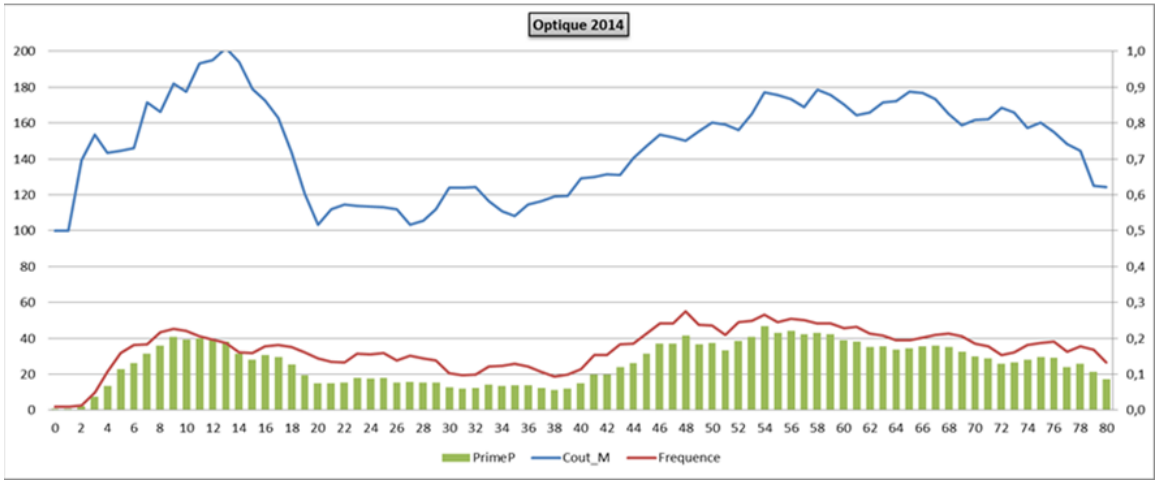


FIGURE 3.3 – Répartition de la Prime Pure, de la Fréquenc et du Coût moyen, en fonction de l'âge

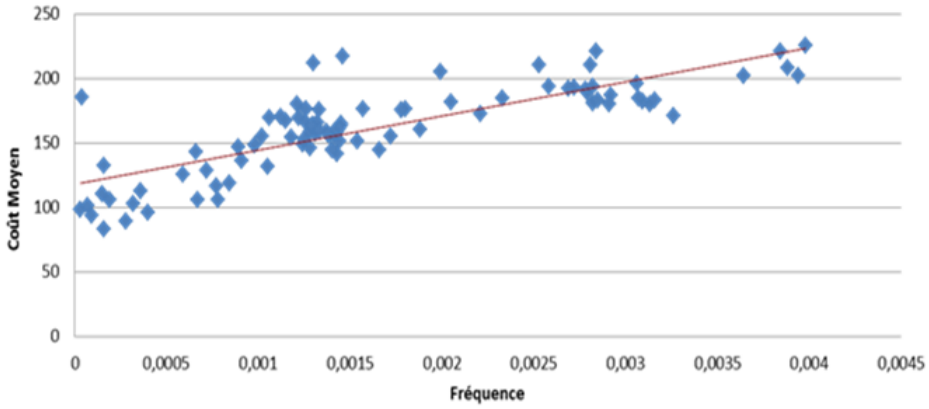


FIGURE 3.4 – Nuage de points Coût moyen Vs Fréquenc du poste Optique

Sur le poste hospitalisation la prime pure en fonction de l'âge est notamment expliquée par la fréquence de la sinistralité : plus un bénéficiaire est âgé plus il a de risque d'être hospitalisé. Cependant le coût moyen d'un sinistre ne révèle pas une tendance en fonction de l'âge de l'assuré. De l'autre côté, les graphiques concernant le poste Optique indiquent clairement la présence de dépendance entre la fréquence et le coût moyen, le nuage de point met en évidence la tendance positive entre eux.

Les analyses de liaison ne sont pas suffisantes pour conclure sur l'indépendance entre la fréquence et le coût moyen d'un sinistre. Cette information est importante pour mieux comprendre le comportement de la sinistralité propre de chaque poste et par la suite arriver à appliquer la modélisation la plus adaptée.

Test d'indépendace de χ^2

Le test de χ^2 vient apporter de l'information supplémentaire, ce test permet de contrôler l'indépendance de deux variables au sein d'un même population. On souhaite vérifier

l'existence ou non d'une dépendance entre la fréquence et le coût moyen. Lorsque les variables sont quantitatives il est nécessaire de regrouper les valeurs, pour cela on doit calculer l'expression présentée ci-après,

Soit (X, Y) une couple des variables quantitatives, on doit placer leurs valeurs dans des intervalles bien répartis, Sous l'hypothèse d'indépendance, la distribution conjointe

		Variable 2			
		Valeur (ou intervalle) 1	Valeur (ou intervalle) 2	Valeur (ou intervalle) 3	
Variable 1	Valeur (ou intervalle) 1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,\cdot}$
	Valeur (ou intervalle) 2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,\cdot}$
		$n_{\cdot,1}$	$n_{\cdot,2}$	$n_{\cdot,3}$	$n_{\cdot,\cdot}$

est donnée par le produit des distributions marginales, c'est-à-dire, $f_{i,j} = f_{i,\cdot} f_{\cdot,j}$, on peut estimer $f_{i,\cdot}$ par $\frac{n_{i,\cdot}}{n_{\cdot,\cdot}}$ et $f_{\cdot,j}$ par $\frac{n_{\cdot,j}}{n_{\cdot,\cdot}}$ de cette manière $n_{i,j} \approx \frac{n_{i,\cdot} n_{\cdot,j}}{n_{\cdot\cdot}}$.

On cherche à calculer l'écart entre les $n_{i,j}$ observés, notés $O_{i,j}$, et les prédis, notés $E_{i,j}$, l'hypothèse d'indépendance sera rejetée si l'écart est trop important.

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{i,j} - n_{i,\cdot} n_{\cdot,j}/n_{\cdot,\cdot})^2}{\frac{n_{i,\cdot} n_{\cdot,j}}{n_{\cdot,\cdot}}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

La statistique Q suit approximativement la loi $\chi^2_{(r-1)(c-1)}$ où r désigne le degré de liberté associé à la variable X et c désigne le degré de liberté associé à la variable Y, elles correspondent au nombre d'intervalles définis sur les variables respectives.

Les résultats obtenus suite à l'application du test sont assez proches des hypothèses conçues à partir de l'analyse des corrélations. Ces calculs ont été faits conditionnellement aux variables tarifaires. A partir des analyses réalisées on est arrivé aux conclusions suivants.

Postes	H_0 := *Fréquence indépendante du Coût moyen*
Hospitalisation	Pas de rejet de H_0 $(*)$
Soins courants	Pas de rejet de H_0
Pharmacie	Rejet de H_0
Dentaire	Rejet de H_0
Optique	Rejet de H_0

TABLE 3.2 – Coefficient de corrélation par poste

$(*)$ Le poste Hospitalisation n'a pas rejetté l'hyphotèse nulle pour toutes les variables tarifaires, sauf sur le test réalisé conditionnellement à la variable Département.

Ces derniers résultats sont fondamentaux pour bien choisir la manière dont on va procéder afin de modéliser la prime pure de façon la plus adaptée. Le tableau suivant résume les types de modélisation à faire sur chaque poste.

Poste|modèle	Fréquence×Coût moyen	Consommation moyenne
Hospitalisation	**X**	**X**
Soins Courants	**X**	
Pharmacie		**X**
Dentaire		**X**
Optique		**X**

TABLE 3.3 – Sélection du type de modélisation par poste

Affectation des lois selon le modèle

— Modèle de la Consommation moyenne : Distribution de Tweedie

$$\Pi = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right)$$

Sous l'hyphotèse d'indépendance,

$$\Pi = E(N_S/n)E(Y).$$

— Modèle de Fréquence : Distribution Poisson
— Modèle de Coût moyen : Distribution Gamma

Les lois de distribution Poisson et Gamma sont couramment utilisées en assurance pour l'estimation du nombre et du coût moyen des sinistres.
Pour une information plus détaillée le lecteur pourra se diriger vers l'annexe « Ajustement des lois ».

3.2.2 Étude de liaison des variables explicatives

Rappelons que les variables explicatives sont des variables telles que leurs valeurs correspondent à des catégories discrètes mutuellement exclusives, mais qu'il n'existe pas de relation d'ordre entre ces modalités. On ne peut donc pas les classer entre elles. Tout ce que l'on peut faire c'est compter combien d'individus tombent dans chaque modalité.

Pour étudier la liaison entre les variables explicatives, on va donc mesurer la corrélation entre toutes les variables. Ce sera un support pour la sélection des variables, car dans le cas où deux variables sont liées avec une corrélation significative, on pourra choisir la variable la plus significative et/ou la plus cohérente en fonction aussi des connaissances a priori.

- L'indice V de Cramer :

$$V = \left(\frac{D^2}{N \cdot \inf\{(r-1);(c-1)\}}\right)^{\frac{1}{2}}, \quad 0 \leq V \leq 1$$

avec,

— N : effectif total (N=somme des cellules de la table de contingence formée par les deux variables étudiées).
— r : nombre de lignes de la table de contingence.

— c : nombre de colonnes de la table de contingence.
— $D^2 = \sum_{i=0}^{r} \sum_{j=1}^{c} \frac{(E_{i,j-O_{i,j}})^2}{E_{i,j}}$, où,

$$E_{i,j} = \frac{(\sum_{n_c=1}^{c} O_{i,n_c}) \cdot (\sum_{n_r}^{r} O_{n_r,j})}{N}$$

avec, $E_{i,j}$ la fréquence théorique en cas d'indépendance des deux variables et $O_{i,j}$ la fréquence observée.

Contrairement au χ^2, les résultats du V de Cramer peuvent être comparés d'un tableau de contingence à l'autre. Nos corrélations sont donc calculées suivant la mesure donnée par le V de Cramer.

L'intérêt de la connaissance des variables corrélées

Si deux variables sont fortement corrélées, une partie de la prime pure estimée sera expliquée par ces deux variables à la fois. Il faudra donc supprimer l'un de ces deux critères pour ne pas se retrouver face à un double effet ou bien dans le cas où les effets ne seraient pas répartis de manière constructives ou cohérentes.

S'il existe des corrélations importantes entre certaines variables explicatives, le calcul de l'inverse de la matrice des produits croisés des variables devient difficile, voire impossible. En effet, une corrélation haute amène à inverser en quelque sorte des nombres très proches de 0, ce qui donne des termes tendant vers l'infini, ou du moins très grands. L'existence de ce problème entraînera des intervalles de confiance plus larges pour les paramètres estimés, car ils sont proportionnels aux éléments de cette matrice inverse, menant ainsi à une perte de précision dans la modélisation.

Tout d'abord, il est important de vérifier que les variables du portefeuille ne sont pas liées entre elles. La plupart d'entre elles sont la base de la tarification des produits considérés, donc l'existence de corrélation pourrait générer des coefficients estimés avec une faible précision. Dans notre étude, les variables du portefeuille sont faiblement liées entre elles.

V de Cramer	Année	Sexe	Age	Régime	Zone	Nb. de bénéf.	Formule
Année (4)	1						
Sexe (2)	0,011	1					
Age (106)	0,053	0,069	1				
Régime (4)	0,012	0,101	0,144	1			
Zone (5)	0,011	0,013	0,044	0,162	1		
Nb. de bénéf. (11)	0,028	0,094	0,219	0,048	0,031	1	
Formule (6)	0,059	0,018	0,077	0,175	0,084	0,035	1

TABLE 3.4 – Coefficient de corrélation Variables du Portefeuille

Ensuite, sur l'ensemble des variables présentes dans la modélisation, les seules corrélations significatives concernent, la **Variable Zone :** elle correspond au zonage actuel.

Cette variable est fortement corrélée à l'ensemble des variables externes. Ces résultats sont très importants car les variables externes viennent dans un premier temps remplacer les effet géographiques anciennement répresentés par les coefficients de cette variable « Zone ».

V de Cramer	Zone
Taux.spécialistes (16)	0,233
Densi.pharmacies (17)	0,467
Densi.etabl.sante (17)	0,412
Densi.laboratoires (17)	0,46
Densite (16)	0,468
DcAvant.65ans (16)	0,255
Infection.par.VIH (17)	0,506
Infec.invasives (17)	0,41
Toxi.infs.alim (17)	0,438
Total.maladies (17)	0,412
Benef.CMU.Comp (16)	0,392
Densi.opticiens (16)	0,314
Ind.vieil.65ans (8)	0,34
Ev.Nais.Hom (8)	0,438
Ev.Nais.Fem (8)	0,399
Ev.65ans.Hom (8)	0,455
Ev.65ans.Fem (8)	0,355
Tx.Pauvrete (8)	0,371
Cohab.coupenf (8)	0,64
Vivant.institi (8)	0,569

TABLE 3.5 – Coefficient de corrélation Variables externes

Les variables externes « Santé »ont été traitées dans le chapitre 2, partie 2.3. Les résultats des ACP ont montré de fortes liaisons entre elles. Suite à leur passage de variables quantitatives à des variables qualitatives puis à leur classification par k-means, le tableau de corrélation de V de cramer permet de voir que leur corrélation persiste.

En considérant l'importance des variables externes pour la modélisation du risque lié à la zone géographique, on est dans l'obligation de garder toutes ces variables pour l'étape suivante, même si certaines d'entre elles sont corrélées. A raison de ce qu'on est amené à faire plusieurs modèles sur les différents postes, il est impossible de connaître à l'avance les variables que nous devrons ou non garder avant la modélisation. Cependant nous devrons toujours garder à l'esprit les corrélations obtenues comme de l'information a priori pour une sélection définitive de variables significatives.

3.2.3 Sélection des variables explicatives

Sélection des variables explicatives tarifaires par la méthode de régression backward

Le risque [1] à modéliser est d'abord estimé pour toutes les variables tarifaires qui sont significativement explicatives. De cette manière on obtiendra le modèle de référence qui servira pour l'étape suivante.

Pour la création de ce modèle de référence, il faut alors conserver uniquement les variables tarifaires qui expliquent au mieux le risque, cela est possible grâce à une procédure de **sélection pas à pas**.

La sélection pas à pas consiste à effectuer des régressions successives afin de sélectionner les variables définitives du modèle. La méthode la plus classique est la régression **backward**, ou descendante. Elle procède par élimination successive de variables. A partir du modèle initial avec toutes les variables comprises, chaque variable est retirée une à une. Celle qui est la moins significativement explicative du risque est retirée du modèle. Le modèle de référence sans cette variable définit alors le nouveau modèle de référence, et l'opération est réitérée, jusqu'à ce que toutes les variables significatives soient conservées dans le modèle.

Critère de sélection

• L'analyse de la déviance, AIC et BIC

Il s'agit d'enlever une à une les variables en regardant à chaque fois la différence de déviance par rapport au modèle de référence. Si la déviance augmente, le modèle est moins bon.

$$D(\beta) = -2[\log \mathcal{L}(\hat{\beta}|Y) - \log \mathcal{L}_*(Y)]$$

Où $log\mathcal{L}(\beta|Y)$ désigne la log-vraisemblance du modèle et $log\mathcal{L}_*(Y)$ est la log-vraisemblance saturée obtenu avec le modèle « saturé », où le nombre des profils d'observations est égal au nombre de paramètres à estimer . Cependant, retirer une variable du modèle aboutit forcément à une perte d'information, donc implique une augmentation de la déviance.

D'autres critères d'information tels quels AIC et BIC existent. Il est possible d'augmenter la vraisemblance du modèle en ajoutant des paramètres, donc pour satisfaire le critère de parcimonie on a considéré aussi ces critères pour l'étude de la significativité du modèle. Ils permettent de pénaliser le modèle en fonction du nombre de paramètres, on choisit le modèle avec les deux critères les plus faibles.

$$\begin{cases} \text{AIC} &= -2\log\mathcal{L}(\hat{\beta}) + 2k \\ \text{BIC} &= -2\log\mathcal{L}(\hat{\beta}) + k\log(n) \end{cases}$$

• La statistique du χ^2

Pour la variable X :

Soient,

— $R = \frac{\text{Vraisemblance du modèle sans } X}{\text{Vraisemblance du modéle avec } X}$.

— $n = \dim(\text{modalités du modèle avec } X) - \dim(\text{modalités du modèle sans } X)$
$= \text{nombre de modalités de } X.$

— $s_{n,95\%}$ le seuil à 95 pour cent d'une χ^2 à n degrés de libertés.

1. Soit les primes pures, les coûts moyens et/ou les fréquences

Sous H_0 : la variable X n'est pas influente dans le modèle, la statistique $S = -2 \ln R$ suit asymptotiquement une loi de χ^2 à n degrés de liberté. Donc si $P[S \leq s_{n,95\%}] > 5\%$, alors les deux vraisemblances sont proches et X n'a pas d'apport significatif dans le modèle, et H_0 est vraie. Ce test permet de fixer une règle pour répondre à la question de significativité.
On part d'un modèle de référence qui contient toutes les variables tarifaires [2] comme variables explicatives. Ensuite on réalise des tests en enlevant chaque fois une variable et on compare avec le modèle de référence pour mesurer leur significativité.

Résultats du poste Hospitalisation

Hospitalisation	Enlever l'âge	Enlever le regime	Enlever le nb de bénéficiaires	Enlever la formule
P-valeur test χ^2	$0,0\%$	$0,0\%$	$0,0\%$	$0,0\%$
Variation Déviance	(+) 448.684,8	(+) 100.079,2	(+) 83.095,2	(+) 239.620,7
Variation AIC	(+) 448.672,8	(+) 100.073,2	(+) 83.089,2	(+) 239.610,7
Variation BIC	(+) 448.604,0	(+) 100.038,8	(+) 83.054,8	(+) 239.553,4

TABLE 3.6 – Tableau de l'apport de chaque variable tarifaire dans le modèle du poste Hospitalisation

Remarque : On ne rentrera pas dans l'explication des valeurs des variations, on se limitera à leur signe et à leur ordre de grandeur. Le signe positif de la variation indique la perte d'information en enlevant la variable, l'ordre de grandeur de la variation représente la quantité d'informations. Par exemple, enlever la variable « âge »du modèle principal génère des variations plus importantes, donc une perte d'information plus importante.

Le test de χ^2 est validé à chaque fois, chaque variable est significative dans le modèle. De plus, on constate qu'en enlevant une variable au modèle, la déviance, AIC et le BIC augmentent. En conséquence il n'est pas judicieux d'enlever une de ces variables du modèle de référence. Cependant, il ne faut pas tirer des conclusions uniquement à partir des tests mathématiques comme le montre l'exemple suivant.

Résultats du poste Optique

Optique	Enlever l'âge	Enlever le regime	Enlever le nb de bénéficiaires	Enlever la formule
P-valeur test χ^2	$0,0\%$	$0,0\%$	$0,0\%$	$0,0\%$
Variation Déviance	(+) 772.546,7	(+) 51.512,5	(+) 3.838,2	(+) 3.177.761,0
Variation AIC	(+) 772.532,7	(+) 51.506,5	(+) 3.818,2	(+) 3.177.749,0
Variation BIC	(+) 772.452,4	(+) 51.472,1	(+) 3.703,5	(+) 3.177.680,0

TABLE 3.7 – Tableau de l'apport de chaque variable tarifaire dans le modèle du poste Optique

D'après le test du χ^2 et les valeurs des variations, il n'est pas approprié d'enlever la variable « Nb. de bénéficiaires », pourtant, en faisant une analyse graphique, on constate que les informations apportées par cette variable sont incohérentes.

2. Variables propres au portefeuille

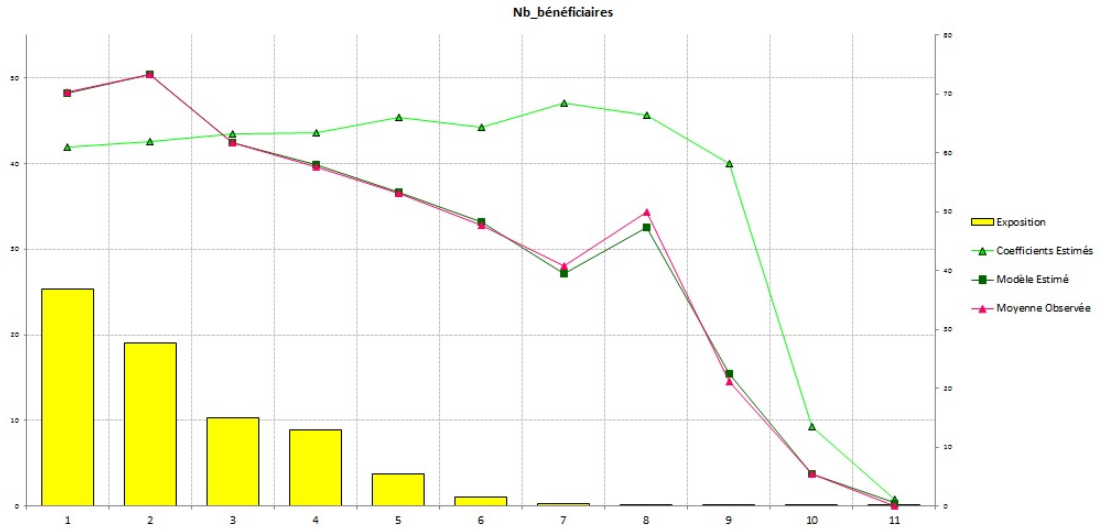


FIGURE 3.5 – Effet de la variable Nombre de bénéficiaires compris dans le modèle

On constate que sur les modalités bien exposées au risque (de 1 à 7), la tendance des coefficients estimés va à l'inverse de celle des primes pures observées, ainsi l'information isolée apportée par la variable « Nb. de bénéficiaires »(courbe verte claire) n'est pas cohérente.

Il faut sélectionner l'effet que l'on souhaite retenir, celui que l'on juge représentatif de la réalité. Dans l'exemple ci-dessus, on ne peut pas raisonnablement dire que l'effet isolé de la variable « Nb de bénéficiaires »sur la prime pure optique est croissant, car on observe l'effet contraire dans l'observé. On souhaite que les autres variables capturent et expliquent entièrement l'effet car elles sont plus cohérentes. C'est ce qu'on fait ici, on ne retient pas la variable « Nb de bénéficiaires », et on laisse les autres variables absorber totalement l'effet sur la prime pure optique. On préfère que les effets marginaux soient tous cohérents, d'autant plus qu'ils ont vocation à être utilisés dans la tarification.

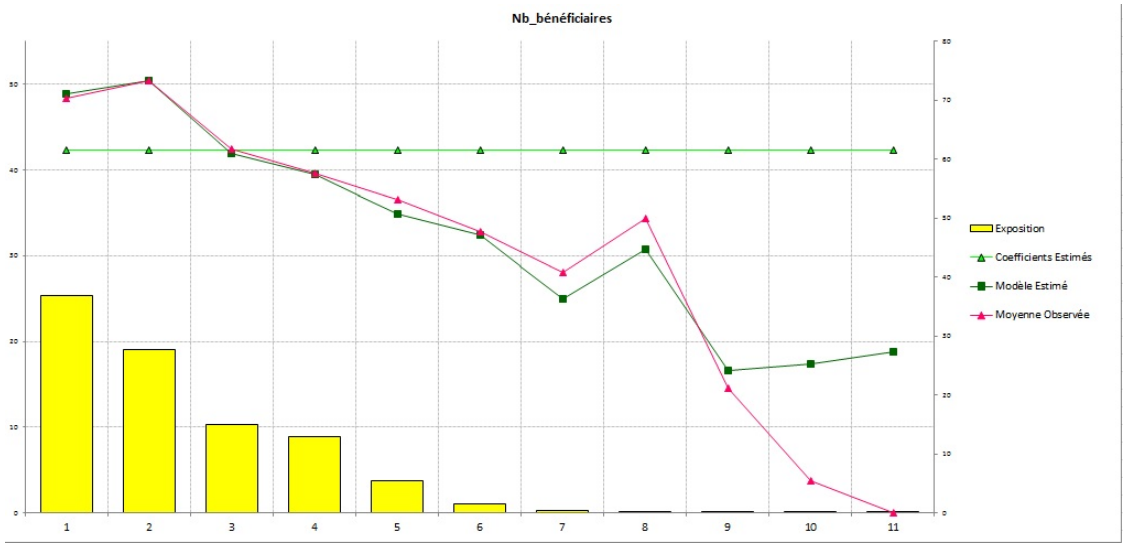


FIGURE 3.6 – Variable Nombre de bénéficiaires non compris dans le modèle

Malgré la suppression de cette variable, on constate que la moyenne estimée par le modèle (courbe vert foncé) reste proche de la moyenne observée (courbe rose) sur les modalités bien exposées au risque. Le tes de χ^2 doit donc bien être complété par une analyse graphique avant de prendre la désicion de garder ou supprimer une variable.

Sélection des variables explicatives externes par la méthode de régression forward

Cette méthode consiste à introduire les variables une par une. On part du modèle de référence obtenu par la méthode précédente avec les variables tarifaires significatives et on ajoute à chaque étape une variable externe. Les sommes des carrés des écarts du Modèle augmentent forcément et le principe est donc de faire entrer à chaque pas la variable qui apportera l'augmentation la plus significative, en utilisant aussi les mêmes critères de sélection déjà mentionés ci-dessus.

Illustration :

On a construit trois modèles différents pour le poste Hospitalisation : Modèle de Prime pure Tweedie, Modèle de Fréquence Poisson et Modèle de Coût moyen Gamma. On va comparer la significativité d'une même variable sur les différents modèles.
Pour chaque modèle, on part de celui de référence obtenu plus haut par la méthode Backward. Ensuite on crée un nouveau modèle en incorporant une à une les variables externes, de cette manière on pourra évaluer l'apport que génère chaque variable dans le modèle de référence. Les variables explicatives n'ont pas le même effet selon la quantité (prime pure, fréquence ou coût moyen) qu'on veut estimer. Dans la suite, comme exemple, on va présenter les résultats de l'insertion de la variable « densité Pharmacie »aux modèles.

Modèle de Fréquence

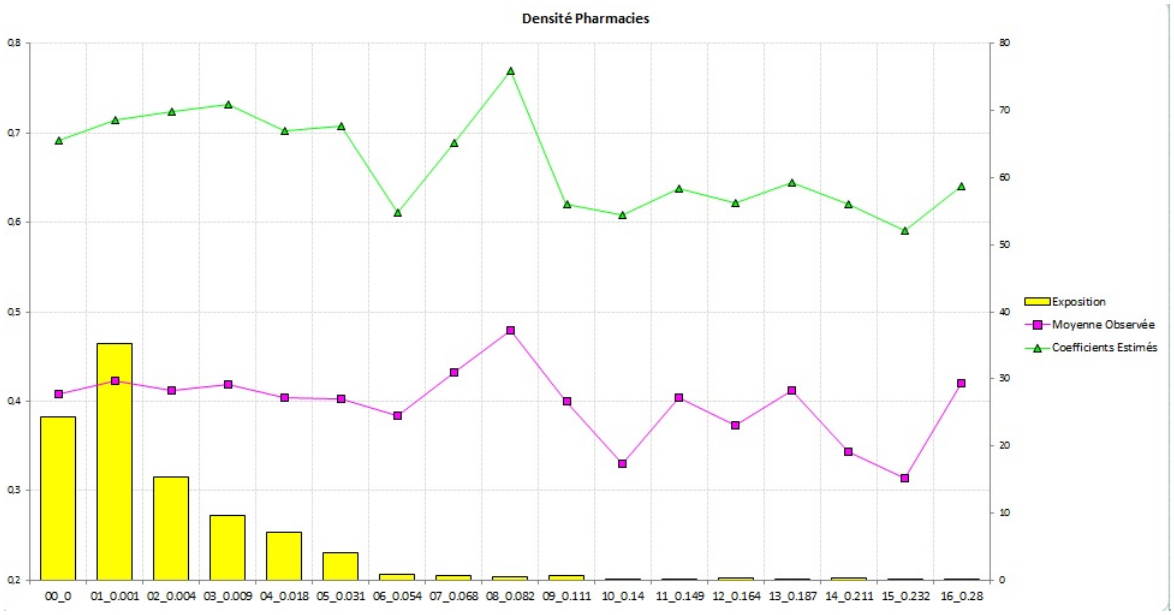


FIGURE 3.7 – Effet variable Densité des Pharmacies dans le modèle Fréquence, poste Hospitalisation

	Ajout de la variable dans le modèle
P-valeur test χ^2	$4,6\%$
Variation Déviance	(-) 172,6
Variation AIC	(-) 140,2
Variation BIC	(+) 43,3

TABLE 3.8 – Apport variable Densité des Pharmacies au modèle Fréquence, poste Hospitalisation

La p-valeur du test de χ^2 est proche de 5%, ainsi l'hypothèse nulle risque de ne pas être rejetée, c'est-à-dire la variable ne semble pas avoir une influence dans le modèle. Graphiquement, la variable n'offre pas un effet satisfaisant, on constate que sont estimation a une tendance croissante (en vert clair), sont effet marginal est donc opposé à ce qui est observé (décroissance en rose). En conséquence, la densité de pharmacies ne semble pas bien expliquer la fréquence en hospitalisation.

Modèle Coût moyen

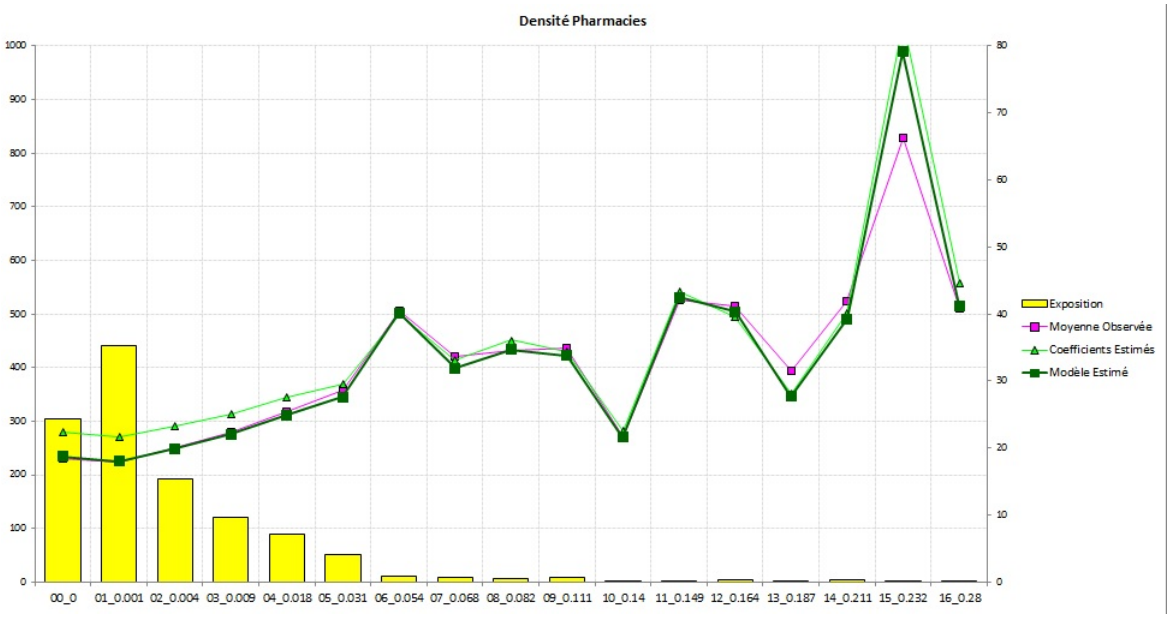


FIGURE 3.8 – Effet variable Densité des Pharmacies compris dans le modèle Coût moyen

	Ajout de la variable dans le modèle
P-valeur test ξ^2	$0,0\%$
Variation Déviance	(-) 3.204,1
Variation AIC	(-) 3.134,5
Variation BIC	(-) 2.983,6

TABLE 3.9 – Apport de la variable Densité des Pharmacies au modèle Coût moyen du poste Hospitalisation

La p-valeur indique que la variable est bien significative dans le modèle, de plus les variations de la déviance, AIC et BIC baissent de manière plus importante. Par ailleurs, le graphique montre que l'effet marginal de la variable est en ligne avec l'effet empirique.

Donc, la densité de pharmacies est une variable qui explique bien le coût moyen en hospitalisation.

Modèle Tweedie

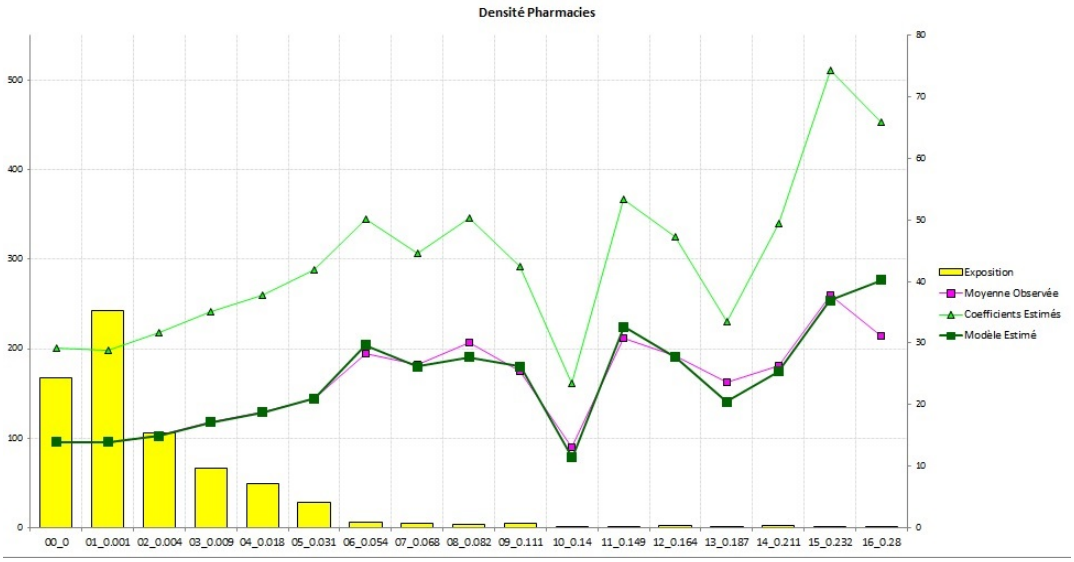


FIGURE 3.9 – Effet variable Densité des Pharmacies dans le modèle Prime Pure direct

	Ajout de la variable dans le modèle
P-valeur test χ^2	$0,0\%$
Variation Déviance	(-) 41.210,08
Variation AIC	(-) 41.178,08
Variation BIC	(-) 40.994,66

TABLE 3.10 – Apport de la variable Densité des Pharmacies au modèle Prime Pure direct

La p-valeur indique que la variable est significative. En parallèle le graphique montre que les effets sont très liés avec des pentes plus élevées, il y a donc une meilleure segmentation entre les coefficients. Si on compare les variations de déviance, AIC et BIC entre Fréquence VS Coût moyen VS Tweedie, on constate que dans ce dernier, la baisse par rapport au modèle de référence est plus significative. En conclusion, la densité de pharmacies est plus adaptée pour expliquer directement la prime pure du poste hospitalisation.

3.2.4 Lissage des coefficients

Afin de confirmer la fiabilité de l'estimation des coefficients, il est nécessaire de vérifier que les modalités sont suffisamment répresentées en terme d'exposition et que les estimations des modalités voisines sont significativement différentes les unes des autres, dans le cas contraire il sera pertinent de regrouper certaines modalités sans perte significative d'information du modèle.

Lorsque deux modalités voisines présentent des coefficients assez proches, il sera nécessaire d'étudier leurs significativité pour prendre, ou pas, la décision de les regrouper.

- **La significativité des coefficients des modalités** : Pour deux classes i et j, l'écart-type de la différence de ces deux classes est défini par :

$$\sqrt{V(\hat{\beta}_i - \hat{\beta}_j)} = \sqrt{V_{i,i} + V_{j,j} - 2V_{i,j}}$$

où $V = [V_{i,j}]$ matrice de variance/covariance.
La matrice V est estimée comme suit : si $V(Y) = \sigma^2\Sigma$, avec Σ connu, alors $V = (X\,{}^tX)^{-1}$, où X matrice des X_i, est de rang p.
Ici $\hat{\sigma}^2 \frac{||e||^2}{n-p}$ est l'estimateur des moindres carrés de σ^2, où e est le vecteur des erreurs résiduelles avec $e = Y - X^{\hat{\beta}}$.
On dispose du coefficient de variation de la différence entre les impacts de 2 modalités.

$$\text{Standard Error (\%)} = \frac{\sqrt{V(\hat{\beta}_i - \hat{\beta}_j)}}{\hat{\beta}_i - \hat{\beta}_j}$$

L'erreur standard (ou Standard Error) met en évidence les classes à rassembler. Cet indicateur a comme objectif de déterminer si la différence $\hat{\beta}_i - \hat{\beta}_j$ est véritablement significative. C'est pourquoi on regarde la variabilité de cette différence, qui cherche à voir si cette différence évolue beaucoup autour de sa moyenne.
Plus la variabilité de la différence est faible, plus l'impact des modalités est significative, c'est pourquoi les deux modalités doivent rester séparées dans le modèle. Au contraire si elle est trop élevée, les modalités i et j pourront raisonnablement être regroupées.

- **L'analyse graphique** : la représentation des coefficients doit être monotone. Dans le cas où les modalités sont suffisamment nombreuses (nombre arbitrairement fixé à 5), on cherche à ce que cette représentation suive une certaine tendance (linéaire, logarithmique, exponentielle, ...), et sinon, à ce que les valeurs prises par les coefficients soient distinctes. Cet outil permet de vérifier visuellement que les lissages effectués sont judicieux.

Les lissages sont effectués à partir de la moyenne des coefficients pondérée par l'exposition et/ou par l'ajustement par polynômes. De cette manière on pense que la réalité correspond à l'effet moyen et pas aux variations d'effets trop importantes entre modalités, ainsi l'intérêt de réaliser des lissages est de baisser la complexité du modéle et en conséquence améliorer la robustesse du modèle.

Le graphique ci-dessous représente l'intervalle de confiance des coefficients de la variable « Nb. de bénéficiaires » [3] compris dans modèle de la prime pure du poste hospitalisation.

3. Définition de la variable dans la section 2.3

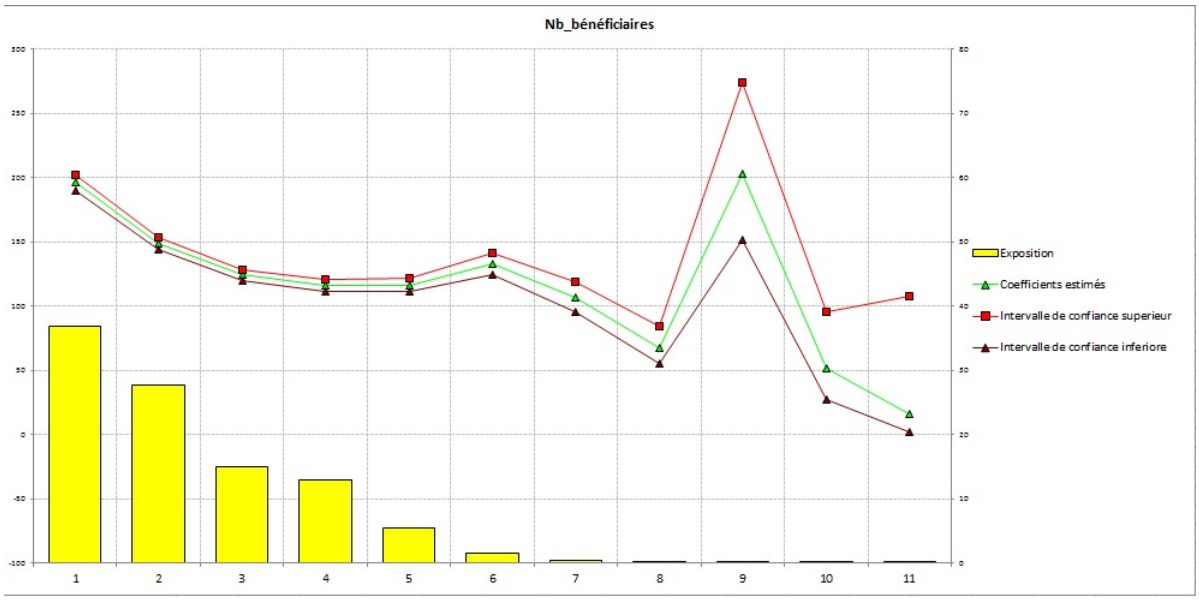


FIGURE 3.10 – Intervalle de Confiance des coefficients de la variable Nombre de bénéficiaires

Suite à la validation du test de significativité de la variable « nombre de bénéficiaires », on obtient l'estimation des coefficients par la méthode de maximum de vraisemblance. La ligne de couleur verte représente les coefficients estimés et les deux autres lignes correspondent à l'intervalle de confiance. Les contrats ayant entre un et quatre bénéficiaires sont les plus représentés et on constate que les résultats de l'estimation sont plus précis sur cette plage car l'intervalle de confiance est plus étroit. A partir de six bénéficiaires, l'exposition est faible, en conséquence, l'estimation est peu robuste et l'intervalle de confiance s'élargit. Ainsi, il sera nécessaire de réaliser un lissage des coefficients pour identifier une tendance.

Au-delà de l'analyse graphique, il est nécessaire de vérifier la significativité des coefficients à partir de l'analyse des erreurs standards. On dispose du coefficient de variation de la différence entre les impacts de 2 modalités :

	nb_beneficiaires (1)	nb_beneficiaires (2)	nb_beneficiaires (3)	nb_beneficiaires (4)	nb_beneficiaires (5)	nb_beneficiaires (6)	nb_beneficiaires (7)	nb_beneficiaires (8)	nb_beneficiaires (9)	nb_beneficiaires (10)	nb_beneficiaires (11)
nb_beneficiaires (1)											
nb_beneficiaires (2)	2										
nb_beneficiaires (3)	2	6									
nb_beneficiaires (4)	2	5	19								
nb_beneficiaires (5)	3	7	27	632							
nb_beneficiaires (6)	7	25	44	22	24						
nb_beneficiaires (7)	8	16	35	64	63	27					
nb_beneficiaires (8)	9	13	17	19	19	16	26				
nb_beneficiaires (9)	402	46	29	26	26	35	24	16			
nb_beneficiaires (10)	23	29	35	38	38	33	43	117	25		
nb_beneficiaires (11)	38	42	46	48	48	45	50	66	37	85	
	nb_beneficiaires (1)	nb_beneficiaires (2)	nb_beneficiaires (3)	nb_beneficiaires (4)	nb_beneficiaires (5)	nb_beneficiaires (6)	nb_beneficiaires (7)	nb_beneficiaires (8)	nb_beneficiaires (9)	nb_beneficiaires (10)	nb_beneficiaires (11)

FIGURE 3.11 – Tableau des rapports entre l'écart-type et espérance entre deux modalités

Ce tableau représente les erreurs standards définies précédemment. Plus il est faible, plus les impacts des 2 modalités sont significativement différents, signe que les 2 modalités doivent rester séparées dans le modèle.

Il faut vérifier qu'entre des classes voisines les valeurs des erreurs standards ne sont pas importantes. Sur le tableau ci-dessus les valeurs des classes 1 à 4 sont faibles, ce qui indique que les coefficients estimés de ces classes sont significativement éloignés, donc il n'y aura pas besoin d'agréger ces classes. Cependant, la valeur 632 en rouge indique qu'il est nécessaire de regrouper les classes 4 et 5. En observant les valeurs des erreurs standards, on arrive aux mêmes conclusions qu'à l'analyse graphique, en conséquence, on rassemble les classes 4 à 11 dans une nouvelle classe appelée « 4+ »dans la suite.

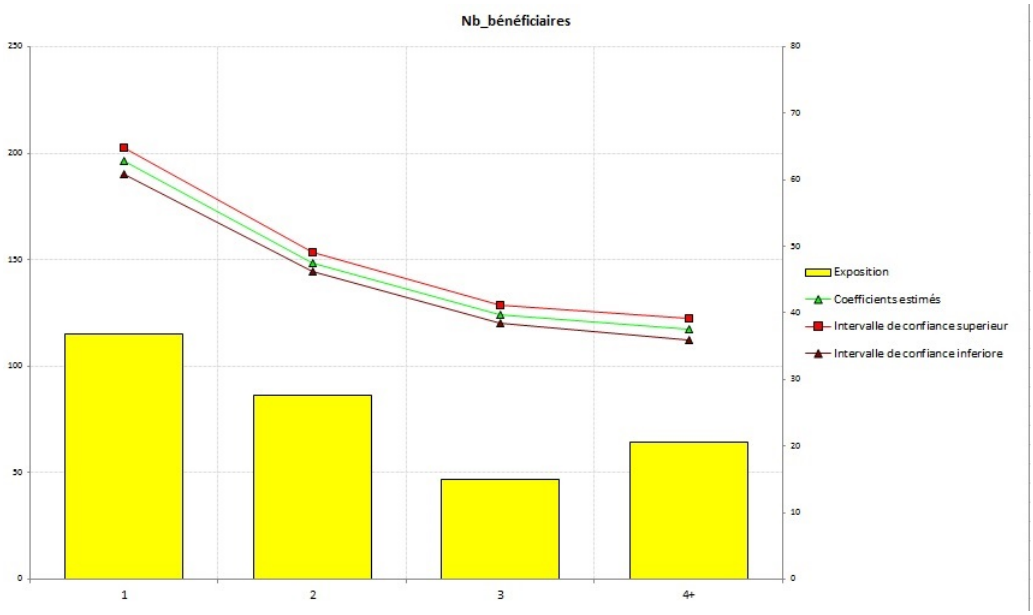


FIGURE 3.12 – Intervalle de confiance après regroupement des modalités

Finalement les classes présentées sur le graphique ci-dessus sont significativement différentes les unes des autres et en plus elles sont équilibrée en terme d'exposition.

3.2.5 Validation du modèle

• **Stabilité dans le temps** : Analyse croisée avec la variable année.
La significativité des coefficients n'est pas suffisante pour valider le modèle. Il faut également vérifier leur stabilité dans le temps, en effet une variable qui n'est pas stable dans le temps ne pourra pas être utilisée dans la modélisation car elle risque de biaiser la prédiction. Dans la suite on présente deux exemples de vérification de stabilité dans le temps.

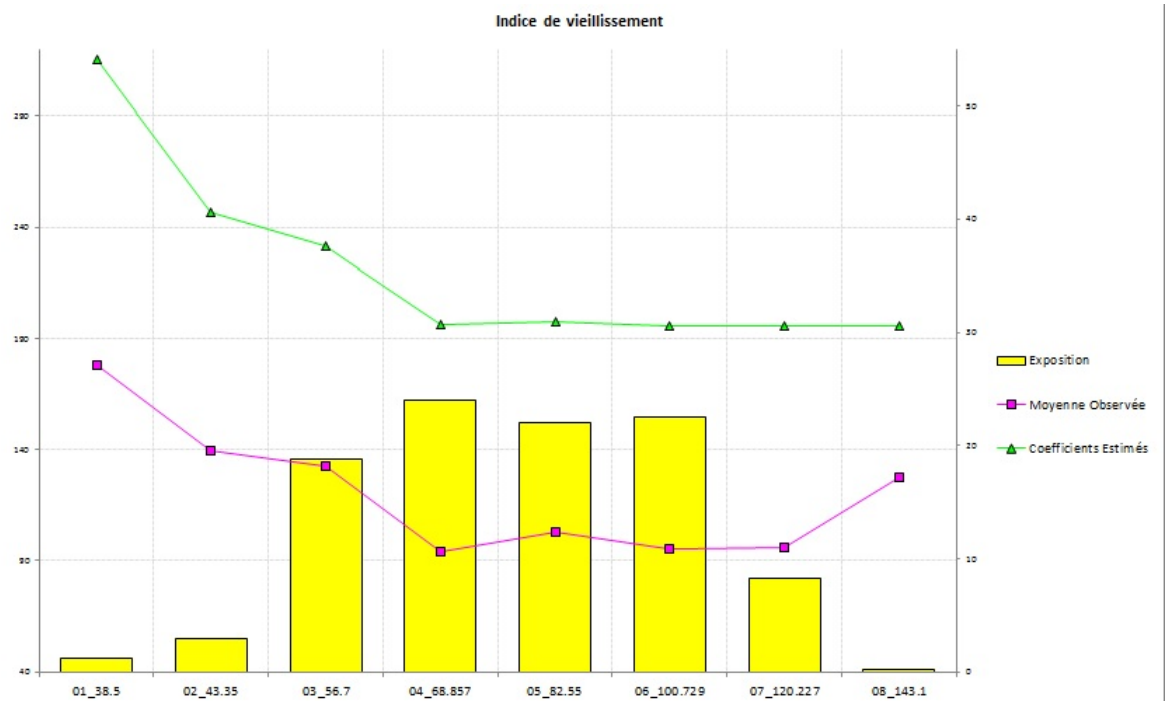


FIGURE 3.13 – Effet de la variable Indice de vieillissement compris dans le modèle Prime Pure direct du poste Hospitalisation

La variable « Indice de vieillissement » [4] a validé le test de significativité et graphiquement on constate que les coefficients estimés suivent la même tendance que celles de la prime pure observée. L'estimation ci-dessus a été réalisée sur tout l'historique des contrats, il reste à vérifier qu'en segmentant par année on retrouve la même information.

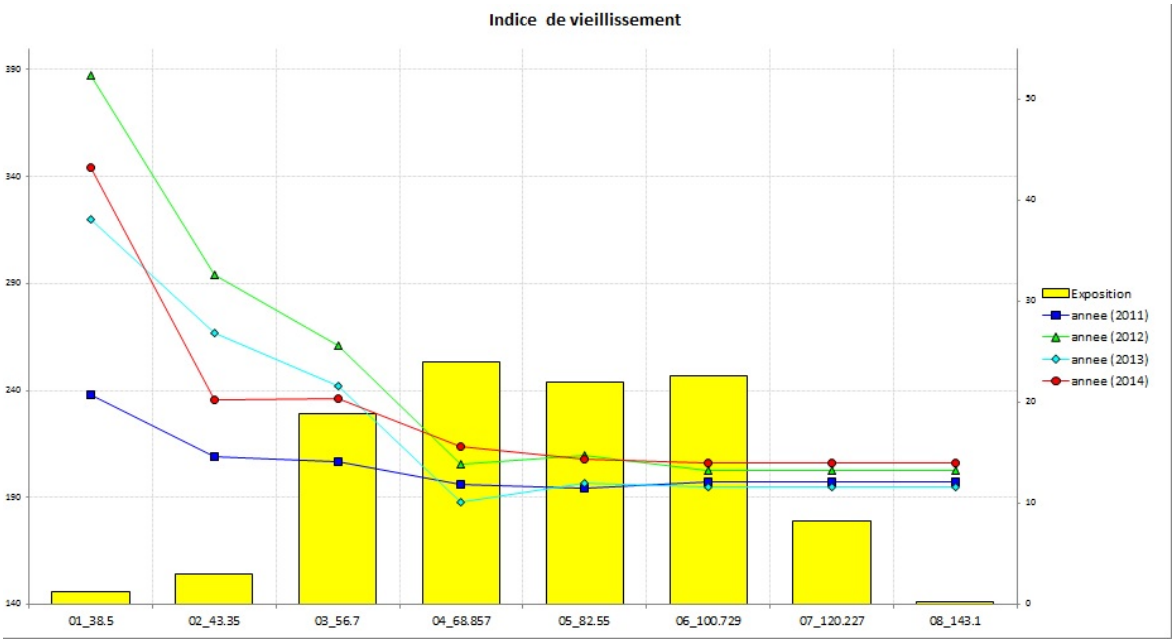


FIGURE 3.14 – Effet de la variable Indice de vieillissement par année d'exercice

4. Définition de la variable dans la section 2.3

On constate que les coefficients sont stables dans le temps, en effet les estimations par année suivent la même tendance, en conséquence ils peuvent être utilisés pour la prédiction des résultats futurs. Cependant, toutes les variables qui ont validé le test de significativité ne sont pas nécessairement stables dans le temps comme le montre l'exemple suivant.

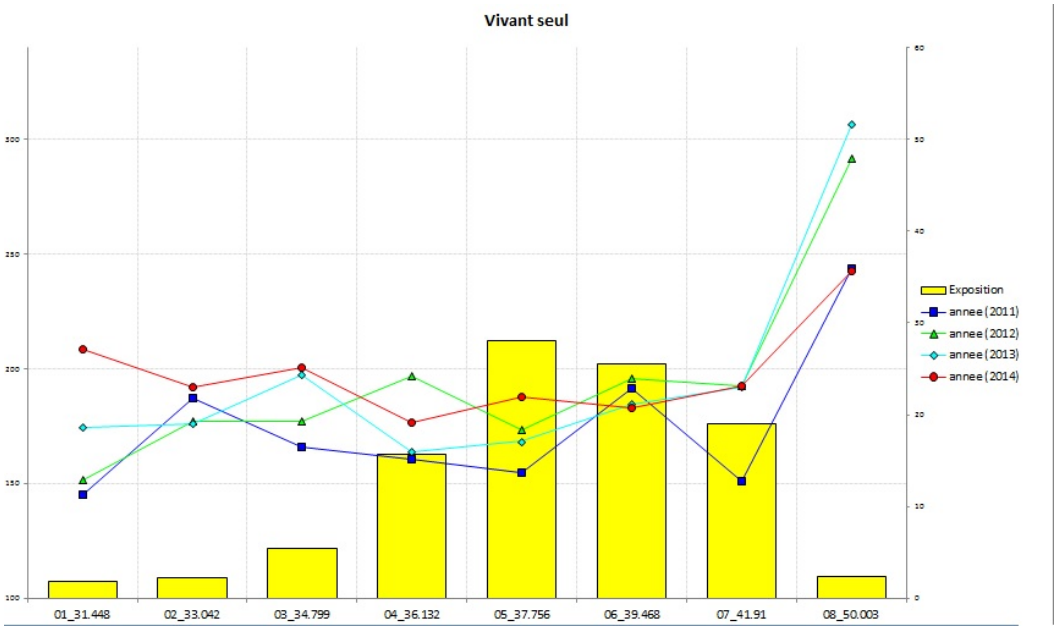


FIGURE 3.15 – Effet de la variable Vivant Seul par année d'exercice

La variable « Vivant seul » [5] a validé le test de significativité du modèle mais on constate sur le graphique ci-dessous que les coefficients estimés par année n'ont pas une tendance générale, donc cette variable ne sera pas intégrée dans la modélisation.

• **Validation sur 20% de la population** : Validation sur un échantillon.
Dans une première étape la modélisation est effectuée sur 80% des données totales, sélectionné de manière aléatoire. Pour vérifier si la modélisation effectuée est la plus ajustée, on vérifie les résultats obtenus sur un échantillon, 20% de la base totale qui n'a pas été utilisé précédemment.
Dans la représentation graphique suivante, on a considéré la variable « Indice de vieillissement »qui plus haut avait validé la stabilité dans le temps.

5. Définition de la variable dans la section 2.3

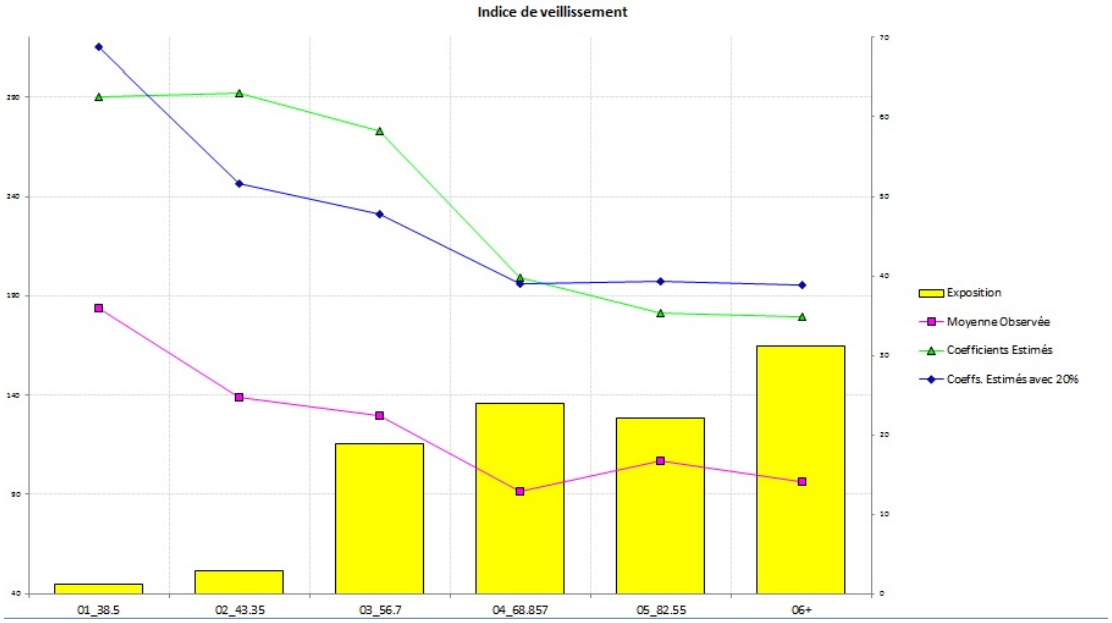


FIGURE 3.16 – Effet estimées de la variable Indice de vieillissement sur 80% et 20% de la population

La ligne verte représente les coefficients estimés avec 80% de la population, la ligne bleue correspond aux coefficients estimés à partir de l'échantillon de 20%. Il est nécessaire que les deux courbes aient la même tendance pour valider la continuité de cette variable dans le modèle.
Suite à la vérification des variables en terme de significativités, stabilité dans le temps et validation sur un échantillon, on arrive à démontrer la robustesse du modèle pour lequel on a conservé uniquement les variables satisfaisant tous les critères.

3.2.6 Analyse des résidus

La validité du modèle se mesure en vérifiant les hypothèses relatives aux résidus. Pour la $i^{\text{ème}}$ observation, les résidus dits « classiques »sont définis par,

$$e_i = |Y_i - \hat{Y}_i|.$$

Cependant, les résidus classiques sont peu intéressants dans la mesure où les observations sont hétéroscédastiques ($\text{Var}(\varepsilon_i) = \sigma_i^2$, où σ_i^2 peut être différent de σ_j^2, pour $i \neq j$). Aussi, l'alternative classique est de les normaliser, et d'étudier plutôt les **résidus de Pearson standardisés**.

Les résidus de Pearson standardisés

Dans le modèle linéaire classique,

$$Y = X\beta + \varepsilon,$$

où X est de rang p, $E[\varepsilon] = 0$ et $\text{V}(\varepsilon_i) = \sigma^2$ pour tout i,

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2 = (X^t X)^{-1} X^t Y,$$

d'où,

$$X\hat{\beta} = X(X^tX)^{-1}X^TY = HY.$$

Ici $X\hat{\beta}$ est le projeté orthogonal de Y sur vect(X), H est donc la matrice de projection dans vect(X). Elle est symetrique, idempotente, sa trace vaut p et ses éléments ne peuvent être supérieurs à 1.

En supposant que $H = [h_{i,j}]$ $i, j = 1, \cdots, p$, la variance de $\hat{\beta}$ est alors, $h_{i,j}\sigma^2$, où $h_{i,i}$ est le coefficient de levier (ou leverage). Dans un modèle Linéaire Généralisé, H continue d'exister avec la forme suivante,

$$H = W^{\frac{1}{2}}X(X'WX)^{-1}X'W^{\frac{1}{2}}.$$

La matrice (X^tX) utilisée dans le cas du modèle linéaire classique et généralisée dans la GLM en faisant intervenir une matrice W de pondération : X^tX devient (X^tWX), calculée à partir de la Hessienne approchée par l'algorithme de maximisation.

Rappel : $V(Y) = \sigma^2\Sigma$, où Σ est connu. Par conséquent $\hat{V}(Y) = \hat{\sigma}^2\Sigma$, avec $\sigma^2 = \frac{||e||^2}{n-p}$, $e = \{e_i\}_i$.

Definition 3.2.1. *Les résidus de Pearson standardisés sont définis par,*

$$\epsilon_i^R = \frac{Y_i - \hat{Y}_i}{\sqrt{\widehat{Var(Y_i)}}\sqrt{1-h_{i,i}}}.$$

On a construit deux modèles différents pour estimer la prime pure du poste « Hospitalisation »avec comme objectif de comparer leur performance et d'ajuster celui qui convient le plus.

	Tweedie	Fréquence × Coût moyen
Écart-type	30,72	31,02
Moyen absolu pondéré	61,13	110,64

TABLE 3.11 – Comparaison des statistiques descriptives des Résidus des modèles du poste Hospitalisation

Les résidus «Crunched»

Les résidus «Crunched» sont calculés sur des groupes d'observations et non sur les observations individuelles. Ils sont particulièrement utiles pour le modèle de fréquence, du fait que la base de données contient un plus grand nombre de cases où la fréquence est soit nulle soit très haute (sur des situations de risque dont l'exposition est très faible). Le résidu Crunched est défini par :

$$r_c = \frac{\sum(\text{Actual} - \text{Expected})}{\sqrt{\text{V}(\sum \text{Expected})}}$$

Où, Actual répresente la valeur observée, Expected la valeur estimée et V(.) la variance. La somme crée 500, 2,500 ou 10,000 groupes d'observations, chaque groupe étant formé des premières valeurs estimées triées par ordre croissant

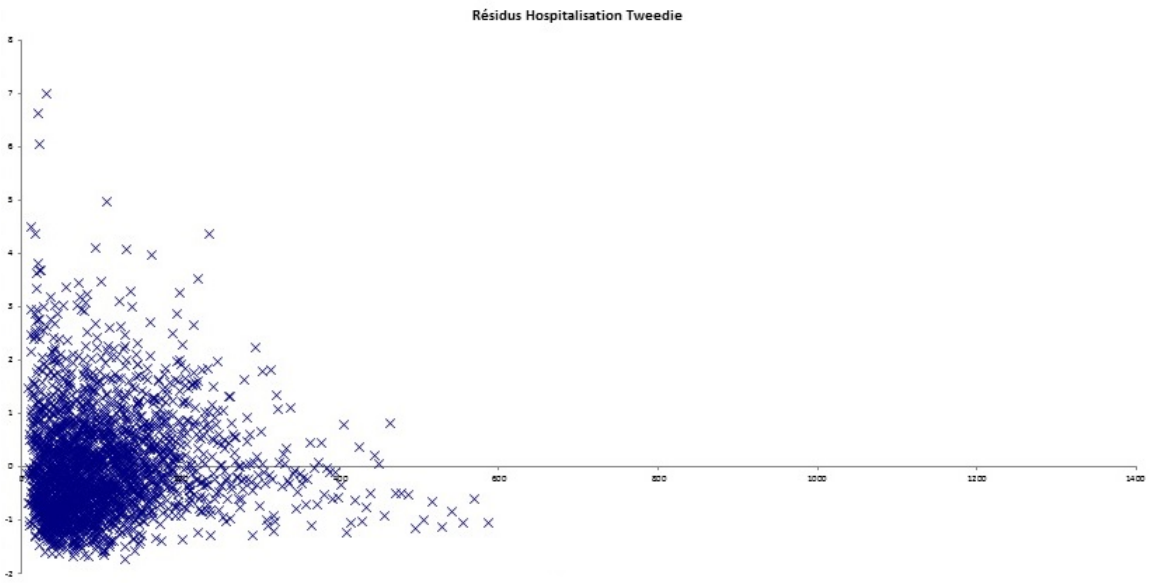


FIGURE 3.17 – Graphique des Résidus obtenus du modéle Tweedie,poste Hospitaliation

Interprétation : Un modèle est considéré comme bon si les graphes de ces résidus présentent une structure centrée autour de l'axe des abscisses.

Lorsqu'un point est isolé sur l'axe des ordonnées, il s'agit d'un fort résidu (valeur extrême). Il convient de retirer les résidus les plus forts, qui peuvent biaiser les coefficients du modèle. Ensuite, si la structure du modèle de résidus est satisfaisante, le modèle est considéré comme étant juste. Dans le cas contraire, il faut analyser la corrélation entre certains facteurs ou revoir les paramètres de la GLM.

Conclusion On a constaté que la grandeur des résidus est moins importante pour le modèle de Tweedie, on choisit de garder ce modèle pour l'estimation de la prime pure sur le poste hospitalisation.

3.2.7 Ajustement du modèle en vue du zonier, Alsace-Moselle

Cas particulier de l'Alsace Moselle

L'Alsace-Moselle bènéficie d'un régime particulier de sécurité sociale. Les zones regroupées sous le nom de l'Alsace-Moselle ont conservé le régime d'assurance maladie obligatoire mis en place par l'Allemagne pendant la guerre, période durant laquelle, ce territoire était annexé à l'Allemagne. Ce régime spécifique est plus avantageux que celui de la Sécurité Sociale, régis sur le reste de la France, le montant des remboursements est en effet plus intéressant. Par exemple, le niveau de remboursement pour une consultation chez un généraliste est de 90% du tarif de convention au lieu de 70% pour le régime général.

Le modèle de prime pure obtenu est globalement séparé en deux parties explicatives, les variables tarifaires et les variables externes. Ces dernières sont là pour capturer une partie de l'effet géographique désiré dans le zonier final. On veut assurer une distinction claire entre l'information propre au risque et l'information géographique. Or on remarque que la variable tarifaire régime possède la modalité Alsace-Moselle. Un assuré appartient à cette modalité si et seulement si il réside dans cette région-là, le critère de définition est

donc purement géographique. Pour rendre indépendante l'information tarifaire de l'information géographique, on est amené à forcer le coefficient de l'Alsace-Moselle pour qu'elle soit neutralisée dans l'effet tarifaire. Une partie de l'information sera réallouée aux autres critères, et en particulier par les variables externes. Quant à l'information perdue, elle est transférée dans les résidus qui seront ensuite traités pour capturer la partie manquante de l'effet géographique (non prise en compte par les variables externes). Ainsi l'information Alsace-Moselle participe pleinement au calcul du zonier.

Remarque : Suite à l'obtention du nouveau zonier et à sa réintroduction dans le modèle de prime pure, il rejoint les autres variables tarifaires, parmi lesquelles le régime est estimé sur toutes ses modalités y compris Alsace-Moselle. Ainsi, au final, l'effet géographique de l'Alsace-Moselle est autant que possible capté par le zonier et par la suite ajusté par la variable régime.

Réajustement pratique

Pour réaliser le forçage sur la modalité Alsace-Moselle, on utilise un « offset »sur la variable régime. On part des coefficients estimés dans le modèle validé (cf. 3.2.5.), et pour la variable régime on conserve les β des modalités sauf celui de l'Alsace-Moselle qui est forcé à 1. Cela signifie que par rapport au profil de référence, cette modalité ne fera ni augmenter ni diminuer le risque estimé. Le graphique suivant montre l'ajout de l'offset sur la modalité Alsace-Moselle de la variable tarifaire régime.

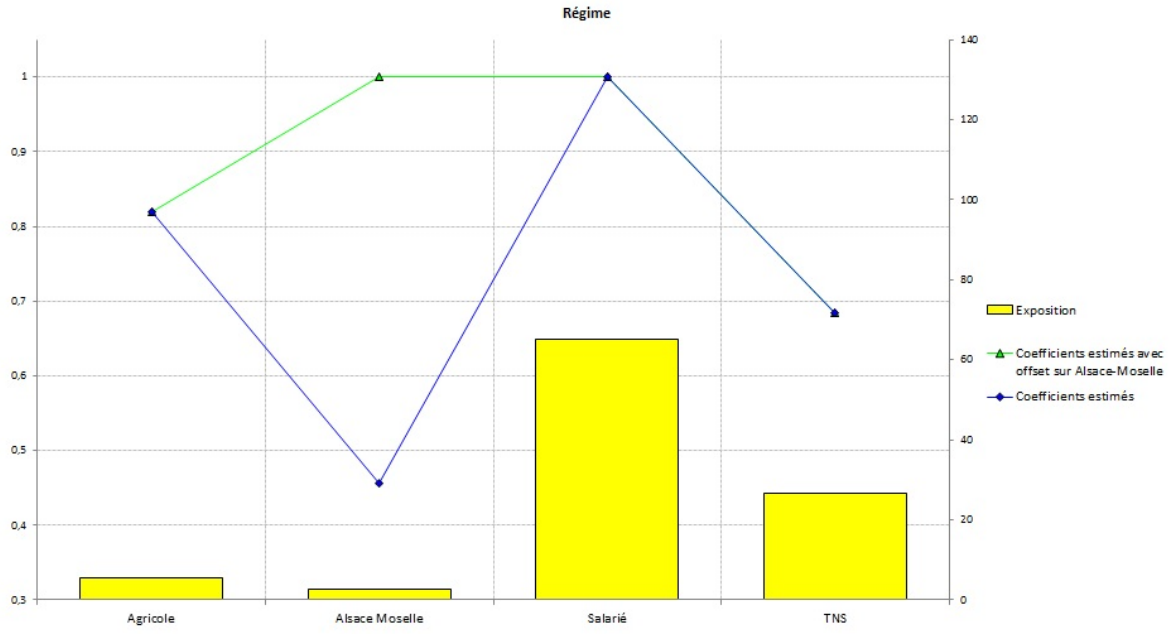


FIGURE 3.18 – Forçage sur la modalité Alsace-Moselle de la variable Régime

A titre d'illustration, ci-dessus deux graphiques qui répresent les impacts du forçage sur le moèle final de la prime pure du poste Hospitalisation. Dans les graphiques, les lignes bleues représentent les coefficients estimés avant offset, et les lignes vertes claires après offset.

Répresentation du forçage sur la variable tarifaire « Age »

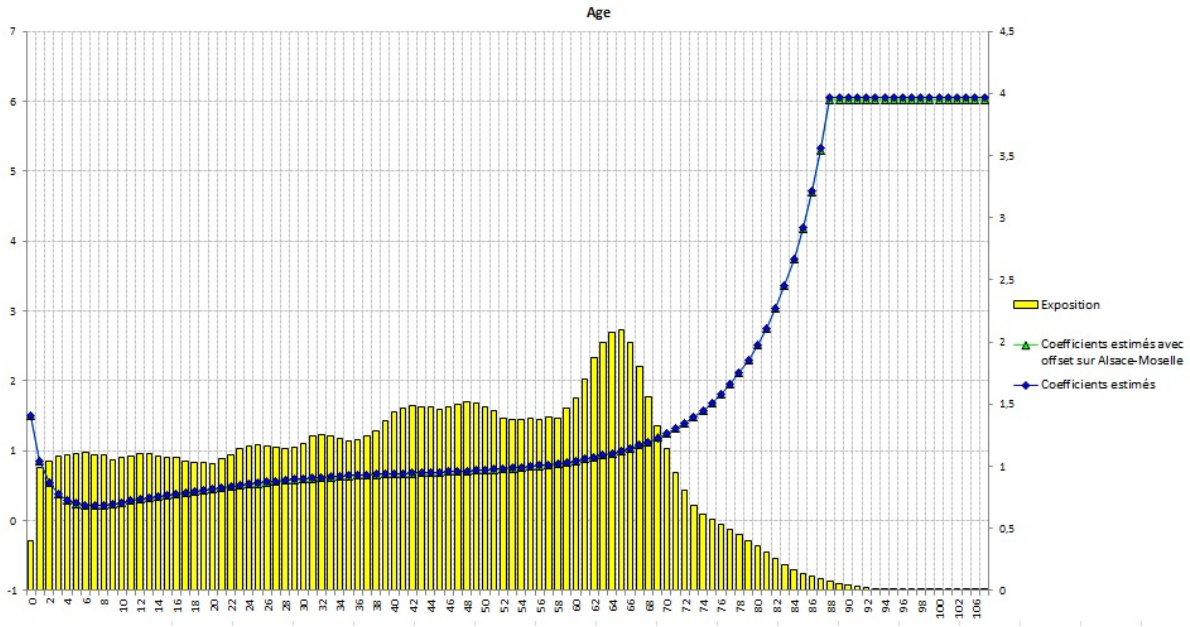


FIGURE 3.19 – Variable Âge

On constate bien que l'impact du forçage de l'Alsace-Moselle est neutre sur la variable tarifaire « Age », les coefficients ne changent quasiment pas entre les deux situations, ce même effet a été constaté sur les autres variables tarifaires.

Répresentation du forçage sur la variable externe « Espèrance de vie à la naissance »

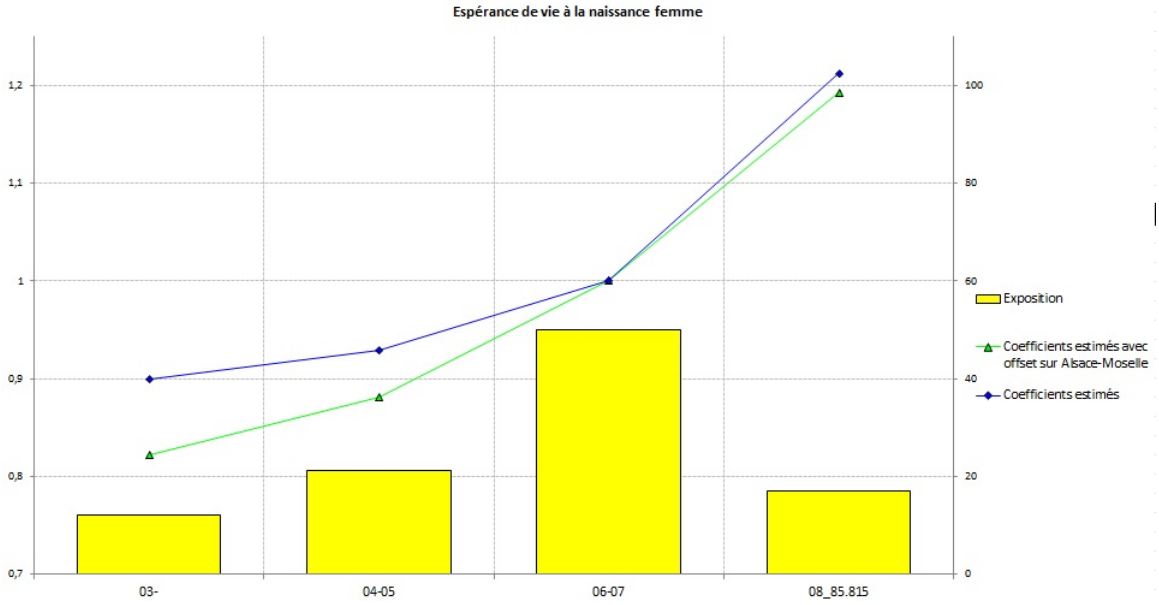


FIGURE 3.20 – Variable espèrance de vie à la naissance du sexe féminin

L'effet de l'Alsace-Moselle est donc bien géographique, car l'ajout de l'offset entraîne

des variations plus importantes sur les coefficients des variables externes, ce même effet a été aussi constaté sur les autres variables. Toutefois, les tendances restent les mêmes sur chaque variable, donc le modèle reste valide.

Impact du forçage sur le modèle

	Ajout de l'offset au modèle
P-valeur test χ^2	$0,0\%$
Variation Déviance	(+) 22.002,98
Variation AIC	(+) 21.996,98
Variation BIC	(+) 21.962,59

TABLE 3.12 – Impact du forçage sur le modèle Prime Pure direct du poste Hospitalisation

Suite à l'ajout de l'offset au modèle, on a globalement perdu de l'information qu'on constate par l'augmentation de la déviance, de l'AIC et du BIC. Cette perte d'information se traduit par une augmentation des résidus. L'information Alsace-Moselle conservée par le modèle se redistribue sur les autres critères, et s'agissant d'un effet géographique, il impacte principalement les variables externes.

Chapitre 4

Construction d'un nouveau zonage

Introduction

L'objectif de ce chapitre est la création et l'insertion du zonier dans la structure du risque.

$$Risque = Effets\ non\ géographiques + nouveau\ Zonier + epsilon.$$

Les effets non géographiques correspondent aux variables tarifaires, zonier actuel exclu, et epsilon à l'erreur finale de la modélisation, sans tendance (le bruit blanc).

La démarche est de modéliser le risque assuré en excluant le zonier actuel, calculer le nouveau zonier par commune, puis l'ajouter au modèle de base pour compléter l'explication du risque.

Il existe une multitude de méthodes pour la création d'un zonier, dans ce mémoire on a choisi d'isoler l'effet géographique contenu dans les résidus du modèle GLM et l'information donnée par les variables externes pour l'obtention du zonier final. On considère l'hypothèse que même en ajoutant des variables externes il existera toujours des effets liés à la zone géographique du risque que le modèle GLM est incapable d'expliquer.

La méthodologie de construction du zonier se résume par les étapes suivantes.

Étape 1

Dans le chapitre trois on a modélisé par la méthode de GLM les effets des variables connues, c'est-à-dire les variables tarifaires et les variables externes sélectionnées. La structure du risque à cette etape etait,

$$Risque = Effets\ connus + résidus_1,$$

où,

- Risque désigne ; soit la prime pure, soit la fréquence, soit le coût moyen.
- Les effets connus contiennent les effets non géographiques (via les variables tarifaires hors zonier actuel) et une partie de l'effet géographique externe (via les variables externes).
- Les résidus contiennent la partie inconnue de l'effet géographique, et l'erreur de modélisation.

Étape 2

L'objectif de cette étape est de lisser les $résidus_1$ projetés sur la carte de France, de manière à ce que l'erreur finale du modèle se rapproche autant que possible d'un bruit blanc.

- Avant lissage :

$$résidus_1 \;=\; effet\ géographique\ inconnu + erreur\ de\ modélisation\ 1.$$

- Après lissage :

$$résidus_1 \;=\; effet\ géographique\ résiduel + bruit\ blanc.$$

Dans cette dernière égalité,

- L'*effet géographique résiduel* s'obtient en lissant les $résidus_1$.

- Le *bruit blanc* ne contient plus aucune tendance géographique, sauf (éventuellement) celle qu'on ne veut pas voir apparaître dans le zonier.

Pour mieux capturer l'effet géographique résiduel, nous allons comparer deux méthodes de lissage. La première est une méthode de lissage stochastique qui utilise la théorie Bayesienne et la deuxième est une méthode déterministe qui utilise la théorie de la crédibilité, elles sont nommées *Adjacency* et *Distance* respectivement.

Le modèle à la fin de cette étape devient :

$$\begin{aligned} Risque = & Effets\ non\ géographiques \\ & + \ effets\ géographiques\ externes + effets\ géographiques\ résiduels \\ & + \ epsilon. \end{aligned} \tag{4.1}$$

Étape 3
L'objectif de cette étape est de construire le zonier à partir des effets uniquement géographiques. En combinant l'effet lissé des résidus à l'effet déjà estimé des facteurs externes, on obtiendra l'effet géographique total duquel on pourra déduire le nouveau zonier, selon une méthode finale de classification choisie.

L'effet géographique total est égal aux effets géographiques externes auxquels sont ajoutés les effets géographiques résiduels. En lui appliquant une classification hiérarchique par la méthode de *Ward*, on obtient le nouveau zonier,

$$Zonier = effet\ géographique\ regroupé\ en\ classes\ .$$

Le zonier est une synthèse des effets géographiques connus à l'étape 1 (impact des variables externes) et dans les $résidus\ lissés_1$ à l'étape 2.

Étape 4
L'objectif de cette dernière étape est d'intégrer le zonier dans le modèle. Le nouveau zonier est obtenu à l'étape 3 à partir des effets géographiques externes compris dans la modélisation des effets connus et les $résidus\ lissés_1$ obtenus à l'étape 2. En appliquant ce zonier au modèle défini par formule **(4.1)**, le risque peut finalement s'expliquer par,

$$Risque = Effets\ non\ géographiques + nouveau\ Zonier + epsilon'.$$

Dans tout ce processus, il est très utile d'avoir intégré certaines variables externes dans le modèle de base (étape 1), car les variables tarifaires sont alors déjà décorrelées avec ces variables externes, donc déjà décorrelées avec une partie de l'effet géographique. Quand

les variables externes sont remplacées par le zonier (étape 4), l'écart d'information est seulement égal à l'effet géographique résiduel, alors qu'il serait égal à l'effet géographique total si nous n'avions pas mis les variables externes au départ. Les coefficients des variables tarifaires sont donc beaucoup moins perturbés par l'arrivée du zonier.

L'étape 1 a été développée dans le chapitre précédent. Dans ce chapitre, nous détaillons les étapes 2 à 4.

Le principal apport de ce mémoire se situe au niveau de l'étape 2, où l'on va appliquer et confronter deux méthodes différentes de lissage, afin de nous orienter vers la méthodologie la plus adaptée pour la réussite de notre zonier en santé.

La première méthode présentée sera une méthode stochastique nommée, *Adjacency*, basée sur l'approche bayésienne, où le lissage est local et prend en compte le risque des communes immédiatement voisines, c'est-à-dire adjacentes. Nous comparerons les résultats obtenus avec une méthode déterministe nommée *Distance*, qui utilise la théorie de la crédibilité. Cette deuxième méthode lisse de manière globale en tenant compte des risques de l'ensemble des communes de France, avec une influence décroissante en fonction de la distance à la commune voisine.

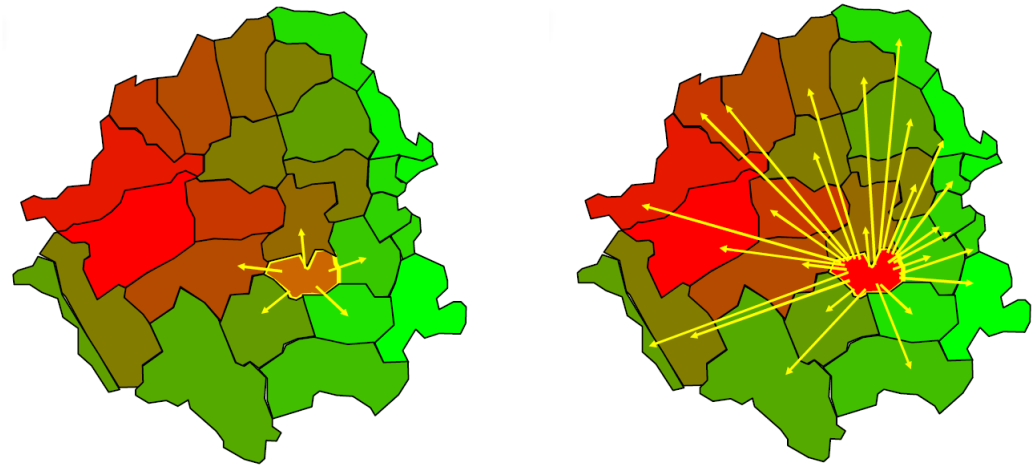


FIGURE 4.1 – À gauche l'idée du lissage spatial par *Adjacency* et à droite celle par *Distance*

A l'issue de cette démarche, le nouveau zonier résumera l'effet des variables externes ainsi que l'effet géographique construit par les deux méthodes de lissages que nous allons donc exposer dans cette partie.

4.1 Lissage spatial bayésien

4.1.1 L'approche bayésienne hiérarchique

L'approche bayésienne

Dans la théorie des modèles linéaires généralisés, la variable qui possède un caractère aléatoire est la variable à expliquer, c'est la seule qui bénéficie d'une loi de distribution attribuée à partir de l'échantillon observé. Cette distribution est caractérisée par un paramètre θ considéré comme fixe mais inconnu. Il est possible d'induire cette caractéristique

inconnue en estimant le paramètre par la méthode du maximum de vraisemblance. Ce type d'estimation s'inscrit dans un cadre d'inférence classique ou fréquentiste.

Dans un tel cadre, la fonction de vraisemblance résume à elle seule toute l'information que les données fournissent sur le paramètre inconnu. On ne fait pas d'hypothèse supplémentaire, θ n'a ainsi aucune structure a priori, aucune composante aléatoire en plus de l'influence directe des données.

Dans certains cas pourtant, on dispose d'informations complémentaires sur ce que doit être θ a priori pour mieux coller à la réalité. Ou bien simplement, on souhaite jouer sur le comportement de θ a priori, pour tester une déformation particulière des données par exemple. La théorie bayésienne permet justement d'intégrer ces informations au modèle, en faisant dépendre θ d'une loi spécifique qui traduit les nouvelles hypothèses.

Le paramètre inconnu n'est donc plus déterminé exclusivement par les données, mais conditionné en plus par les hypothèses a priori faites sur lui.

La théorie bayésienne ajoute un degré d'information, en considérant θ comme une variable aléatoire munie d'une loi propre $p(\theta)$. Les probabilités s'étendent ainsi à θ, et en inversant les rôles de y et de θ dans le conditionnement probabiliste, on obtient le paramètre θ au vu de la réalisation aléatoire. C'est donc pourquoi elle est naturellement liée au théorème de Bayes qui formalise l'inversion des conditionnements dans les probabilités.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

Cette probabilité est connue *a posteriori*, et apparaît comme le comportement des données observées (fonction de vraisemblance) influencé par l'opinion a priori. Cet impact a priori sur le modèle distingue l'inférence bayésienne de l'inférence classique.

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$
$$\text{A posteriori} \propto \text{Vraisemblance} * \text{A priori}.$$

Remarquons que le problème reposant sur une estimation de θ, le dénominateur ne représente qu'une normalisation.

Pour estimer le paramètre inconnu θ, on doit donc calculer l'impact $p(y|\theta)$ des données, ainsi que l'impact $p(\theta)$ de notre a priori sur θ.

L'inférence bayésienne hiérarchique

La théorie qui suit est en grande partie inspirée de, **[26]**, chapitre 10, section 1.

L'intérêt de l'analyse bayésienne hiérarchique est d'ajouter des hypothèses supplémentaires au modèle bayésien. Il s'agit de modéliser l'information a priori en la décomposant en plusieurs niveaux de distributions a priori conditionnelles, donc d'introduire des niveaux d'incertitude supplémentaires. Suivant la logique bayésienne, l'incertitude a tous les niveaux est prise en compte au moyen de lois a priori additionnelles.

Les hypothèses a priori sont emboîtées les unes dans les autres, «hiérarchisées» selon plusieurs niveaux. Dans les cas les plus simples, la structure hiérarchique ne contient que deux niveaux, les paramètres du premier étant associés a une distribution a priori définie dans le second, où la distribution de premier niveau est en général une loi a priori conjuguée.
Ce modèle bayésien hiérarchique simple peut donc se résumer en,

$$y|\theta \sim p(y|\theta)$$
$$\theta|\epsilon \sim p(\theta|\epsilon)$$
$$\epsilon \sim p(\theta),$$

avec y la variable aéatoire observée de paramètré θ et d'hyperparamètre ϵ.

Cette méthode nécessite de fixer l'hyperparamètre qui caractérise la loi a priori de θ, et on manque parfois d'information pour le faire correctement. Dans ce cas, on ajoute un second niveau de probabilité en considérant à son tour l'hyperparamètre comme une variable aléatoire dont on fixe la loi en amont.

Du fait de l'incertitude croissante avec le degré d'aléa présent dans le modèle, il faut prendre soin de bien définir chacun des niveaux d'hypothèses a priori.

4.1.2 Modélisation formalisée

Dans tout ce qui suit, nous prenons l'exemple de la modélisation de la fréquence. Le raisonnement est identique pour la prime pure ou le coût moyen.

Avant d'aller plus loin, faisons un rappel sur la modélisation effectuée dans le chapitre trois.

Au chapitre trois nous avons expliqué la fréquence y par les variables tarifaires et externes x, en décorrélant tous les effets. Nous avons pour cela utilisé les modèles linéaires généralisés. L'équation,

$$Risque \ = \ effets \ connus \ + \ résidus,$$

se traduit en :

$$g(y) = \beta x + \varepsilon_1,$$

où,

- La variable y suit une loi de Poisson (ou plus exactement : le nombre de sinistres suit une loi de Poisson, et nous normalisons par l'exposition), le paramètre lié à son espérance est β.
- La fonction de lien g est le logarithme népérien. De cette manière on aura une structure multiplicative des coefficients.
- La variable $\eta := \beta x$ est le prédicteur linéaire des effets connus (tarifaires et externes)
- ε_1 est l'erreur de modélisation à cette étape.

Le paramètre β est fixe et estimé par la méthode du maximum de vraisemblance, qui maximise la probabilité que la densité de la distribution empirique se rapproche de la densité de la loi de Poisson. Le combinaison des β correspond à une modélisation de chaque «profil de risque». L'effet d'un profil particulier part toujours de l'effet du profil de référence, multiplié par les effets des modalités appartenant au profil particulier. Avec tous ces choix nous obtenons une fréquence estimée de $y = \exp(\beta X) = \exp(\eta)$.

L'étape suivante consiste à la mise en place d'une méthode qui fournira un estimateur du risque associé à chaque commune. Les composantes tarifaires (hors zonier) et externes ont déjà été expliquées, le nouvel espace explicatif du modèle est donc la carte des communes de la France métropolitaine.

L'hypothèse de base utilisée par la méthode de lissage bayésienne hiérarchique est que les communes qui sont proches les unes des autres ont plus de chance de présenter un risque similaire que celles qui sont plus éloignées.
Le fait d'intégrer cet argument contribue à l'homogénéité du zonier. Cela se traduit par le fait que le risque d'une commune doit être influencé par ses communes voisines, et on attend comme résultats que deux communes adjacentes soient de la même couleur, ou proches, de manière à avoir une carte lissée.

L'approche bayésienne nous offre justement la possibilité d'ajouter cette hypothèse a priori dans le modèle, grâce à la définition d'une densité a priori du risque géographique.

Mise en œuvre

Nous avons déjà estimé la volatilité de la fréquence liée aux variables connues, il suffit donc de pondérer par l'exposition de chaque commune pour obtenir le nombre de sinistres prédit : $r_i \exp(\eta_i)$.

Le niveau du risque de la commune i, est ainsi composé des effets connus, auxquels s'ajoutent les effets décorrélés de variation spatiale et de variation inexpliquée.

Le modèle du niveau de risque réel y_i de la commune i s'écrit donc,

$$\begin{aligned} y_i &= e_i \exp(u_i)\exp(v_i) \\ &= r_i \exp(\eta_i + u_i + v_i), \end{aligned} \tag{4.2}$$

où,

r_i est l'exposition au risque de la commune i ;

η_i est le prédicteur linéaire des effets connus (tarifaires standardisées et externes) ;

u_i variable aléatoire qui représente l'effet géographique ;

v_i variable aléatoire représentant l'effet inexpliqué du modèle.

En plus de ce que présentent les données, nous faisons l'hypothèse a priori sur les x_i vue plus haut. Cette dernière implique donc que les x_i sont des variables aléatoires identiquement distribuées selon une loi a priori $p(x_i)$. On peut alors conditionner par les x_i, et en utilisant le théorème de Bayes, on obtient l'expression de la densité a posteriori :

$$p(x|y) \propto p(y|x)p(x).$$

où $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_n)$, avec n le nombre de communes.

Nous cherchons à estimer le risque moyen de la commune, ce qui correspond à l'espérance de $x|y$ (car après observation des données et impact de l'a priori).

Pour construire le zonier, nous avons besoin de l'effet géographique total. Nous connaissons déjà l'effet géographique issu des variables externes, et nous avons maintenant besoin d'estimer l'effet géographique résiduel. Nous aurons ainsi une expression de la densité a posteriori que nous simulerons avec la méthode de Monte Carlo qui simule un nombre important de réalisation d'une densité, et permet d'en déduire l'espérance de la densité empirique.

Fonction de vraisemblance

Conditionnellement à x_i, nous supposons que les y_i sont mutuellement indépendantes et distribuées selon une loi de Poisson de paramètre x_i,

$$p(y_i|x_i) = \exp(-x_i)\frac{x_i^{y_i}}{y_i!}, \quad y_i \in \mathbb{N}.$$

D'où,

$$\begin{aligned} p(y|x) &= \prod_{i=1}^{n} p(y_i|x_i) \\ &= \prod_{i=1}^{n} \exp(-x_i)\frac{x_i^{y_i}}{y_i!}. \end{aligned}$$

Distributions a priori

Dans l'équation **(4.2)** nous avons vu que le niveau de risque dépendait de deux variables aléatoires, u et v correspondant respectivement à l'effet géographique et à l'effet inexpliqué du modèle. Ces variables aléatoires étant supposées indépendantes, on a $p(u, v) = p(u)p(v)$, les densités a priori de u et v pourront être calculés séparément.

Nous définissons la densité a priori de la variable u conformément à l'hypothèse a priori que nous voulons faire : une commune i doit être plus influencée par les communes avoisinantes ayant un risque similaire. On peut donc raisonnablement supposer une distribution d'écart centrée sur 0 et tendant vers 0. Autrement dit, on suppose que le risque de la commune i et les communes au voisinage de i est distribué selon une loi normale $N(\overline{u}_j, \tau)$, avec $\overline{u}_j$ la moyenne des u_j voisins de i et τ inconnu (deuxième hyperparamètre). Ainsi, en notant δ_i le voisinage de la commune i, on peut écrire la densité a priori de u_i sachant τ et δ_i :

$$p(u_{i|u_{-i}}, \tau) \propto \tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau}\sum_{j\in\delta_i}(u_i - u_j)^2\right),$$

où $u_{-i} = u_1, \cdots, u_{i-1}, u_{i+1}, \cdots, u_n$.

Au voisinage de i, plus le risque de la commune voisine j est éloigné du risque de i, moins il influe sur la valeur finale du risque. On oriente ainsi le risque pour qu'il ressemble au risque proche voisin. Ce processus correspond bien à un lissage spatial, puisque la valeur de u_i est modifiée selon une loi déterminée par des données extérieures à u_i, les valeurs de risque voisines.

Pour faire intervenir ce conditionnement dans le calcul de la densité a priori de u, on utilise la définition de la densité conditionnelle, en conditionnant chaque fois par les communes manquantes :

$$\begin{aligned} p(u|\tau) &= p(u_n|u_1, \cdots, u_{n-1}, \tau)p(u_1, \cdots, u_{n-1}|\tau) \\ &= p(u_n|u_1, \cdots, u_{n-1}, \tau)p(u_n|u_1, \cdots, u_{n-2}, \tau)p(u_1, \cdots, u_{n-2}, \tau) \\ &= p(u_n|u_1, \cdots, u_{n-1}, \tau)\cdots p(u_2|u_1), \tau)p(u_1|\tau), \end{aligned}$$

on obtient ainsi pour la densité jointe de u l'égalité,

$$p(u, \tau) \propto \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau}\sum_{i\sim j}(u_i - u_j)^2\right),$$

où $i \sim j$ signifie que nous sommons sur l'ensemble des régions voisines, en ne diférenciant pas le couple (i, j) du couple (j, i). Ainsi, chaque paire de régions voisines ne sera utilisée qu'une seule fois, conformément aux chaînes de conditionnement ci-dessus.

Passons maintenant à l'étude plus détaillée des variables aléatoires $v_i,\ i = 1, \cdots, n$, ces variables sont supposées être des bruits blancs, elles suivent donc une loi normale $N(0, \lambda)$, avec λ inconnu. Il s'agit de l'hyperparamètre du modèle bayésien hiérarchique. Les variables $v_i,\ i = 1, \cdots, n$ supposées de plus indépendantes, on peut écrire la densité a priori de $v = v_{i\,i=1}^{n}$ sachant λ, par :

$$\begin{aligned} p(v|\lambda) &= \prod_{i=1}^{n} p(v_i|\lambda) \\ &\propto \lambda^{-\frac{n}{2}} \exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2\right). \end{aligned}$$

Ne disposant pas de sufisamment d'informations concernant les hyperparamètres τ et λ, qui déterminent les variances des lois normales a priori de u et v, nous les considérons à leur tour comme étant des variables aléatoires, en leur attribuant une loi proche des distributions non informatives habituelles,

$$p(\tau, \lambda) \propto \exp(-\frac{\epsilon}{2\tau} - \frac{\epsilon}{2\lambda}),$$

avec ϵ une petite constante positive (nous commencerons par 0,01 avant de tester la sensibilité de cette constante).

Ainsi, le modèle bayésien hiérarchique se résume par,

$$\begin{aligned} y|u, v &\sim p(y|u, v), \\ u|\tau &\sim p(u|v), \\ v|\lambda &\sim p(v|\lambda), \\ \tau, \lambda &\sim p(\tau, \lambda). \end{aligned}$$

Densité a posteriori recherchée

Dans l'équation,

$$x_i = r_i \exp(\eta_i + u_i + v_i),$$

les paramètres r_i et η_i sont connus. Les seules paramètres qui ont un caractère aléatoire sont $u_i \sim N(0, \tau)$ et $v_i \sim N(0, \lambda)$. Ce qui nous amène à définir la densité a posteriori de notre modèle par,

$$\begin{aligned} p(u, v, \tau, \lambda|y) &\propto p(y|u, v, \tau, \lambda)p(u, v, \tau, \lambda) \\ &\propto \left[\prod_{i=1}^{n} p(y_i|x_i)\right] p(u|v)p(v|\lambda)p(\tau|\lambda) \\ &\propto \left[\prod_{i=1}^{n} \exp(-x_i)\frac{x_i^{y_i}}{y_i!}\right] \tau^{-\frac{n}{2}} \exp\left(\sum_{i\sim j}(u_i - u_j)^2\right) \cdot \\ &\qquad \lambda^{-\frac{n}{2}} \exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2\right) \exp(-\frac{\epsilon}{2\tau} - \frac{\epsilon}{2\lambda}). \end{aligned}$$

En théorie bayésienne, l'estimateur classique est celui du maximum a posteriori. Le but est ainsi d'estimer les paramètres qui maximisent cette densité a posteriori. Mais en tant que telle, la maximisation n'est pas réalisable car la densité en question est une fonction non linéaire définie sur $2n + 2$ variables.

Finalement, remarquons que la génération des variables aléatoires à partir de la méthode de Monte-Carlo requière des lois de distribution de probabilité standard. Pour pallier à cette difficulté technique, nous allons détailler dans la section suivante l'utilisation de méthodes dites de Monte Carlo par Chaînes de Markov.

4.1.3 Simulation par la méthode d'échantillonage de Gibbs

La simulation de la densité a posteriori nécessite des méthodes de Monte Carlo par Chaînes de Markov que nous allons expliciter ci-dessous. En particulier, nous présenterons la méthode de l'échantillonnage de Gibbs qui en est un cas particulier et que nous utiliserons pour la résolution finale.

Dans un cadre bayésien, il est parfois très difficile de calculer l'intégralité des probabilités a posteriori. Un algorithme de simulation de Monte-Carlo permettant d'estimer ces distributions est cependant disponible pour approcher la sélection bayésienne des jeux de variables explicatives les plus probables.

Les chaînes de Markov

Nous rappelons ici quelques définitions et propriétés essentielles des chaînes de Markov qui nous serviront par la suite.

Soit I un ensemble fini ou dénombrable.

- Un **état** est un élément de I.
- Une **distribution de probabilité** est une suite $\lambda_i = (\lambda_i)_{i\in I}$ telle que pour tout $i \in I$

$$0 \leq \lambda_i \leq 1, \quad \text{et} \quad \sum_{i\in I} \lambda_i = 1.$$

- Une **matrice stochastique** $P = (p_{ij})_{i,j\in I}$ est telle que chaque ligne est une distribution de probabilité.
- Une chaîne de Markov est un ensemble de variables aléatoires $X_0, X_1, \cdots$ à valeur dans I tel que,

$$P(X_{n+1} = x_{n+1}|X_n = x_n, \cdots, X_0 = x_0) = P(X_{n+1} = x_{n+1}|X_n = x_n).$$

Dans une chaîne de Markov, la seule information passée nécessaire pour connaître le futur est l'information présente. Les expériences passées sont absorbées par l'expérience présente.

- Une chaîne de Markov est **homogène** si la probabilité $p_{i,j}$ de passer en une étape d'un état i à un état j ne dépend pas de l'instant n,

$$P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i)$$

- Les $p_{i,j}$ forment une matrice stochastique $P = (p_{ij})_{i,j\in I}$ appelée **matrice de passage** de $X_0, X_1, \cdots$.
- Les probabilités de passer de l'état i à j en n étapes sont notées,

$$P(X_n = j|X_0 = i) = p^n_{i,j}.$$

- La distribution des états de la chaîne de Markov X après n étapes, i.e. la distribution de X_n, est notée $\mu^{(n)}$, telle que pour tout $i \in I$:

$$\mu_i^{(n)} = P(X_n = i).$$

La distribution initiale de X est donc $\mu^{(0)}$, et correspond à la loi de X_0.

D'où pour tout $i \in I$,

$$\begin{aligned}\mu_i^{(n)} &= P(X_n = i) \\ &= \sum_{j \in I} \mu_j^{n-1} p_{j,i}.\end{aligned}$$

Et sous forme matricielle,

$$\begin{aligned}\mu^{(n)} &= \mu^{(n-1)} P \\ &= (\mu^{(n-2)} P) P \sum_{j \in I} \mu_j^{n-1} p_{j,i} \\ &\vdots \\ &= \mu^{(0)} P^n.\end{aligned}$$

La loi de X_n dépend donc uniquement de la loi initiale de X et de la matrice de passage d'un état à un autre de la chaîne.

Pour appréhender le théorème central utilisé par la suite, nous détaillons maintenant le comportement de $\mu^{(n)}$, en particulier les conditions sous lesquelles X converge vers un état stable indépendant de X_0.

Soient $X = (X_n)_{n \in \mathbb{N}}$ une chaîne de Markov de matrice de transition P, et $i, j \in I$ deux états quelconques.

- X est **irréductible** si tous les états de I communiquent entre eux, s'ils peuvent tous être atteints les uns par les autres, i.e. pour tout $n, m > 0$,

$$p_{i,j}^{(n)} > 0 \quad \text{et} \quad p_{j,i}^{(m)} > 0$$

- X est **périodique** si les états reviennent cycliquement selon une même période (au bout de 8 étapes par exemple). Sinon elle est dite **apériodique**.
- X est **récurrente positive** si le temps moyen pour revenir à l'état i sachant que l'on est parti de i est fini.
- Une chaîne de Markov est **ergodique** si elle est irréductible, apériodique et récurrente positive.
- Une distribution **stationnaire** π de la chaîne de Markov est une distribution telle que,

$$\pi = \pi P.$$

La distribution stationnaire π existe car P est stochastique donc sa plus grande valeur propre est 1, ce qui définit π. Une distribution stationnaire ne change pas suite au passage d'un état à un autre.

Si P est une chaîne de Markov ergodique, alors sa distribution stationnaire est unique et P^n converge vers une matrice dont chaque ligne est égale à cette unique distribution. C'est-à-dire, la distribution de X_n converge vers la distribution stationnaire.

Théorème 4.1.1. *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov ergodique de distribution stationnaire π et f une fonction réelle définie sur l'espace des états I de la chaîne. Alors,*

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} f(X_t) = \sum_{i\in I} \pi_i f(i) = E_\pi(f(x)),\ \ p.s$$

Donc en particulier (avec la fonction identité) si X a un nombre suffisant d'états on peut approximer son espérance par la distribution empirique après qu'elle ait atteinte la distribution stationnaire.

Nous disposons à présent des définitions et propriétés nécessaires pour aborder les techniques dites de Monte Carlo par Chaînes de Markov.

Les méthodes de Monte Carlo par Chaînes de Markov

Ces techniques sont utilisées dans les situations où la densité est trop complexe pour être simulée en pratique via Monte Carlo standard.

L'idée consiste à construire une chaîne de Markov ergodique de distribution stationnaire π, au lieu de simuler directement un échantillon de loi π. En partant d'une chaîne arbitraire dont on ajuste les états au cours du processus, on arrive à la stationnariser et ainsi atteindre π. Les conditions sont alors remplies pour simuler des échantillons de cette loi sur chaque commune.

Contrairement à la méthode d'estimation du maximum a posteriori qui donne d'un bloc les paramètres estimés, les méthodes de Monte Carlo par Chaînes de Markov fournissent un échantillon entier de la densité a posteriori sur chaque commune. Les paramètres que nous cherchons u, v, λ et τ pourront ainsi être estimés en prenant la moyenne de la densité empirique simulée. Pour réaliser cette estimation, nous allons utiliser l'échantillonnage de Gibbs, que nous détaillons ci-dessous.

L'échantillonnage de Gibbs

L'échantillonnage de Gibbs est particulièrement adapté aux modèles graphiques et aux modèles hiérarchiques. La méthode repose sur la simulation successive suivant les diverses lois conditionnelles associées à la loi a posteriori cible. Avec cette méthode, nous obtenons des réalisations de la densité a posteriori sans la calculer explicitement. En simulant des densités correctement conditionnées, elle permet de transformer notre estimation d'une fonction complexe et multivariée en l'estimation de fonctions simples et univariées.

Passons maintenant à la description des étapes de cette méthode :

On part du vecteur aléatoire $X = (u, v, \tau, \lambda, y)$. La densité a posteriori $p(u, v, \tau, \lambda|y)$ est celle que nous voulons estimer, mais sa complexité nous empêche d'obtenir un échantillon sur chaque commune. C'est ici que l'échantillonnage de Gibbs intervient : il s'agit d'une méthode itérative où à chaque étape de l'algorithme, la valeur de chaque paramètre est remplacée par une valeur choisie aléatoirement à partir de sa distribution conditionnelle totale. Pour une étape, la valeur d'un paramètre change, les autres restent constants.

Le principe de l'algorithme est de simuler une chaîne de Markov ergodique de distribution stationnaire $p(u, v, \tau, \lambda|y)$. On doit initialiser l'algorithme, c'est-à-dire fixer les valeurs de $X(0)$.

A chaque étape, on simule les distributions conditionnelles suivantes :

$$\begin{aligned}
\tau^{(t+1)} &\sim p(\tau|_u^{(t)}, v^{(t)}, \lambda^{(t)}, y) \\
\lambda^{(t+1)} &\sim p(\lambda|_u^{(t)}, v^{(t)}, \tau^{(t+1)}, y) \\
u_1^{(t+1)} &\sim p(u_1|_{u_{-1}}^{(t)}, v^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y) \\
&\vdots \\
u_n^{(t+1)} &\sim p(u_n|_{u_{-n}}^{(t)}, v^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y) \\
v_1^{(t+1)} &\sim p(v_1|_{v_{-1}}^{(t)}, u^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y) \\
&\vdots \\
v_n^{(t+1)} &\sim p(v_n|_{v_{-n}}^{(t)}, u^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y).
\end{aligned}$$

Ce nouvel état actualise au fur et à mesure les composantes de X. Par exemple, les $u_1^{(t+1)}, \cdots, u_{n-1}^{(t+1)}$ sont créés en premier, donc on peut tout de suite les utiliser en conditionnement pour simuler $u_n^{(t+1)}$.

La chaîne de Markov ainsi créée va converger vers la distribution stationnaire correspondante à la densité recherchée $p(u, v, \tau, \lambda|y)$. Grâce au théorème ergodique, les états d'après la distribution stationnaire permettent de simuler des échantillons de la densité empirique indépendants sur chaque commune.

En effet, supposons que la chaîne atteigne le stade stationnaire à la $k^{\text{ème}}$ simulation. Si nous disposons d'un nombre suffisant d'états T, le théorème ergodique nous donne un estimateur de l'espérance de X, qui après la $k^{\text{ème}}$ simulation suit la loi a posteriori $p(u, v, \tau, \lambda|y)$:

$$\hat{X} = \frac{1}{T-k} \sum_{t=k+1}^{T} X^{(t)}.$$

Au lieu d'avoir à simuler directement une densité complexe, chaque itération de l'algorithme de Gibbs simule les densités univariées de chaque paramètre conditionné par tous les autres. Il nous faut donc déterminer l'expression de ces densités conditionnelles.

Les «densités conditionnelles des paramètres conditionnellement à tous les autres», appelées aussi densités conditionnelles totales, se calculent à l'aide du théorème de Bayes et la notation proportionnelle vus plus haut :

- Densité conditionnelle totale de u_i :

$$\begin{aligned}
p(u_i|u_{-1}, v, \tau, \lambda, y) &\propto p(y_i|x_i)p(u_i|u_{i-1}, \tau) \\
&\propto \exp(-x_i)x_i^{y_i} \exp\left(-\frac{1}{2\tau}\sum_{j\in\delta_i}(u_i - u_j)^2\right) \\
&\propto \exp(\varepsilon_i \exp(u_i + v_i)) \exp(y_i \log(-\varepsilon_i \exp(u_i + v_i))) \exp\left(-\frac{\#\delta}{2\tau}(u_i - \overline{u}_j)^2\right) \\
&\propto \exp(\varepsilon_i \exp(u_i + v_i)) \exp(y_i \log(-\varepsilon_i) + y_i(u_i + v_i)) \exp\left(-\frac{\#\delta}{2\tau}(u_i - \overline{u}_j)^2\right) \\
&\propto \exp\left(\varepsilon_i \exp(u_i + v_i) + u_i y_i - \frac{\#\delta}{2\tau}(u_i - \overline{u}_j)^2\right),
\end{aligned}$$

avec $\varepsilon_i = r_i \exp(\eta_i)$, $\#\delta_i$ le nombre de communes voisines de i et $\overline{u}_j$ la moyenne des u_j voisins de i.

- Densité conditionnelle totale de v_i :

$$\begin{aligned} p(v_i|v_{-1},u,\tau,\lambda,y) &\propto p(y_i|x_i)p(v_i,\lambda) \\ &\propto p(y_i|x_i)\exp(-\frac{1}{2\lambda}v_i^2) \\ &\propto \exp\left(\epsilon_i \exp(u_i+v_i)+v_i y_i-\frac{1}{2\lambda}v_i^2\right), \end{aligned}$$

où une nouvelle fois $\varepsilon_i = r_i \exp(\eta_i)$.

- Densité conditionnelle totale de τ :

$$\begin{aligned} p(\tau|u,v,\lambda,y) &\propto p(u|v,\tau,\lambda,y)p(\tau|v,\lambda,y) \\ &\propto \tau^{-\frac{n}{2}}\exp\left(-\frac{1}{2\tau}\sum_{i\sim j}(u_i-u_j)^2\right)\exp\left(-\frac{\epsilon}{2\tau}-\frac{\epsilon}{2\lambda}\right) \\ &\propto \tau^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\tau}\left(\epsilon+\sum_{i\sim j}(u_i-u_j)^2\right)\right\}. \end{aligned}$$

- Densité conditionnelle totale de λ :

$$\begin{aligned} p(\lambda|u,v,\tau,y) &\propto p(v|u,\tau,\lambda,y)p(\lambda|u,\tau,y) \\ &\propto \lambda^{-\frac{n}{2}}\exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n}(v_i)^2\right)\exp\left(-\frac{\epsilon}{2\tau}-\frac{\epsilon}{2\lambda}\right) \\ &\propto \lambda^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\lambda}\left(\epsilon+\sum_{i=1}^{n}(v_i)^2\right)\right\}. \end{aligned}$$

Les densités conditionnelles totales de τ et λ peuvent être simulées directement, car leur expression peut s'identifier à une loi χ^2.

Les densités conditionnelles totales de u_i et v_i n'étant pas standards, on ne peut pas les simuler directement. Nous utiliserons la méthode de rejet adaptatif de Gilks & Wild, appelée Adaptative Rejection Sampling.

La méthode de rejet adaptatif de Gilks & Wild

Les techniques de simulation élémentaires de type acceptation-rejet sont prises souvent pour des états difficiles à parcourir.

L'étape de proposition vise à explorer l'espace, en proposant des états aléatoires selon une transition markovienne. Comme son nom l'indique, l'étape d'acceptation-rejet consiste à accepter la proposition précédente avec une certaine probabilité, dans le cas contraire l'algorithme retourne à son état précédent la proposition. Cette exploration aléatoire est le plus souvent dictée par une idée de voisinages sur l'espace d'état. L'algorithme propose alors aléatoirement un état voisin au précédent.

L'algorithme d'acceptation-rejet est économique et rapide à condition que la fonction instrument soit une bonne approximation de la fonction cible. En pratique, il est généralement difficile de trouver une telle fonction.

Gilks et Wild ont introduit la méthode Adaptative Rejection Sampling en 1992, elle permet d'obtenir des échantillons indépendants de lois non standards, la seule contrainte étant que la fonction densité cible à simuler soit log-concave et univariée.

La méthode de rejet adaptatif utilise les propriétés des fonctions log-concaves pour construire une fonction enveloppant la densité que l'on souhaite simuler, ce qui permet de minimiser le nombre de candidats rejetés. Les gains en terme de performance viennent en fait d'une augmentation de la vitesse de convergence de la chaîne grâce à une adaptation de la forme des fonctions servant à tester les valeurs candidates à la forme de la fonction que l'on cherche à approximer.

L'idée de cette méthode est d'utiliser deux fonctions qui enveloppent la fonction densité cible, et convergent vers elle grâce â l'ajustement des deux fonctions à chaque étape du processus.

Principe de la méthode :

Le but de cette méthode est d'approximer une densité $f(x)$, que l'on ne sait pas simuler directement, par une densité g plus simple à simuler. Il s'agit alors de construire des enveloppes supérieures et inférieures à f et de tester si les valeurs candidates générées selon la fonction g sont comprises entre ces deux enveloppes et peuvent donc être considérées comme provenant de f.

Ces enveloppes supérieures et inférieures de la fonction f sont construites de manière séquentielles, et s'adaptent selon qu'il y a eu ou non acceptation de la valeur candidate générée.

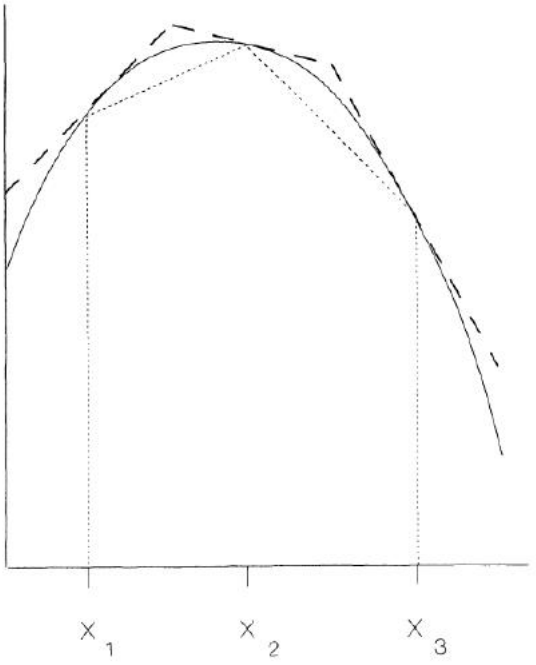


FIGURE 4.2 – Approximation donnée par la méthode de Rejet

Cette méthode nous permet ainsi de simuler les densités conditionnelles totales de u_i et v_i qui nous manquaient pour compléter l'estimation de tous les paramètres inconnus du modèle.

4.1.4 Résumé des étapes d'estimation

Étape 1. Modélisation du risque géographique par commune i, pour estimer au mieux l'effet géographique u_i non expliqué dans 1. Par l'utilisation de la théorie des modèles bayésiens hiérarchiques u_i et v_i deviennent aléatoires, et les paramètres respectifs de leur distribution τ et λ aussi. Ainsi le modèle s'écrit :

$$\begin{aligned} y|u,v &\sim p(y|u,v), \\ u|\tau &\sim p(u|\tau), \\ v|\lambda &\sim p(v|\lambda), \\ \tau,\lambda &\sim p(\tau,\lambda). \end{aligned}$$

L'estimation des paramètres inconnus u, v, τ et λ nécessite de simuler la loi du modèle, représentée par sa densité a posteriori. La densité a posteriori a une forme complexe, donc le problème d'estimation des paramètres se ramène à un problème de simulation de cette densité a posteriori.

Étape (2) Simulation de la densité a posteriori $p(u, v, \tau, \lambda|y)$ du modèle Utilisation de l'échantillonnage de Gibbs, méthode de type Monte Carlo par chaînes de Markov. Simuler $p(u, v, \tau, \lambda|y)$ revient à simuler les densités conditionnelles totales des paramètres $p(u_i|u_{-i}, v, \tau, \lambda, y)$, $p(v_i|v_{-i}, u, \tau, \lambda, y)$, $p(\tau|u, v, \lambda, y)$, $p(\lambda|u, v, \tau, y)$, qui ont des formes plus simples.

1. Simulation des densités conditionnelles totale de τ et λ Simulations par des lois du χ^2.
2. Simulation des densités conditionnelles totale de u_i et v_i Simulation par la méthode de rejet adaptatif de Gilks & Wild.

4.2 Lissage spatial par la théorie de la crédibilité

4.2.1 Théorie de la crédibilité (Modèle de Bülmann-Straub)

La théorie de la crédibilité est une technique mathématique fréquemment utilisée par les actuaires pour la tarification et la modélisation des risques. Son objectif est de proposer une tarification adaptée aux groupes à partir des données historiques.

Les principaux avantages de cette théorie sont la possibilité de modéliser un portefeuille hétérogène en terme de risque et l'application des modèles de crédibilité sur les différentes cibles à modéliser, soit la Prime pure, le coût moyen ou bien le nombre de sinistre, entre autres.

Parmi les différents modèles de crédibilité, nous allons présenter le modèle de Bühlmann-Straub. Ce modèle est une généralisation du modèle de Bühlmann tenant compte de l'exposition au risque des assurés, qui permet ainsi d'être plus adéquat à la performance du risque et au propos recherché.

Dans tout ce qui suit notons par I le représentant du portefeuille de risque, par $X_{i,j}$ la fréquence de sinistres du risque i pendant l'année j et par $\omega_{i,j}$ le poids associé. Notons que le risque i est caractérisé par le profil de risque θ_i qui est une réalisation de l'espace Θ_i. Afin de pouvoir employer le modèle de Bülmann-Straub il nous faut poser les deux hypothèses suivantes,

H1. Les variables aléatoires $X_{i,j}$ sont conditionnellement à Θ_i, indépendantes, et de moments conditionnels,

$$\mu(\Theta_i) = E[X_{i,j}|\Theta_i],$$
$$\sigma^2(\theta_i) = \omega_{i,j}\text{Var}[X_{i,j}|\Theta_i].$$

H2. Pour un j fixe, les couples $(\Theta_1, X_{1,j}), \cdots, (\Theta_I, X_{I,j})$ sont indépendants et les Θ_i sont indépendants et identiquement distribués pour $i = 1, \cdots, I$.

Remarquons que l'indépendance conditionnelle peut être relâchée en exigeant seulement la non-corrélation conditionnelle, $E[\text{Cov}(X_{i,k}, X_{i,l}|\Theta_i)] = 0$ pour $k \neq l$.

Avant de continuer avec la description du modèle on aura besoin d'introduire la notation suivante :

- Fréquence individuelle,
$$\mu(\Theta_i) := E[X_{i,j}|\Theta_i].$$
- Risque individuel normalisé (par $\omega_{i,j} = 1$),
$$\sigma^2(\Theta_i) := \omega_{i,j}\text{Var}[X_{i,j}|\Theta_i].$$
- Fréquence collective,
$$\mu_0 := E[\mu(\Theta_i)].$$

Alors pour la variance on dispose de la décomposition suivant :

$$\begin{aligned}\text{Var}[X_{i,j}] &= \text{Var}[E[X_{i,j}|\Theta_i]] + E[\text{Var}[X_{i,j}|\Theta_i]] \\ &= \text{Var}[\mu(\Theta_i)] + E\left[\frac{1}{\omega_{i,j}}\sigma^2(\Theta_i)\right].\end{aligned}$$

Notons,

$$\tau^2 := \text{Var}[\mu(\Theta_i)] \quad \text{et} \quad \sigma^2 := E\left[\frac{1}{\omega_{i,j}}\sigma^2(\Theta_i)\right].$$

Dans cette décomposition σ^2 est une mesure du risque interne au risque individuel alors que τ^2 mesure l'hétérogénéité du portefeuille.

Fixons maintenant i. Le meilleur estimateur de la fréquence individuelle sans biais de $\mu(\Theta_i)$ est,

$$\overline{X}_i = \sum_j \frac{\omega_{i,j}}{\omega_i} X_{j,j}, \quad \text{avec } \omega_i = \sum_j \omega_{i,j}.$$

Et,

$$\text{Var}[X_i] = \tau^2 + \frac{\sigma^2(\Theta_i)}{\omega_i}.$$

Pour chaque risque i, on cherche l'estimateur de crédibilité de la fréquence de sinistres individuelle $\mu(\Theta_i)$. Les risques étant indépendants, hypothèse (H2), $\widehat{\mu(\Theta_i)}$ dépend uniquement des observations X_i. Soit,

$$\widehat{\mu(\Theta_i)} = \beta_{i,0} + \sum_j \beta_{i,j} X_{i,j}.$$

L'estimateur de crédibilité est donc de la forme :

$$\mu(\widehat{\Theta_i}) = Z_i \overline{X}_i + (1 - Z_i)\mu_0.$$

Lequel doit satisfaire,

$$\mathrm{Cov}(\mu(\widehat{\Theta_i}), X_i) = Z_i \mathrm{Con}(X_i, X_i) = \mathrm{Cov}(\mu(\Theta_i), X_i).$$

Par conséquent pour le facteur de crédibilité on a l'expression suivante,

$$\begin{aligned} Z_i =& \frac{\mathrm{Cov}(\mu(\Theta_i), X_i)}{\mathrm{Cov}(X_i, X_i)} \\ =& \frac{E[\mathrm{Cov}(\mu(\Theta_i), X_i|\Theta_i)] + \mathrm{Cov}(\mu(\Theta_i), E[X_i|\Theta_i])}{\mathrm{Var}[X_i]} \\ =& \frac{E[\mathrm{Cov}(\mu(\Theta_i), X_i|\Theta_i)] + \mathrm{Var}[\mu(\Theta_i)]}{\mathrm{Var}[X_i]} \\ =& \frac{0 + \tau^2}{\tau^2 + \frac{\sigma^2(\Theta_i)}{\omega_i}} = \frac{\omega_i}{\omega_i + \frac{\sigma^2}{\tau^2}}. \end{aligned}$$

L'erreur quadratique moyenne de l'estimateur de la crédibilité est donc donnée par :

$$E\left[(\overline{\mu(\Theta_i)} - \mu(\Theta_i))^2\right] = (1 - Z_i)\tau^2 = Z_i \frac{\sigma^2}{\omega_i}.$$

Dans cette formule τ^2 correspond à l'erreur quadratique moyenne de μ_0 et $\frac{\sigma^2}{\omega_i}$ correspond à l'erreur quadratique de $\overline{X}_i$.

Passons maintenant à l'estimation empirique des variances. On a,

$$\begin{aligned} \overline{\sigma^2} &= \frac{1}{I(S-1)} \sum_{i=1}^{I} \sum_{j=1}^{S} \omega_{i,j} (X_{i,j} - \overline{X}_i)^2 \\ \overline{\tau^2} &= \left[\sum_{i=1}^{I} (X_i - \overline{X})^2 - (I-1)\overline{\sigma^2} \right]. \end{aligned}$$

Ces estimateurs sont sans biais convergents si $\Sigma_i \left(\frac{\omega_i}{\omega}\right)^2 \to 0$ lorsque $I \to \infty$ pour $\overline{\tau^2}$.

4.2.2 Adaptation de la théorie au lissage spatial «Distance»

Cette deuxième méthode a le même objectif que le lissage précédent, c'est-à-dire capturer l'effet géographique résiduel non expliqué par la méthode du GLM, qui permet ensuite de créer le nouveau zonier qui résume les effets externes et l'effet géographique résiduel.

Le lissage Distance, qui utilise la méthode de crédibilité exposée ci-dessus, lisse le risque d'une commune à partir de la moyenne de risque des autres communes pondérée par l'exposition et une fonction de la distance. La fonction en question assigne une influence plus grande aux communes plus proches. Lorsque nous appliquons le lissage «Distance» par la méthode de crédibilité, le risque d'une commune est lissée par ceux de toutes les autres communes en même temps, avec une influence décroissante avec l'éloignement des communes voisines. L'expérience lissée est basée sur une moyenne pondérée par l'exposition du code et des codes aux alentours, avec une influence plus grande assignée aux codes plus proches.

En formalisant le lissage des résidus qui nous intéressent à partir de la théorie de la crédibilité vue ci-dessus, on obtient, pour une commune $i = 1, \cdots, n$, que le lissage Distance du résidu r_i est donné par la formule suivante,

$$R_i = Z_i r_i + (1 - Z_i)\overline{r}_i.$$

Dans cette formule,

- La variable Z_i correspond au facteur de crédibilité. Lequel est donné par,

$$Z_i = \left(\frac{\omega_i}{\omega_i + \omega_0}\right)^l.$$

  Ici ω_i est le poids de la crédibilité, correspondant donc à l'exposition de chaque commune et ω_0 qui compense le poids de la crédibilité, correspondant donc au rapport des variances interclasses et intra-classes, qui seront estimées de manière empirique.
  La puissance de la crédibilité est mesurée par le paramètre l, à partir duquel on peut contrôler l'importance donnée au risque individuel par rapport au collectif.

- Le risque individuel r_i correspond au résidu de la commune i.
- Le risque collectif,

$$\overline{r}_i = \frac{\sum_{j \neq i} r_j d_{i,j}^P \omega_j}{\sum_{j \neq i} d_{i,j}^P \omega_j}, \quad j = 1, \cdots, i-1, i+1, \cdots, n$$

  avec $d_{i,j}$ la distance euclidienne entre le point recherché et les points connus aux alentours,

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

  $\overline{r}_i$ correspond donc à la moyenne des résidus voisins non lissés, pondérés par leur distance avec les autres communes et leur poids.

La formule du risque collectif est donnée par la méthode déterministe de l'interpolation pondérée par l'inverse de la distance.

Cette méthode repose principalement sur l'inverse de la distance élevée à une puissance mathématique, dont la valeur à estimer en un point de la zone d'étude est déterminée à l'aide de la moyenne pondérée des valeurs des points, où leur l'influence décroît avec la distance par rapport au point échantillonné. Le paramètre Puissance P, contrôle la signification des points connus sur les valeurs interpolées en fonction de leur distance par rapport au point en sortie. (Il s'agit d'un nombre positif et réel dont la valeur par défaut est 2). La définition d'une puissance plus élevée permet une concentration sur les points les plus proches. Lorsque la puissance augmente, les valeurs interpolées commencent à approcher la valeur du point d'échantillonnage le plus proche. Ainsi, les données proches auront plus d'influence et la surface comportera davantage de détails (sera moins lisse). Une valeur de puissance moins élevée accorde au contraire plus d'influence aux points environnants les plus éloignés, ce qui génère une surface plus lisse.

4.3 Lissages en pratique

Pour constituer le zonier, dans cette étude on utilise soit la théorie bayésienne, soit la théorie de la crédibilité, toutes deux présentées dans les sections précédentes. Concrètement, il considère les effets tarifaires et externes déjà estimés dans l'étape 1 grâce aux modèles linéaires généralisés, et commence à l'étape 2 où l'on projette le risque sur les communes de France.

L'effet géographique non expliqué à l'étape 1 se retrouve dans les *résidus 1* :

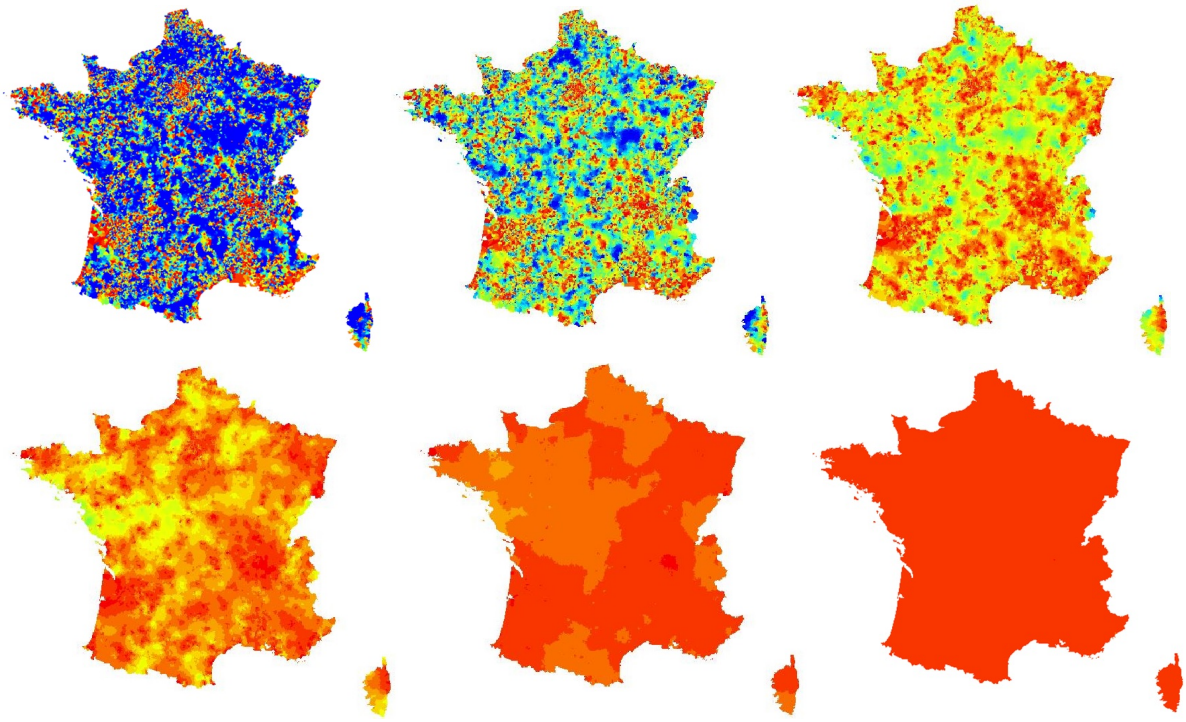
$$\textit{Résidus 1} = \textit{effet géographique inconnu} + \textit{erreur de modélisation.} \tag{4.3}$$

L'objectif est d'isoler dans ces résidus 1 un effet géographique significatif, que l'on appelle effet géographique résiduel. Ce dernier est obtenu grâce au lissage des résidus 1, tel que :

$$\textit{Résidus}1 = \textit{effet géographique résiduel} + \textit{bruit blanc.} \tag{4.4}$$

L'importance du lissage

Lorsqu'on projette les *résidus 1* sur la carte des communes, la plupart du temps on constate qu'ils ne sont pas distribués de manière aléatoire, car on visualise des zones de risque proche. A l'intérieur de ces zones, les valeurs des résidus sont parfois très hétérogènes entre deux communes voisines. Face à ce problème, on est conduit à réduire ces fortes variations pour mettre en évidence l'effet géographique réel de la zone. Si on ne fait aucun lissage, les variations sont trop fortes, et on pense que le caractère chaotique de ces variations est dû à l'échantillon : ces variations particulières ne se reproduisent pas sur un autre échantillon de données, car elles ne font pas parti du phénomène moyen. Autrement dit, si on laisse les résidus tels quels, le modèle n'est pas prédictif.



Les différences de risque entre les communes diminuent à chaque lissage, donc pour un

lissage total (dernier graphique) elles sont entièrement effacées, comme le montre le dernier graphique. Les risques sont alors égalisés les uns les autres à partir de leurs voisins, jusqu'à devenir entièrement égaux sur toute la carte. Cela équivaut à éliminer tout effet géographique, puisqu'une commune ne représente un surplus de risque par rapport à une autre. Cette situation n'est naturellement pas celle que l'on veut retenir, puisqu'on sait qu'il existe des différences de risque intrinsèques à chaque commune, le zonier servant justement à les synthétiser

À partir des graphes, on voit que le « bon »lissage est le résultat d'un juste milieu. D'un côté, lisser suffisamment pour constituer des différences de risque représentatives d'un effet prédictif. De l'autre côté, prendre garde à ne pas uniformiser le risque avec un lissage trop fort car on éliminerait trop d'information sur le risque géographique réel, alors qu'elle devrait être intégrée au zonier.

À noter : le zonier étant destiné à intégrer la structure tarifaire, il y a souvent une nécessité d'avoir des zones particulièrement homogènes, de manière à ne pas observer de zones radicalement différentes pour deux communes proches. Dans ce cas, on va privilégier des zones plus homogènes pour que les différences de prix soient plus acceptables sur le terrain.

4.3.1 Critères de décision

Somme des carrés des écarts, Standard Square Errors (SSE)

Les résidus lissés ont pour but de représenter l'effet géographique résiduel. Le point optimal de lissage explicatif est celui où l'information (les regroupements de risques découlant du lissage) est la plus prédictive. Pour choisir le meilleur lissage, on doit comparer les effets de prédiction des résidus lissés et identifier le lissage qui soit le plus juste. L'idée est donc de lisser les résidus sur une partie des données et vérifier si ils modélisent bien les résidus d'une autre partie des données. Le fait de comparer avec un autre échantillon de données nous permet de voir si cet effet est bien prédictif, c'est-à-dire si on a bien capturé l'effet moyen réel des résidus.

Soit $\hat{r}_l$ l'effet géographique résiduel capturé, il correspond aux résidus lissés d'une partie des données et r_p les résidus non lissés d'une autre partie des données. Le lissage à privilégier est ainsi celui qui minimise la différence entre eux. Le premier critère est donc le SSE appliqué aux résidus, parmi les l différents niveaux de lissages, on choisira a priori le lissage $\hat{r}_l$ qui minimise l'erreur sur l'ensemble des communes i :

$$\text{Min} \sum_i (\hat{r}_{i,l} - r_i)^2.$$

Voici le schéma utilisé pour obtenir le niveau de lissage optimal :

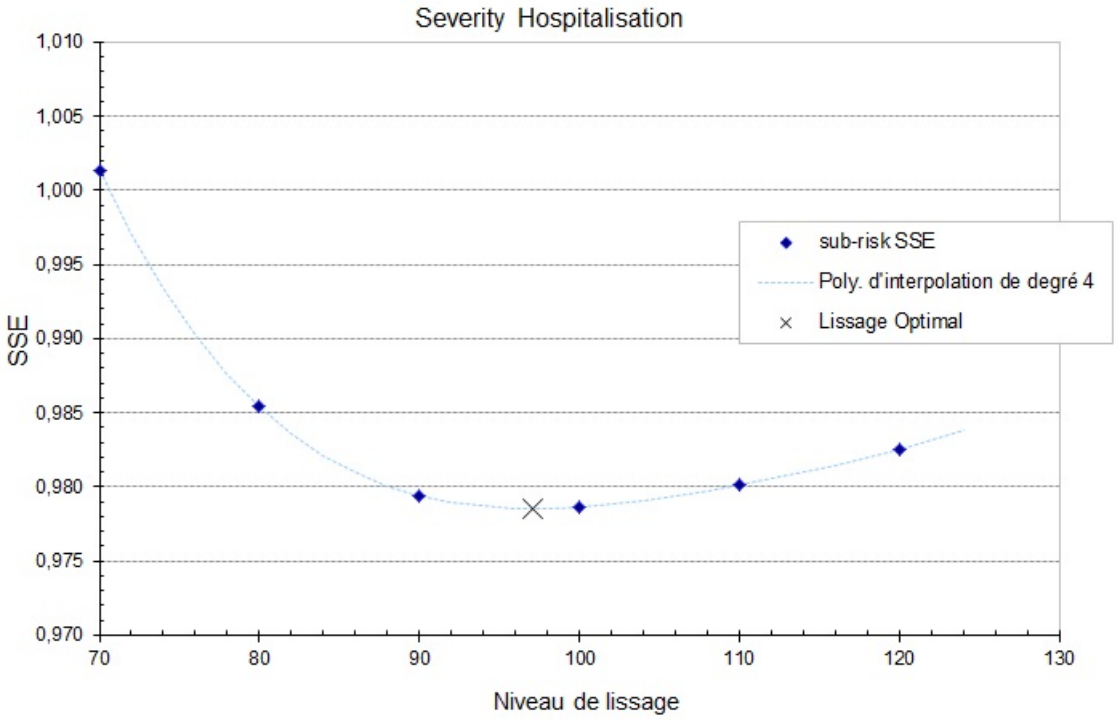


FIGURE 4.3 – Polynôme d'interpolation de degré 4

Le point de lissage optimal est obtenu en interpolant les points par un polynôme de degré quatre, dont on détermine le minimum en annulant sa dérivée première. L'idée de cette méthode correspond à celle exposée plus haut. En partant des résidus (lissage nul), augmenter le niveau de lissage fait d'abord gagner du pouvoir prédictif, ce qui se traduit donc par une courbe de SSE décroissante. Au bout d'un moment, si on continue à augmenter le lissage, on détruit de l'information en égalisant excessivement les résidus entre eux, la courbe des SSE se met alors à croître. Ainsi, le point optimal correspond bien au minimum de la courbe des SSE, dont le minimum du polynôme interpolateur fournit une bonne approximation. La forme de l'évolution est toujours la même : une décroissance suivie d'une croissance. Pour donner plus de flexibilité à l'approximation, on a utilisé un polynôme d'ordre quatre plutôt qu'un polynôme d'ordre deux. Cette forme étant régulière, avec l'intervalle par pas de dix choisi, les six points sont suffisants pour que l'approximation du polynôme soit performante.

QQPlot sur les résidus restants

Le deuxième critère répose sur l'analyse des erreurs de modélisation, correspondant au deuxème composant de la somme 4.3. On cherche cette fois-ci à démontrer que la structure de ces erreurs suit bien celle d'un « bruit blanc », correspondant au deuxème composant de la somme 4.4. En effet, si l'effet géographique résiduel est bien capturé, le résidu restant doit autant que possible être un bruit blanc, donc sa distribution empirique doit se rapprocher d'une distribution normale $N(0, \lambda)$. On va donc utiliser le critère de QQPlot pour comparer, empirique VS normale et ainsi choisir le lissage optimal.

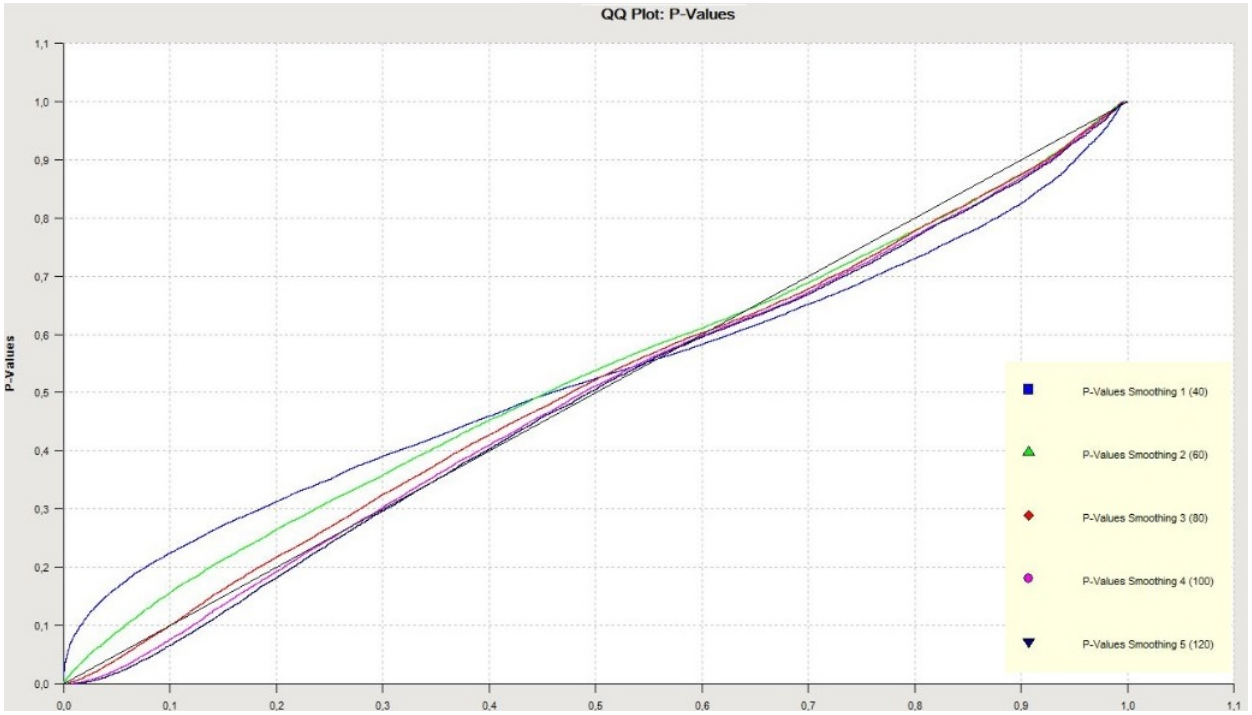


FIGURE 4.4 – QQPlot des différents niveaux de lissages

Pour la création du nouveau zonier, on a besoin de l'effet des variables externes combiné à l'effet géographique résiduel. Dans les étapes précèdentes on a réussi à identifier et estimer cette effet géographique total, pour la construction du zonier il est encore nécessaire de créer des classes qui regroupent les communes qui possèdent des risques similaires. Le regroupement sera effectué grâce à la méthode de classification suivante.

4.3.2 Classification CAH, Méthode de Ward

On souhaite rassembler les communes selon un critère de risque semblable, pour relever des zones relativement homogènes.

La classification hiérarchique ascendante effectue une classification en partant du côté le plus fin où chaque commune représente une classe. Puis en plusieurs étapes il réalise des regroupements jusqu'à l'obtention d'un seule classe qui contient tous les communes. Le principe de la méthode est le suivant,

- Initialement, les n communes constituent des classes à elles seules.
- On calcule les distances deux à deux entre communes et leur niveau de risque, et les deux communes les plus semblable sont réunis dans une même classe.
- La distance et le risque entre cette nouvelle classe et les communes restantes est ensuite calculée, et à nouveau les deux éléments (classes ou communes) les plus proches sont réunis.

Si on n'arrête pas ce processus, il continuera à regrouper jusqu'à ce qu'il ne reste plus qu'une unique classe constituée de toutes les communes.

Dans cette méthode il est nécessaire de définir deux distances.

1 **Distance entre communes** : Il semble donc préférable d'utiliser une distance euclidienne pondérée par le poids des communes.

2 **Distance entre classes** : La méthode de Ward est la distance choisit, il consiste

à regrouper les classes de façon à ce que l'augmentation de l'inertie interclasse soit maximum [1] . La distance entre deux classes est celle de leurs barycentres au carré, pondérée par les effectifs des deux clusters.

Definition 4.3.1. *La distance de Ward entre deux centres de classes* (C_j, C_l) *de barycentres respectifs* x_{C_j} *et* x_{C_l} *est définie par,*

$$D_W^2(C_j, C_l) = \frac{n_j n_l}{n_j + n_l} ||x_{C_j} - x_{C_l}||^2.$$

Il faut noter que lorsque l'on réalise une classification en utilisant la distance de WARD, à chaque étape on effectue une fusion optimale au sens de la conservation de l'inertie intra-classe. Cette optimalité locale (à chaque étape) ne garantie pas l'optimalité globale, c'est à dire l'identification du nombre de classes optimal. Cependant, on fait l'hypothèse que la partition trouvée à partir du regroupement optimal local à chaque étape est proche de la partition optimale globale.

4.4 Comparaison des méthodes de lissage Adjacency VS Distance

Dans un premier temps, on projète les *résidus 1* sur la carte des communes de France, pour vérifier si il y a réellement besoin d'effectuer un lissage.

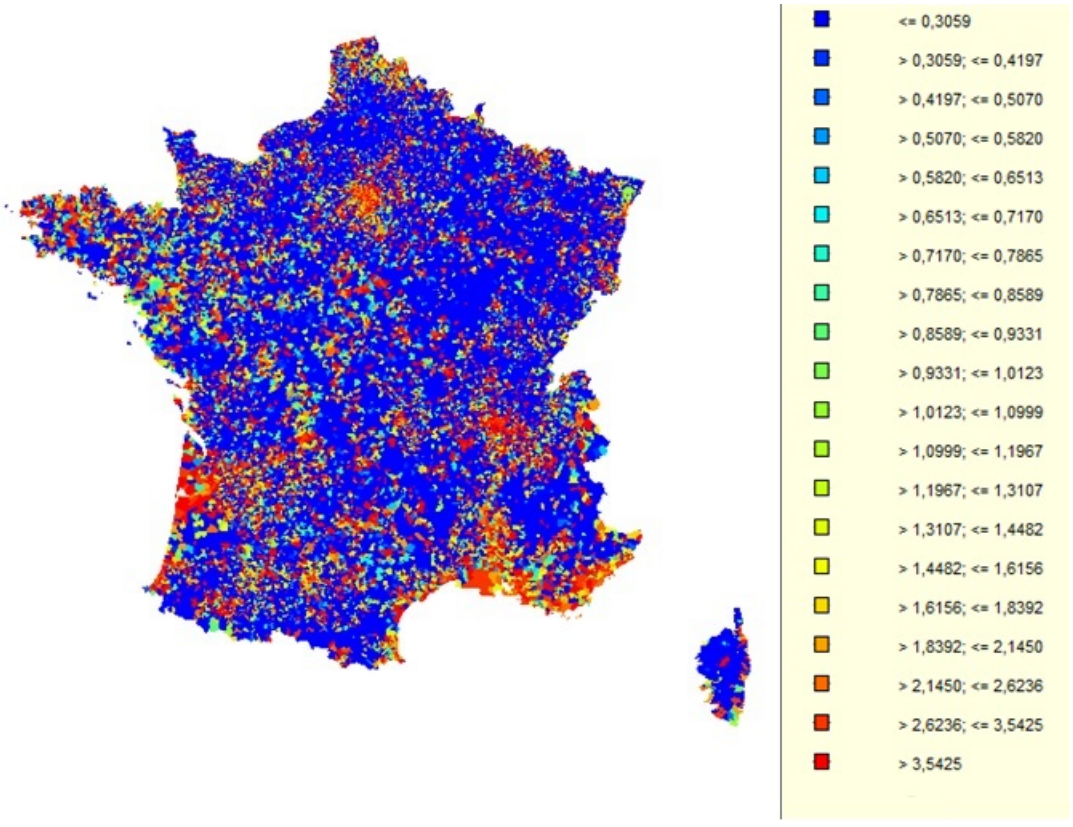


FIGURE 4.5 – Projection des Résidus sur la carte des communes de France

1. L'augmentation de l'inertie interclasse revient au même d'après le théorème de Huygens, à ce que l'augmentation de l'inertie intraclasse soit minimum

On constate que l'effet des *résidus* 1 n'est pas uniformément réparti sur la carte, car on distingue des groupes de risque. Il y a donc bien un effet géographique non expliqué par les variables externes. Néanmoins, les groupes ne sont pas homogènes, il reste du bruit qui perturbe la carte qu'il faut séparer pour bien distinguer l'effet géographique résiduel. C'est pourquoi il est nécéssaire de lisser ces *résidus* 1.

Scénarios des lissages

Le lissage consiste à laisser les résidus se faire plus ou moins influencer par leurs voisins, directs pour les cas Adjacency et plus largement étendus pour les cas Distance. En pratique, on va donc jouer sur l'intensité de cette influence, jusqu'à obtenir l'effet géographique jugé représentatif et explicatif de la réalité, l'effet qui se réalise en moyenne. Le problème est donc de mieux visualiser l'impact du lissage sur la carte, et de savoir sur quels critères on peut se baser pour juger que le lissage est suffisant. On s'intéresse à l'effet qui compose le zonier en plus de celui issu des variables externes, il s'agit de l'effet géographique résiduel. Il est noté u_i dans la méthode « Adjacency », et r_i dans la méthode « Distance ». L'initialisation du processus de lissage pose les *résidus 1* comme les observations des variables aléatoires u_i pour la méthode « Adjacency », et comme les valeurs de r_i pour la méthode « Distance ».

Ensuite, on veut jouer sur l'intensité du lissage. Pour Adjacency, on répéte le processus de lissage via un nombre de simulations de u_i plus important. Pour Distance, on applique différents valeurs pour le paramètre de puissance de distance. Les valeurs des résidus lissés changent et se rapprochent donc tous un peu de leur voisins, pour projeter des risques moyens semblables. Dans la méthode Adjacency, on part de ces nouvelles valeurs et en lissant à nouveau, les u_i se rapprochent de plus en plus de leurs voisins. La variance des u_i diminue ainsi à chaque itération. Ainsi de suite, jusqu'à ce que de proches en proches, la carte tende à être uniforme en appliquant un grand nombre de lissage successifs. Dans la méthode Distance, un paramètre de distance de plus en plus petit conduit de même à uniformiser la carte. La difficulté du lissage n'est pas seulement dans leur application, mais aussi dans le choix du lissage le plus juste.

4.4.1 Comparatif au niveau du lissage

Lissage optimal

Pour l'obtention du lissage optimal pour chaque méthode, on a testé 6 paramètres de lissage différents, qu'on applique dans un premier temps sur un échantillon témoin correspondant aux résidus des années 2011 à 2013. Pour évaluer leur capacité de prédiction, on les compare avec l'échantillon de validation correspondant aux résidus de l'année 2014. On rappelle que l'effet sélectionné doit représenter un phénomène moyen, qui puisse se reproduire dans le temps.

Grâce aux critères de décision exposés dans la partie 4.3.1, on sélectionne le paramètre où l'erreur de prédiction est la moins grande et celle dont la distribution est la plus proche de la loi normale, c'est-à-dire celle qui minimise les SSE et qui ajuste le mieux le graphique QQplop.
Les tableaux suivants représentent les SSE moyens associés à chaque lissage :

Adjacency paramètre	SSE moyen
50	15,16
60	14,49
70	14,11
80	13,96
90	13,95
100	14,01

TABLE 4.1 – SSE moyen Méthode Adjacency

Distance paramètre	SSE moyen
0, 01	18,19
0, 005	18,23
0, 001	17,54
0, 0005	17,07
0, 0001	16,68
0, 00005	16,72

TABLE 4.2 – SSE moyen Méthode Distance

Les 2 paramètres font varier l'intensité du lissage. Pour la méthode Adjacency, le paramètre est lié au nombre d'itérations de simulation de lissage, tandis que pour la méthode Distance, le paramètre est lié à la puissance de l'inverse de l'équation de distance. En interpolant ces paramètres par un polynôme de degré 4 et en vérifiant les QQPlot, on conclut que les niveaux de lissage optimaux sont de 86 pour Adjacency et 0,0001 pour Distance.

On remarque que les SSE Adjacency sont plus bas, ce qui signifie que les résidus sont plus prédictifs avec cette méthode. Toutefois, cet indicateur n'est à lui seul pas suffisant pour choisir la méthode optimale, il faut le compléter par l'analyse graphique détaillée ci-dessous.

Cartes des effets géographiques totaux

On applique les valeurs de paramètre de lissage optimal sur les *résidus* 1, et on ajoute les résidus lissés à l'effet des variables externes. On aboutit ainsi aux cartes des effets géographiques totaux de chaque méthode, respectivement Adjacency et Distance représentées ci-dessous.

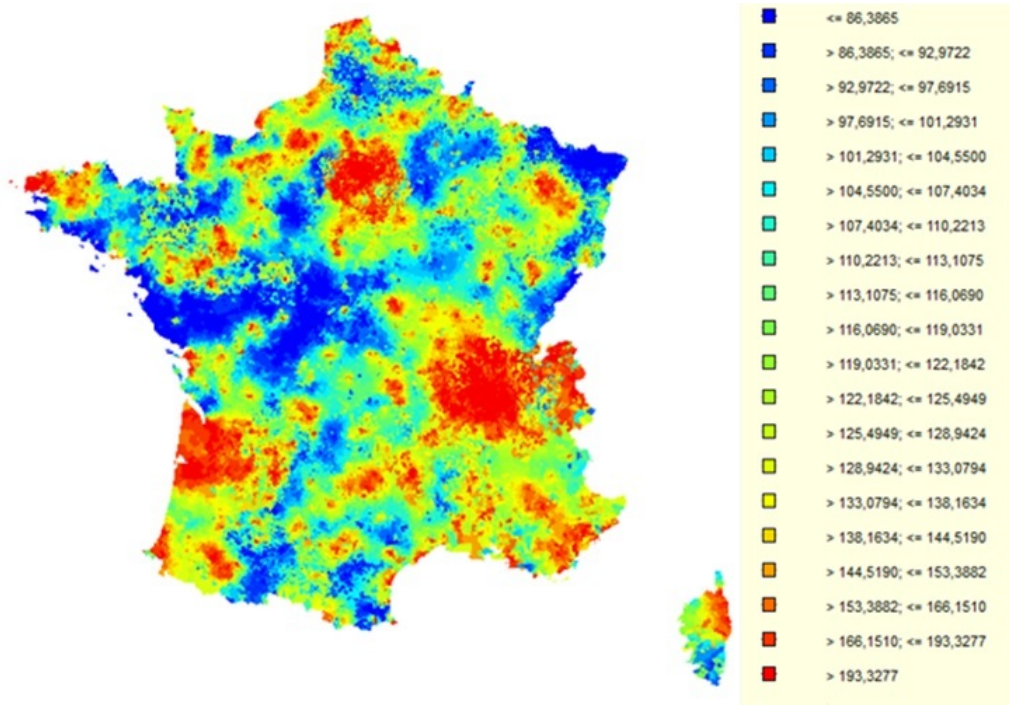


FIGURE 4.6 – Carte Méthode Adjacency

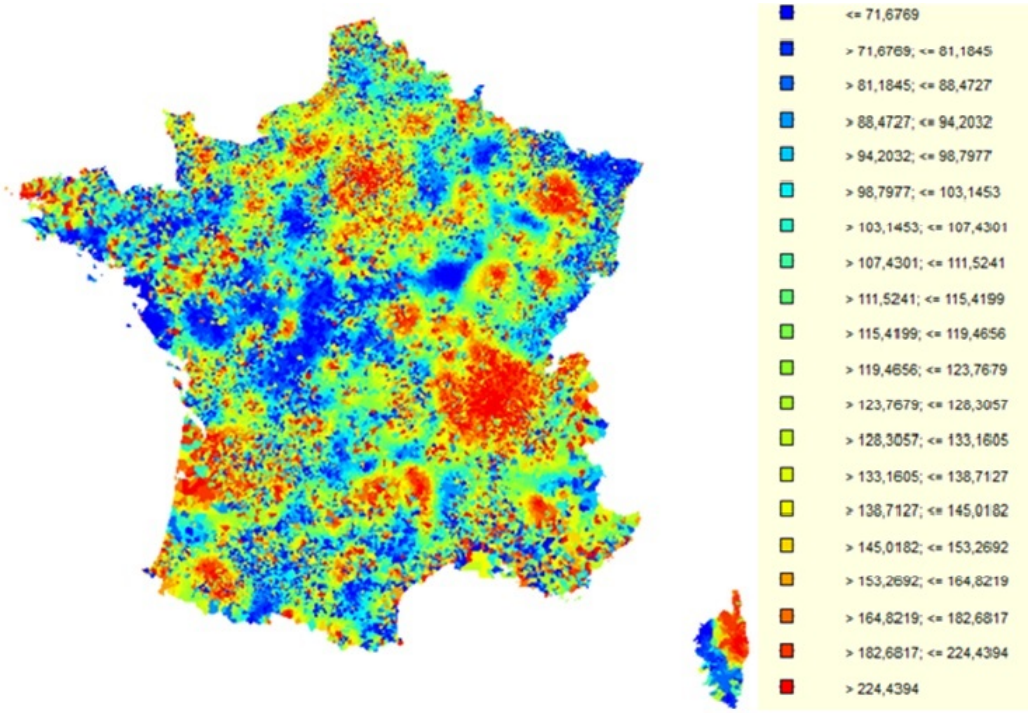


FIGURE 4.7 – Carte Méthode Distance

On remarque que les zones qui se détachent sont globalement les mêmes dans les deux cartes, pourtant elles sont plus homogènes sur la première carte. Dans la méthode Distance, même avec un paramètre optimal, on constate que les zones ne sont en effet pas aussi bien distinguées que dans la méthode Adjacency. Il est possible que ce soit dû au fait que la

méthode Distance est déterministe.

Au niveau de l'effet géographique total, en combinant les analyses SSE et les cartes, la méthode Adjacency semble être la plus satisfaisante.

4.4.2 Comparatif au niveau du zonier

On complète la comparaison des deux méthodes en faisant l'analyse descriptive des effets géographiques au sein de chaque zone finale :

Zone Adjacency	Poids	Moyenne	Ecart-type
1	6,6%	80,95	6,21
2	15,9%	97,53	4,53
3	23,9%	112,07	4,68
4	16,6%	126,29	3,77
5	14,2%	141,84	5,08
6	5,5%	157,19	3,48
7	6,2%	170,43	4,16
8	5,7%	186,95	5,75
9	2,5%	214,62	8,57
10	2,0%	243,22	10,95
11	0,6%	291,15	11,48
12	0,2%	344,13	10,15
13	0,0%	434,50	17,92
14	0,0%	485,08	4,84
15	0,0%	868,29	0,00

TABLE 4.3 – Tableau descriptif Méthode Adjacency

Zone Distance	Poids	Moyenne	Ecart-type
1	3,9%	57,39	6,39
2	15,5%	79,96	6,50
3	34,0%	107,12	9,39
4	21,2%	136,18	7,98
5	13,5%	164,88	7,82
6	4,0%	190,31	6,55
7	2,7%	217,32	8,51
8	2,4%	257,80	15,67
9	1,2%	305,84	16,41
10	0,6%	373,78	16,97
11	0,5%	466,78	34,01
12	0,3%	633,17	67,37
13	0,2%	858,33	48,80
14	0,0%	1060,75	109,20
15	0,0%	1643,37	125,67

TABLE 4.4 – Tableau descriptif Méthode Distance

En comparant les deux premiers tableaux, on remarque que les zones sont réparties

différemment entre les deux méthodes. En effet, 89% de l'exposition Adjacency est concentrée sur les zones de 1 à 7, tandis que cette même proportion est déjà atteinte sur les zones de 1 à 5 avec la méthode Distance. La méthode Adjacency présente donc une meilleure répartition des zones. Cette remarque est importante sur des zones de faible poids, sur lesquelles on ne peut pas faire confiance en leurs moyennes car elles risquent de ne pas se reproduire dans le temps. Ainsi, même si l'intervalle [moyenne min ; moyenne max] est plus grand avec la méthode Distance, il n'est pas pertinent de conclure que cette dernière offre une meilleure segmentation, car les moyennes y sont moins bien exposées. En outre, les variances intra-zones sont plus faibles que celles de la méthode Distance.

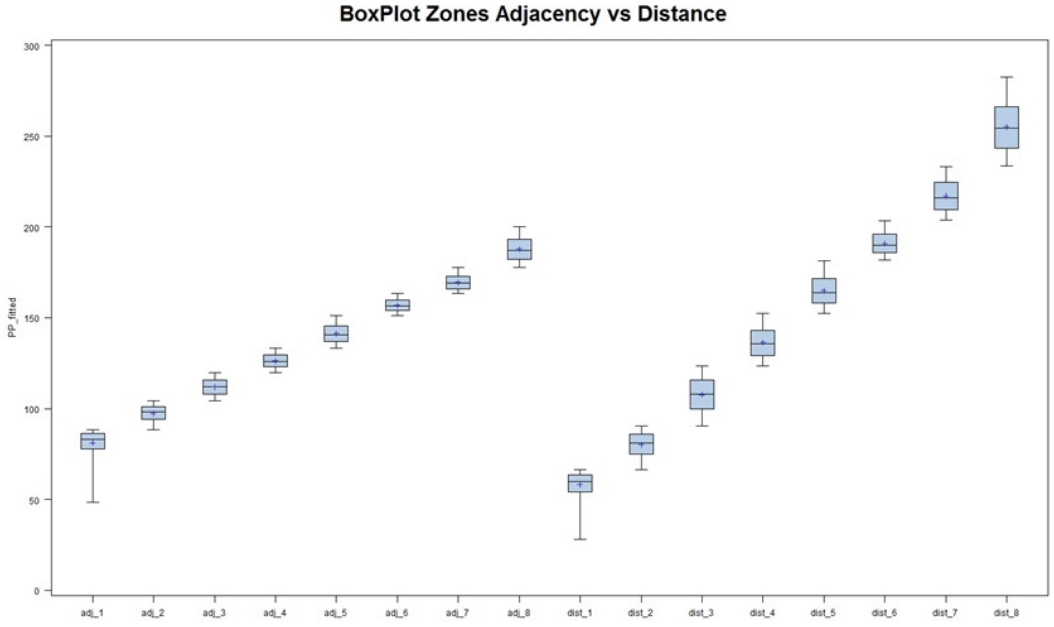


FIGURE 4.8 – Comparaison des Boît à moustaches

En analysant le box plot sur les huit premières zones (celles où l'exposition est significative), on remarque que la répartition de l'effet Adjacency est plus homogène à l'intérieur de chaque zone, puisque la boîte à moustache est moins étendue qu'avec la méthode Distance.

Une analyse a également été menée pour tester l'homogénéité des zones à l'intérieur de chaque département, le support des résultats sont dans l'annexe section **A.6** :

- 17,7% du total des départements de la France présentent plus de 9 zones différentes avec la méthode de lissage Adjacency ;
- 79,2% du total des départements de la France présentent plus de 9 zones différentes avec la méthode de lissage Distance ;
- En moyenne, à l'intérieur d'un même département, la méthode Distance présente 3 zones de plus que la méthode Adjacency.

Suite à l'analyse descriptive, la méthode Adjacency semble toujours la plus cohérente pour représenter le risque géographique.

4.4.3 Comparatif au niveau du modèle de prime pure

Comparons maintenant l'impact de la zone lorsqu'on la réintroduit dans le modèle de prime pure. Dans un premier temps, on veut voir si le modèle final, composé par les variables tarifaires (hors actuel zonier) plus le nouveau zonier est plus ou moins adapté que

le modèle avant lissage des *résidus*1, c'est-à-dire celui avec variables tarifaires (hors actuel zonier et avec forçage sur Alsace-Moselle) plus variables externes. L'intérêt est de voir si le lissage des résidus a permis de capturer une information supplémentaire significative par rapport à celle déjà contenue dans les variables externes, et si, remplacer l'ensemble des segmentations externes par la segmentation de la zone ne fait pas perdre trop d'information.

	Apport des Zones Adjacency dans le modèle
P-valeur test χ^2	$0,0\%$
Variation Déviance	(-) 92.792,6
Variation AIC	(-) 57.736,5
Variation BIC	(-) 57.749,4

TABLE 4.5 – Tableau d'apport du lissage Adjacency au modèle

	Apport des Zones Distance dans le modèle
P-valeur test χ^2	$0,0\%$
Variation Déviance	(-) 427.186,7
Variation AIC	(-) 287.229,0
Variation BIC	(-) 287.274,9

TABLE 4.6 – Tableau d'apport du lissage Distance au modèle

Indépendamment de la comparaison des méthodes, on constate qu'échanger les variables externes par le nouveau zonier fait gagner de l'information au modèle. L'étape du lissage a donc permis de récupérer une information géographique significative pour l'ajustement du modèle de prime pure, en plus de celle des variables externes dont l'effet est aussi synthétisé dans le zonier.

La déviance baisse beaucoup plus avec la méthode de la Distance. Cela est dû au fait que le lissage conserve plus de variations résiduelles, donc l'estimé est au final plus proche de l'observé. Mais ce meilleur ajustement aux données de l'échantillon se paie par un pouvoir prédictif moins fort, comme l'a prouvé l'analyse des SSE au chapitre 4.2.1.

Ci-dessous l'effet marginal de la zone dans les modèles de prime pure, décliné sur les deux méthodes :

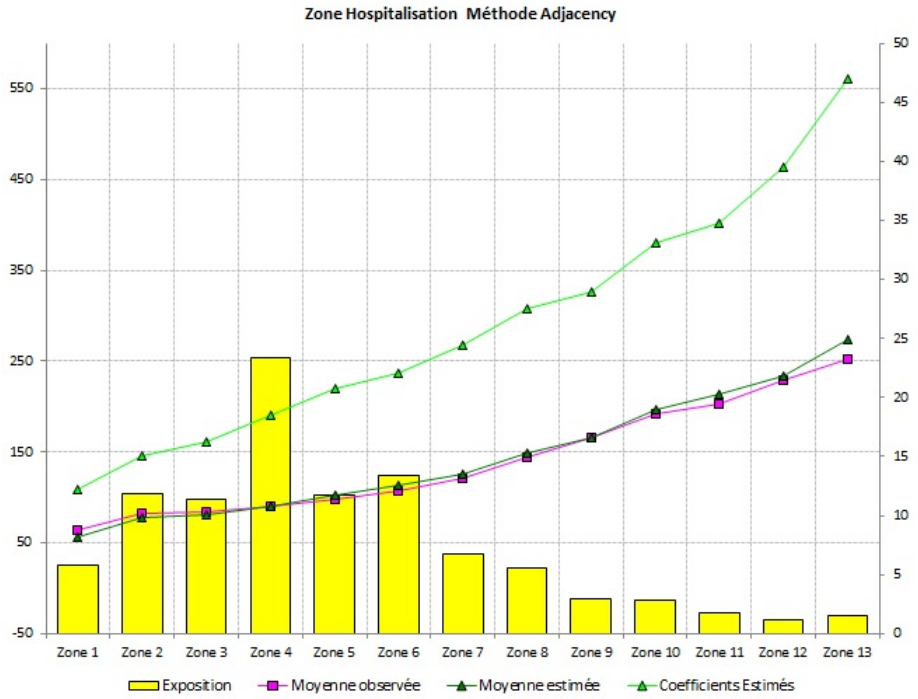


FIGURE 4.9 – Modèle de Prime pure Hospitalisation Méthode Adjacency

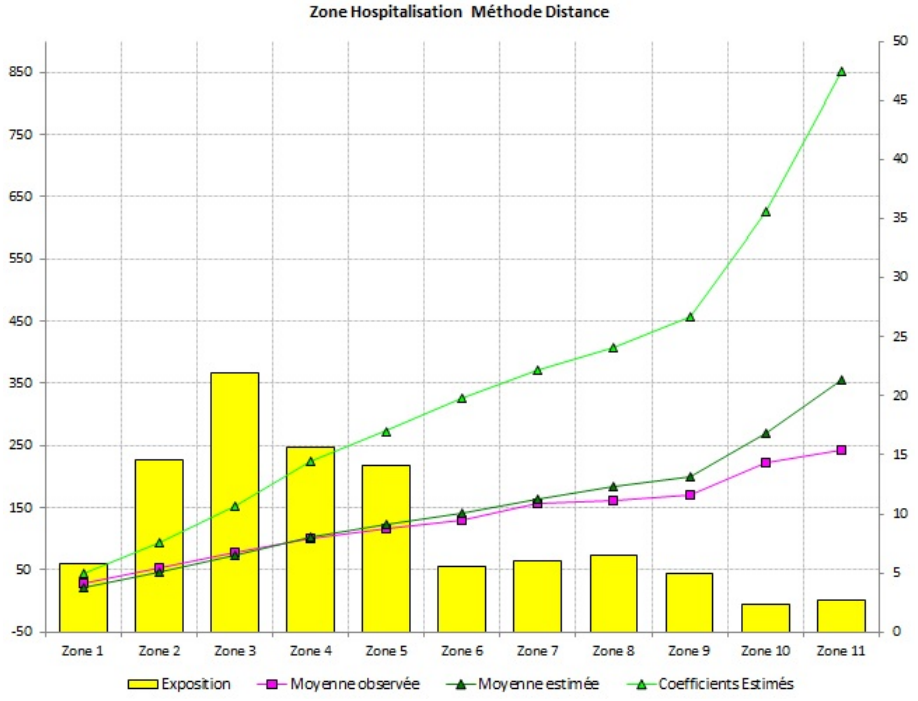


FIGURE 4.10 – Modèle de Prime pure Hospitalisation Méthode Distance

Les Betas de la zone Adjacency sont plus réguliers que ceux de la distance. De plus, comme on l'a indiqué lors de l'analyse descriptive, la méthode Adjacency inclut une meilleure distribution de l'exposition entre les zones, chacun des coefficients est donc plus fiable.

Conclusion

Pour la comparaison des deux méthodes de lissage, les analyses suivantes ont été réalisées :

- Choix du lissage optimal :
  - SSE de l'effet géographique résiduel ;
  - Analyse graphique : cartes de l'effet géographique total.
- Analyse descriptive par zone : statistiques de base et Box-Plot ;
- Impact de la zone sur le modèle de prime pure.

A chaque analyse, la méthode Adjacency s'est révélée la plus robuste avec un pouvoir prédictif plus important, regroupements de risques mieux distingués, zones mieux distribuées, risque géographique plus homogène à l'intérieur de chaque zone, linéarité de l'effet marginal de la zone dans le modèle de la prime pure. C'est cette méthode qui a donc été utilisée pour estimer la zone de chaque poste Santé.

4.5 Valorisation des hypothèses de modélisation

4.5.1 Apport des variables externes

Dans la construction du zonier, l'effet géographique est partagé en deux : une partie expliquée par les variables externes, et l'autre résultant du lissage de résidus du modèle de prime pure. L'objectif est ici de mettre en relief l'impact d'insérer ou non ces variables externes, au niveau du zonier hospitalisation.

Ci-dessous, à gauche on constate l'effet observé moyen non expliqué par les variables tarifaires, ou dit autrement, l'observé moyen standardisé de l'effet des variables tarifaires. On confronte cet observé moyen avec l'effet marginal des variables externes qui correspond au cumul des coefficients estimés pour les variables externes dans le modèle de prime pure (carte de droite).

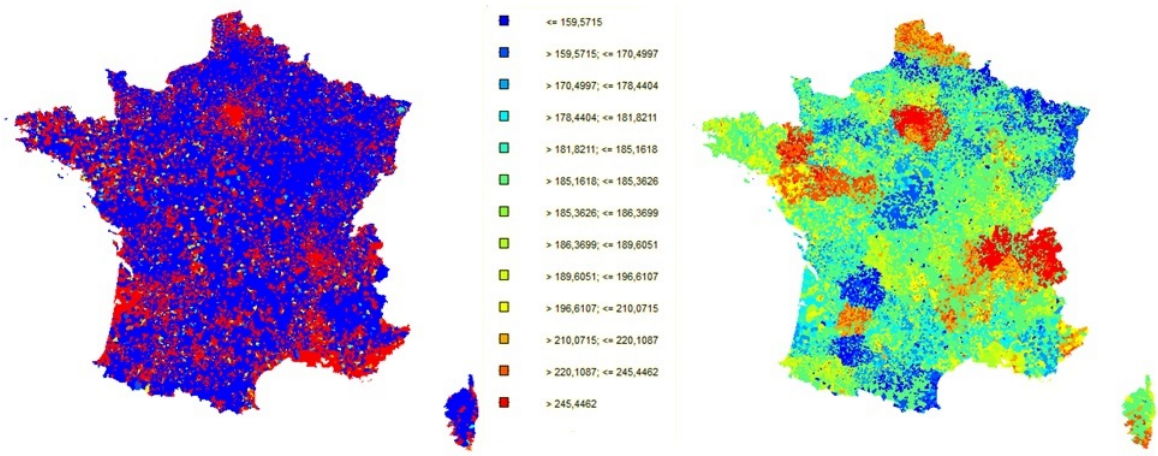


FIGURE 4.11 – Cartes effet obervé moyen vs effet marginal des variables externes

En ajoutant des variables externes au modèle de prime pure, une partie de l'information géographique est déjà intégrée au modèle, l'effet non expliqué est donc moins important, ce

qui équivaut à des résidus avec moins de variations. Il y a alors moins de lissage à effectuer sur les résidus par la suite.

Concernant l'effet des variables externes en lui-même on voit que l'information a une segmentation visible par département. Cela est dû à ce qu'une partie des variables externes sélectionnées sur le modèle hospitalisation est renseignée au niveau départemental. Dans

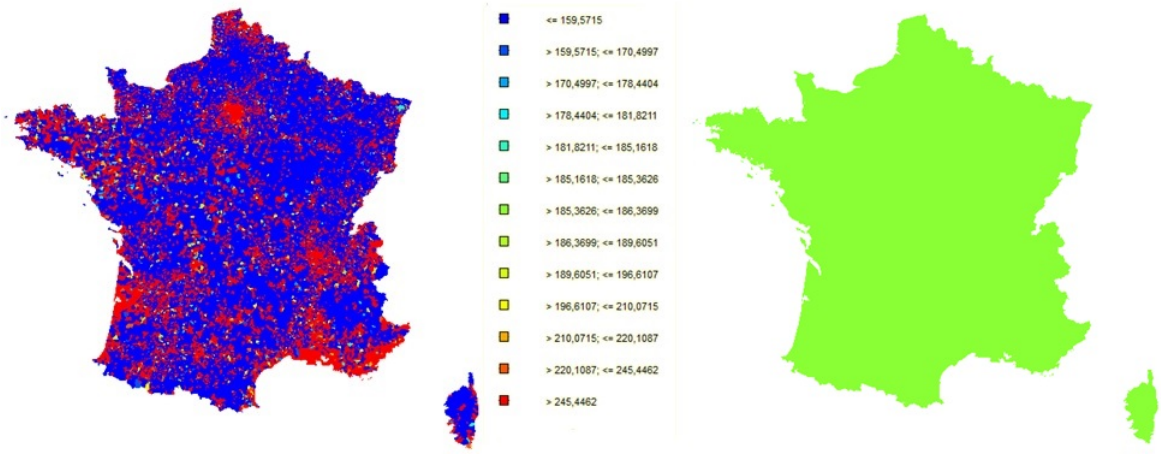


FIGURE 4.12 – Cartes effet obervé moyen vs effet marginal sans variables externes

le cas où les variables externes ne sont pas présentes, l'effet géographique est à construire en intégralité grâce au lissage des résidus.

A partir du lissage des résidus, on arrive à construire les classes d'effet géographique total. On compare ensuite le zonier final dans les deux situations suivantes :

$$\begin{aligned}\text{Zonier Final avec variables externes} &= \text{effet externes} + \text{effet } \textit{résidus 1} \text{ lissés},\\ \text{Zonier Final sans variables externes} &= \text{effet externes} + \text{effet } \textit{résidus*} \text{ lissés}.\end{aligned}$$

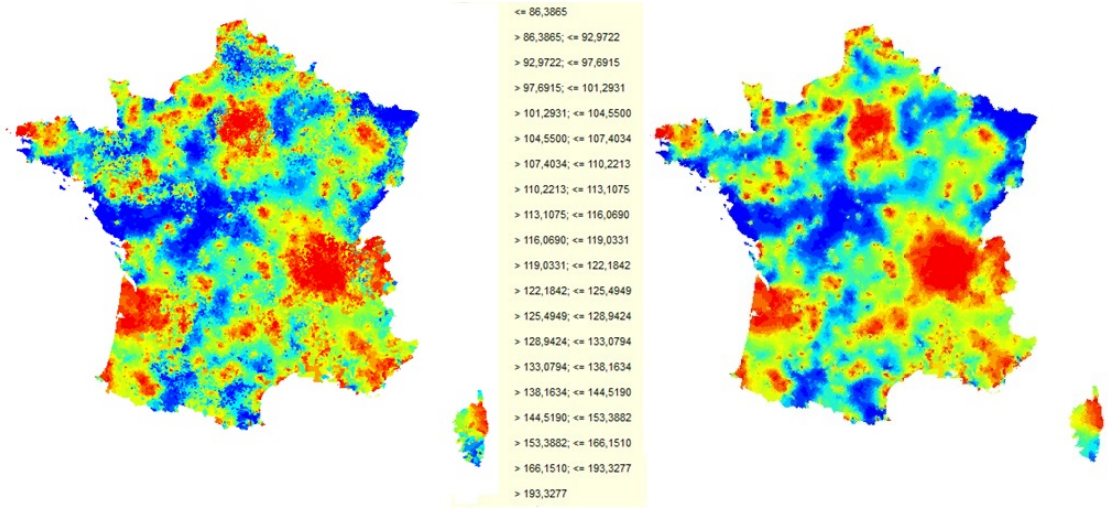


FIGURE 4.13 – Cartes Zonier Finale avec variables externes VS Zonier Finale sans variables externes

On remarque que l'effet géographique est plus lisse sans variables externes. Ce résultat était attendu, car dans ce cas le lissage s'applique sur une plus large étendue de résidus,*résidus**, et l'effet géographique provient uniquement des simulations de lissage faites sur eux. Par contre, lorsque les variables externes sont présentes, l'effet total doit composer avec l'effet estimé par les variables externes en plus de celui des résidus lissés.

Par ailleurs, on arrive à comparer l'effet marginal de chaque zonier sur le modèle de prime, visible dans le graphique suivant :

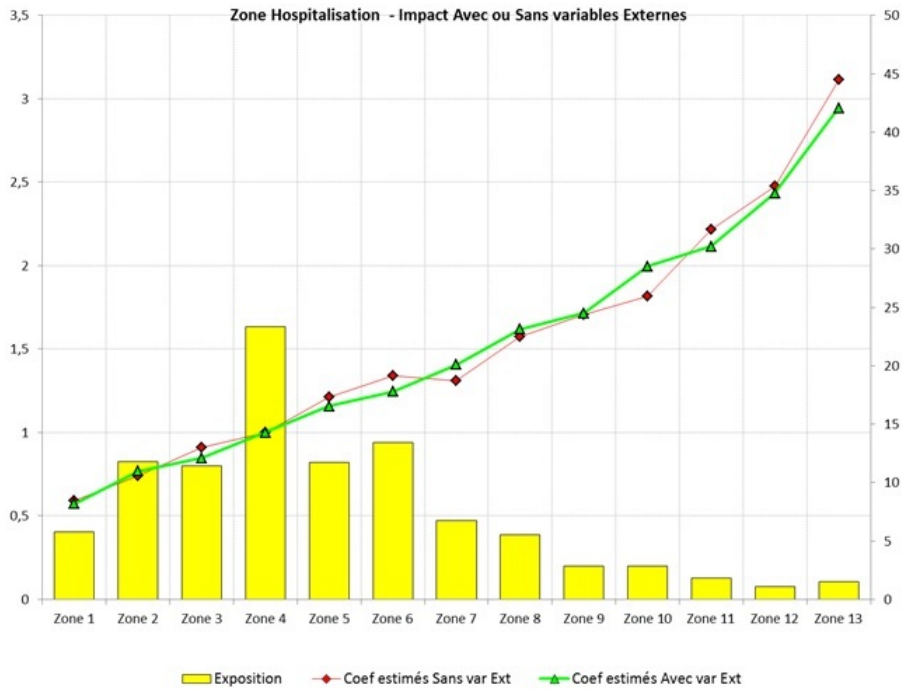


FIGURE 4.14 – Zones Finales avec variables externes VS Zones Finales sans variables externes

On constate que la tendance des coefficients estimés est comparable, mais celle avec variables externes reste plus linéaire que celle sans variables externes, avec une croissance des coefficients plus homogène. Le zonier avec variables externes apparaît donc plus cohérent.

L'intérêt des variables externes est de pouvoir donner une connaissance plus riche des facteurs globaux qui expliquent le risque. Si ces facteurs changent, on saura ainsi que le risque géographique est lui aussi susceptible de changer, ce qui aide à contrôler si le zonier est encore fonctionnel.

En conclusion, la présence des variables externes dans le modèle de prime pure en amont de la construction du zonier constitue un choix optimal. L'effet marginal de la zone se retrouve plus linéaire grâce à elles, et ces dernières constituent de plus un bon outil de pilotage du zonier.

4.5.2 Valorisation du forçage sur l'Alsace Moselle

Un ajustement indispensable à la construction du zonier a été la neutralisation du coefficient estimé de l'Alsace-Moselle présente dans la variable tarifaire régime du modèle de prime pure. L'idée était de transférer l'estimation de cette modalité sur la partie géographique du modèle. On va illustrer dans cette partie l'impact de cet ajustement sur le zonier

hospitalisation, en le regardant avec ou sans forçage.

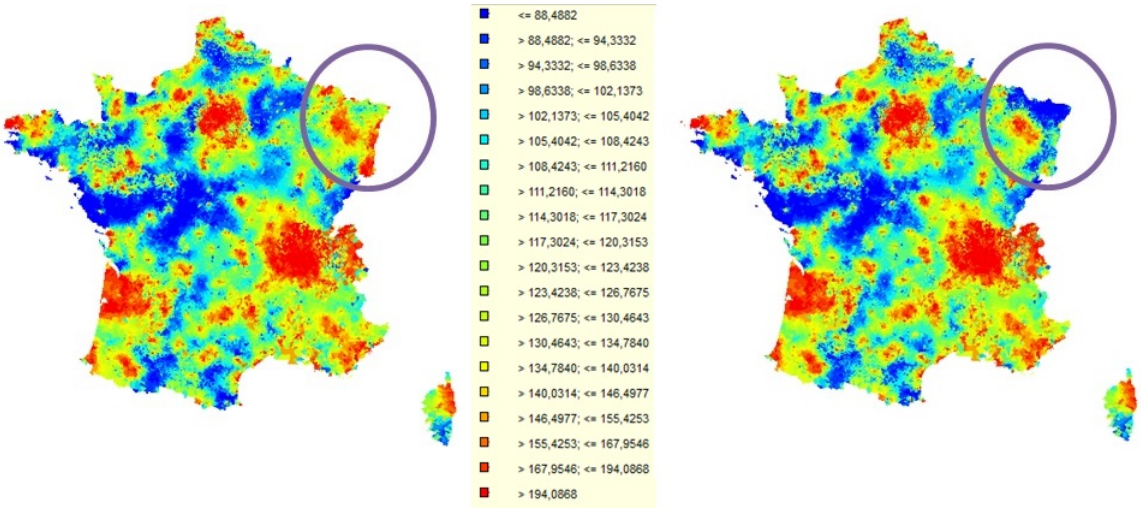


FIGURE 4.15 – Zones Finales à partir du modèle sans forçage VS Zones Finales à partir du modèle avec forçage

Les cartes sont sensiblement identiques, comme prévu les seuls départements qui se distinguent sont ceux de l'Alsace-Moselle :

- Sans forçage, l'Alsace-Moselle prend sur lui toute la baisse du risque, et la moyenne estimée se retrouve plus basse que la moyenne observée. Ainsi le modèle sous-estime le risque Alsace-Moselle. Les résidus ensuite lissés doivent impulser un effet à la hausse pour mieux coller à la réalité, ce qui donne comme résultat un risque géographique élevé pour l'Alsace-Moselle, visible en rouge sur la carte de gauche.
- Avec forçage, l'impact du risque Alsace-Moselle est synthétisé dans le zonier, on retrouve donc le faible risque géographique hospitalisation propre à cette région, effectivement coloré en bleu sur la carte de droite.

Les graphiques ci-dessous illustrent ces deux cas de figure :

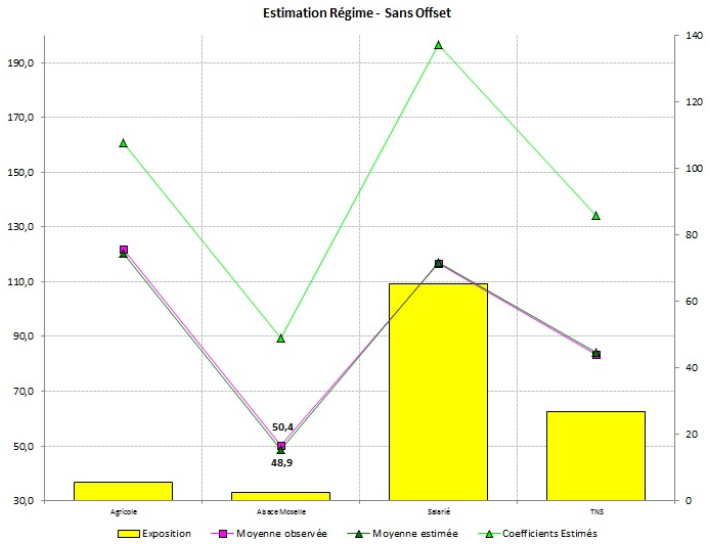


FIGURE 4.16 – Coefficients variable Régime à partir du modèle sans forçage

Le modèle sans forçage sous-estime la modalité Alsace-Moselle, la moyenne estimée (couleur vert foncé) est en effet en dessous de la moyenne observée (couleur rose). Ce qui se traduit par les résidus supérieurs à 1 précédemment observé sur la carte.

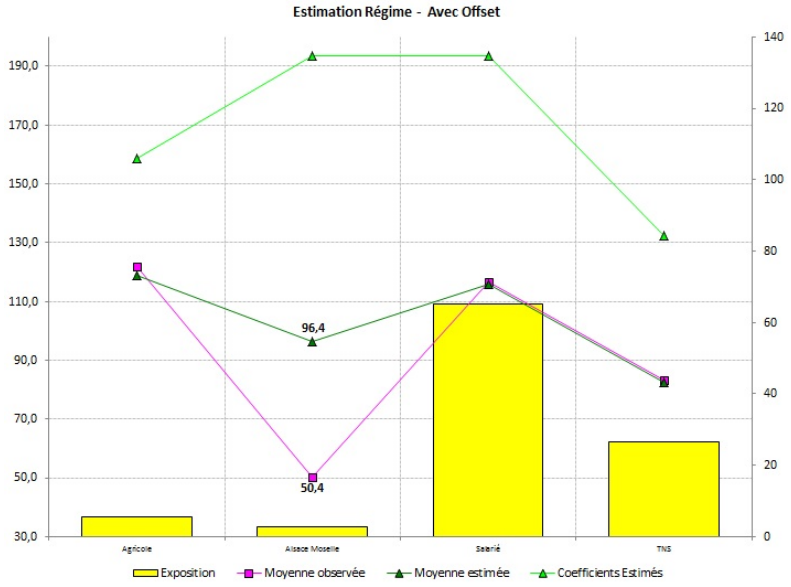


FIGURE 4.17 – Coefficients variable Régime à partir du modèle avec forçage

Dans le modèle avec forçage, l'Alsace-Moselle est surestimée, de manière à ce que son effet d'abaissement du risque se répercute dans un résidu alors faible. Plus tard, cette information est récupérée dans le zonier qui lui affecte une zone de faible risque.

Enfin, on détaille plus précisément les zones et leur répartition à l'intérieur de l'Alsace-Moselle, selon que le forçage Alsace-Moselle est présent ou non :

	Zones sans forçage	Zones avec forçage
Mode	6	1
Minimum	2	1
Quantile à 5%	3	1
Quantile à 25%	5	1
Quantile à 50%	6	2
Quantile à 75%	8	3
Quantile à 95%	10	5
Maximum	13	7
Moyenne	6,36	2,52
Ecart-type	1,78	1,19

TABLE 4.7 – Tableau comparatif des distributions des zones

On observe bien une position et une répartition des zones très différentes. Dans le cas sans forçage, la zone la plus représentée dans les communes Alsace-Moselle est la zone 6. Alors qu'avec forçage la zone 1 correspond à la plus représentée, ce qui reflète donc mieux le vrai risque géographique associée à cette région.

4.6 Zonier Final, tous postes confondus

4.6.1 Composition du zonier

La construction du zonier global s'effectue en suivant différentes étapes. D'abord, la sélection du lissage optimal appliqué aux résidus de chaque poste. Cette étape permet d'obtenir l'effet géographique total par poste, qui est le résultat de l'effet marginal des variables externes auquel on ajoute le résidu lissé.

Ensuite, on récupère l'effet géographique total tous les postes confondus. Cet effet s'obtienne en faisant la sommes des effets géographiques obtenues à partir de l'estimation sur chaque poste.

$$\Pi_{Globale} = \Pi_{Hospitalisation} + \Pi_{Soinscourants} + \Pi_{Pharmacie} + \Pi_{Dentaire} + \Pi_{Optique} \tag{4.5}$$

Rappelons que la prime pure du poste Soins Courants est obtenue par la combinaison, du modèle Fréquence avec le Moèle Coût Moyen. La prime $\Pi_{Globale}$ est répartie géographiquement, et finalement un classement est effectue par la méthode de Ward. Cette étape donne les 14 zones constituant le zonier final tous postes.

L'effet marginal du zonier avant regroupement :

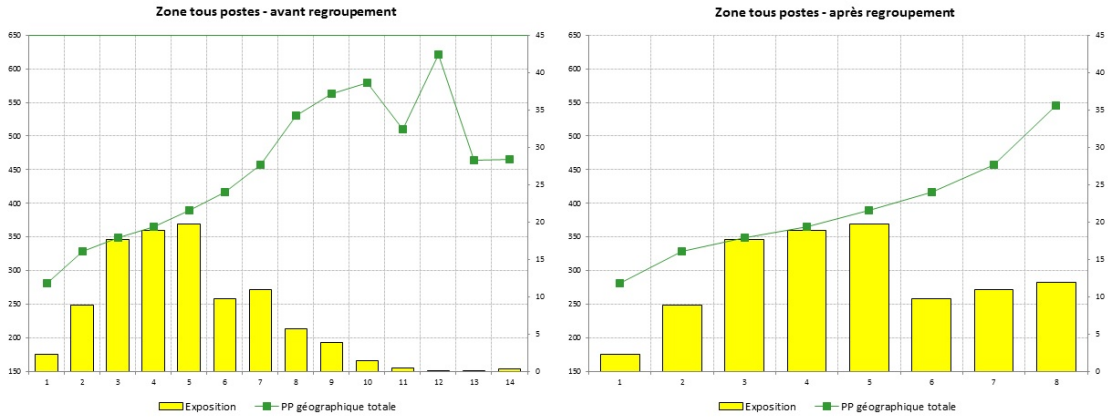


FIGURE 4.18 – Distribution des zones finaux

Il est plus cohérent que la prime pure observée, soit croissante avec la zone, c'est pourquoi on choisit de regrouper les zones de 8 et plus. On regarde ensuite comment la prime pure géographique totale est segmentée par les zones :

Zone Tous Postes	Poids	Moyenne	Ecart-type
1	3,2%	280,11	27,01
2	20,6%	350,42	15,77
3	32,9%	398,29	14,65
4	19,2%	447,74	13,94
5	12,0%	502,25	17,33
6	4,3%	550,51	11,29
7	5,4%	600,89	15,50
8	2,2%	848,00	362,40

TABLE 4.8 – Tableau des statistiques descriptives des zones finaux

4.6.2 Impact du zonier global sur chaque poste

Dans un contexte d'entreprise, on est souvent amené à utiliser un unique zonier santé, la plupart du temps en raison de contraintes informatiques qui limitent le nombre de variables. Si on doit choisir un seul zonier, il doit donc être un bon représentant du risque géographique de chacun des postes. On veut donc tout d'abord voir dans quelle mesure ce zonier recoupe les zoniers par poste, par une analyse des corrélations :

V de Cramer	1.	2.	3.	4.	5.	6.
1. Zones Hospitalisation	1,00	0,00	0,00	0,00	0,00	0,00
2. Zones Pharmacie	0,12	1,00	0,00	0,00	0,00	0,00
3. Zones Dentaire	0,15	0,11	0,00	0,00	0,00	0,00
4. Zones Optique	0,14	0,12	0,19	1,00	0,00	0,00
5. Zones Soins Courants	0,17	0,12	0,16	0,12	1,00	0,00
6. Zones Tous postes	0,41	0,24	0,55	0,25	0,29	1,00

TABLE 4.9 – Tableau de corrélation des zoniers par poste et final

On remarque tout d'abord que les zoniers par poste sont très peu corrélés entre eux. A l'inverse, la zone globale se trouve assez bien corrélée avec tous les postes, en particulier avec hospitalisation et dentaire. On s'attend donc à un effet marginal de la zone globale proche d'une tendance linèaire sur ces deux postes.

On incorpore la zone globale dans le modèle de prime pure de chaque poste, en remplaçant le zonier du poste par le zonier global. De cette manière, on peut tester si la zone globale est aussi bien adaptée pour approcher le risque géographique du poste.

Poste Hospitalisation

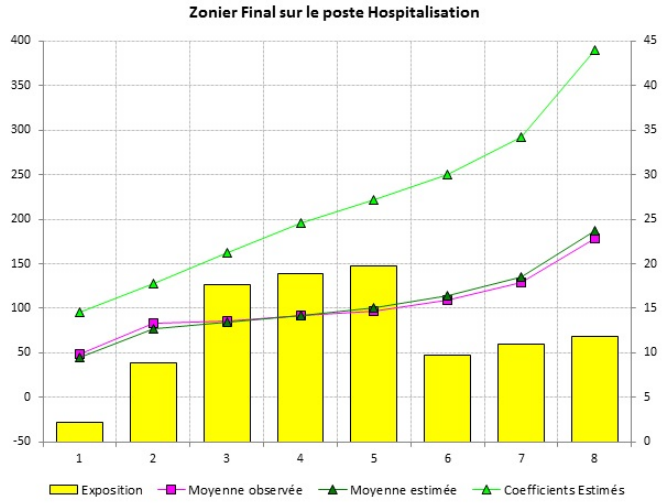


FIGURE 4.19 – Effet marginal du Zonier Final sur la prime pure du poste Hospitalisation

L'effet marginal de la zone globale sur la prime pure hospitalisation est quasiment linéaire et la segmentation est très satisfaisante, elle est croissante et régulière. Le zonier global s'adapte donc très bien au poste hospitalisation.

Poste Pharmacie

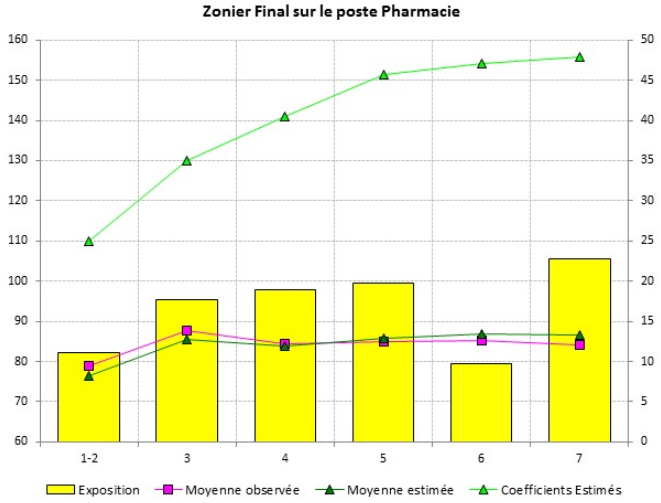


FIGURE 4.20 – Effet marginal du Zonier Final sur la prime pure du poste Pharmacie

Sur le poste Pharmacie, l'effet marginal n'est pas cohérent avec l'observé, on a donc ressemblé d'une part les zones de 7 et 8, et d'autre part les zones 1 et 2. Suite au regroupement, la tendance est bien croissante, mais on remarque que le zonier global sous-estime les zones 1 à 3, tandis qu'il surestime les zones 6 et 7. Il manque un effet marginal décroissant seulement présent dans le zonier propre à la pharmacie. Le zonier global semble donc y être moins bien adapté, mais on remarque aussi que l'effet géographique semble difficile à estimer. En effet, l'observé ne présente pas de tendance réelle, il semble donc difficile de distinguer une composante géographique pour la consommation moyenne en pharmacie.

Poste Dentaire

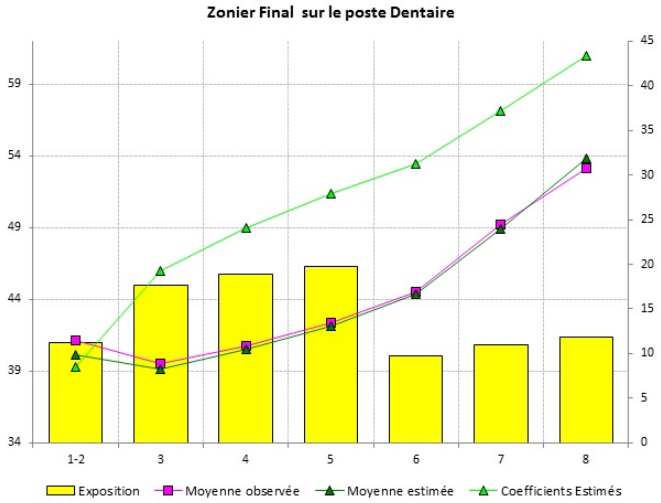


FIGURE 4.21 – Effet marginal du Zonier Final sur la prime pure du poste Soins Dentaire

L'effet de la zone globale est bien linéaire sur le poste dentaire, seule la zone 1 présente un effet incohérent, on a dû la regroupe donc avec la zone 2.

Poste Optique

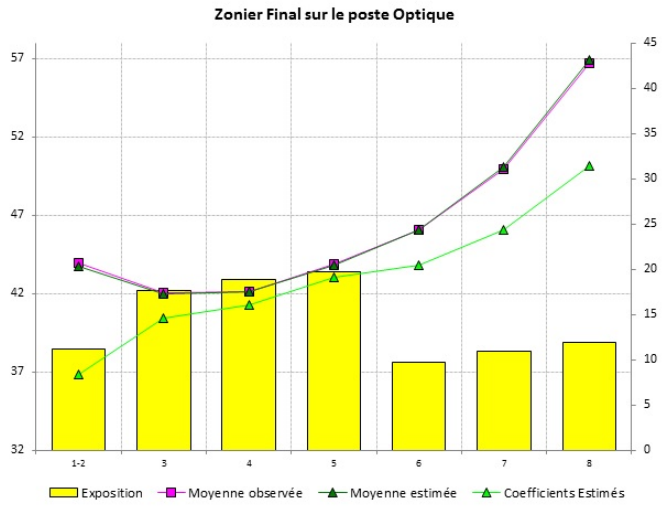


FIGURE 4.22 – Effet marginal du Zonier Final sur la prime pure du poste Optique

Le poste optique présente la même particularité que le poste dentaire, on procède donc au même regroupement des zones 1 et 2. On ne veut pas aller plus loin dans les regroupements, même si l'effet sur la zone 1-2 semble étrange, car il n'est pas conseillé qu'une zone regroupe présente trop d'exposition à elle seule.

Poste Soins Courants

Modèle de Fréquence :

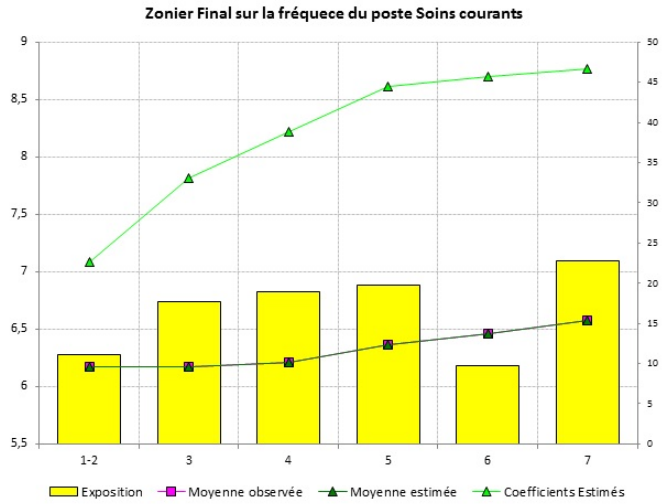


FIGURE 4.23 – Effet marginal du Zonier Final sur la fréquence du poste Soins courants

L'effet marginal est globalement croissant, mais contraire à l'observé sur les très petites et très grandes zones, on a donc fait les regroupements des zones 1 + 2 et 7+8.

Modèle de Coût moyen :

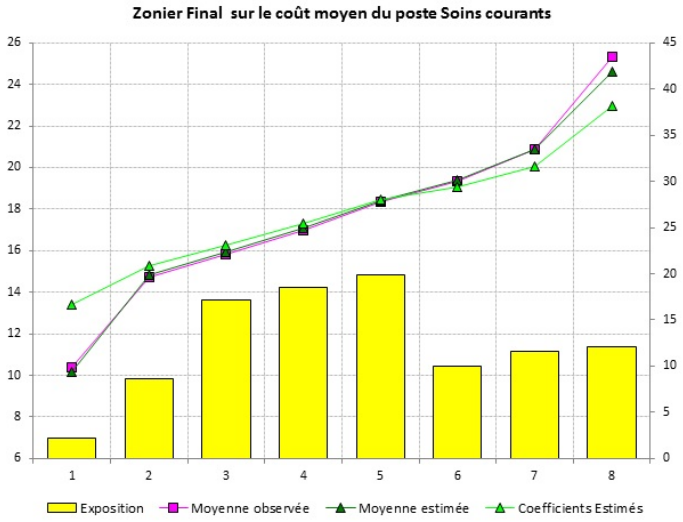


FIGURE 4.24 – Effet marginal du Zonier Final sur le coût moyen du poste Soins courants

La zone globale s'adapte bien à l'explication du coût moyen soins courants, aucun regroupement n'est donc nécessaire.

Exemple d'effet du zonier propre au poste vs zonier final

En supplément de l'analyse graphique ci-dessus, on présente sur le poste hospitalisation l'évolution des indicateurs de significativité du modèle lorsqu'on remplace la zone du poste par la zone globale :

	Echange zones hospitalisation par zones globales
P-valeur test χ^2	$0,0\%$
Variation Déviance	(+) 61.773,5
Variation AIC	(-) 38.808,1
Variation BIC	(-) 38.762,3

Remplacement de l'effet de gauche (propre au poste) par l'effet de droite (tous postes confondus) :

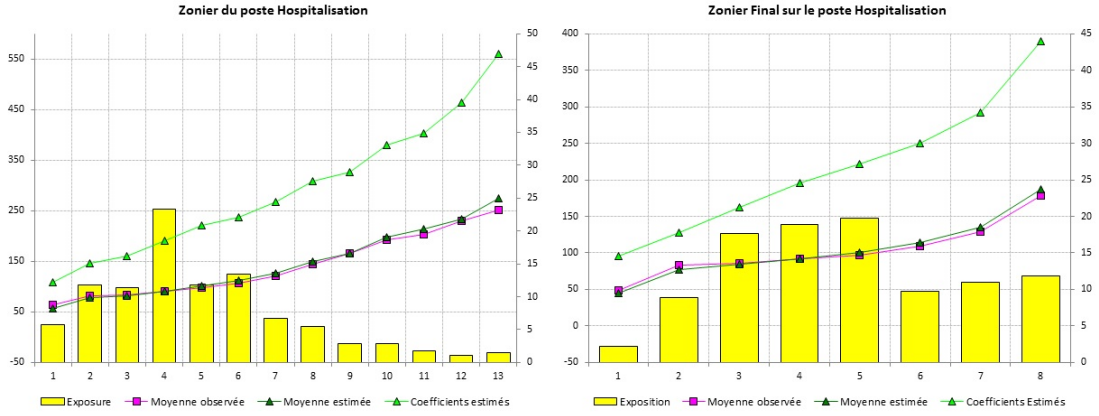


FIGURE 4.25 – Zonier propre au poste VS Zonier final sur la prime pure Hospitalisation

On observe comme attendu une perte d'information au modèle. Il est toujours mieux de privilégier la zone du poste puisqu'elle est plus proche du risque propre au poste, mais on a vu que graphiquement la zone globale reste cohérente et offre une bonne segmentation même à l'intérieur de chaque poste.

4.6.3 Comparaison du nouveau zonier VS l'actuel zonier

La meilleure estimation de la prime pure totale s'obtient en sommant les modèles de primes pures de chaque poste. Ainsi on peut obtenir les distributions de prime pure estimée suivantes, en fonction du zonier utilisé sur chacun des postes :

• Prime pure estimée à partir de la somme des primes pures des postes en utilisant son propre zonier :

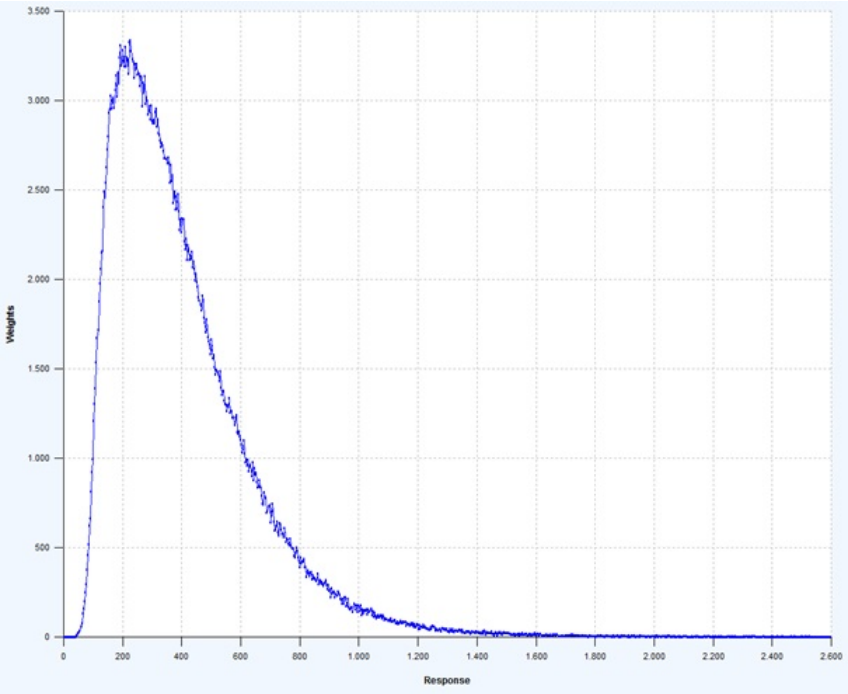


FIGURE 4.26 – Distribution de la prime pure avec zoniers propres au poste

• Prime pure estimée à partir de la somme des primes pures des postes en utilisant le zonier final :

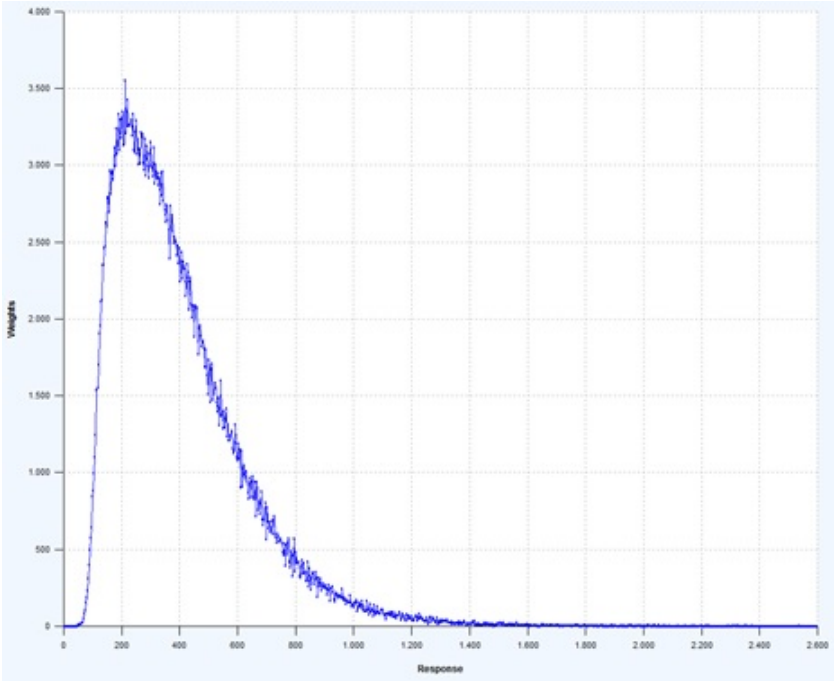


FIGURE 4.27 – Distribution de la prime pure avec zonier globale

• Prime pure estimée à partir de la somme des primes pures des postes en utilisant l'actuel zonier :

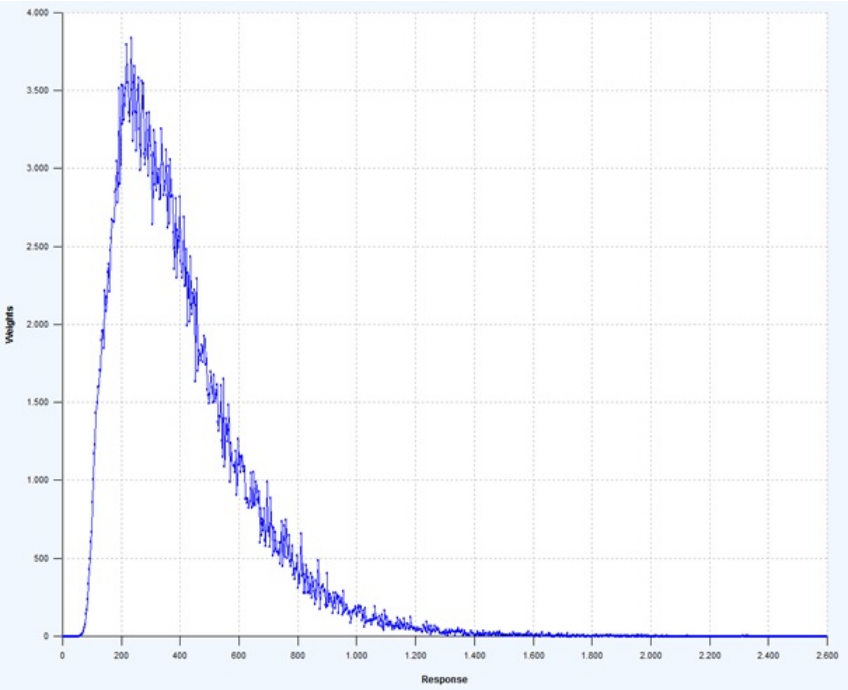


FIGURE 4.28 – Distribution de la prime pure avec actuel zonier

A noter : la distribution de la prime pure estimée avec l'actuelle zone présente plus de

variations que lorsqu'on utilise le nouveau zonier global. En outre, l'estimation sur la base des zoniers par poste reste la plus lisse.

Ces distributions sont le résultat des cinq ajustements de prime pure pour chacun des postes. Pour comparer la performance des zones, on devrait donc analyser l'apport de la zone pour chaque poste. Néanmoins, du fait de contraintes de temps et informatiques, la révision de la tarification est faite sur la base de l'estimation directe de la prime pure totale, sans passer par l'estimation des primes pures de chaque poste. De cette manière, la variable à expliquer est directement la consommation moyenne globale en santé.

Pour l'approbation du nouveau zonier, on a effectué la comparaison entre l'apport de l'actuel zonier versus l'apport du zonier global. Pour la comparaison souhaitée, on a estimé la prime pure en utilisant l'actuel zonier et d'autre part, on a estimé la prime pure en utilisant le zonier global qu'on a obtenu comme résultat dans ce mémoire.

On regarde l'impact de la nouvelle zone VS l'actuelle zone au niveau des indicateurs de significativité :

	Echange actuelles zones par les zones finales
P-valeur test χ^2	$0,0\%$
Variation Déviance	(-) 954,21
Variation AIC	(-) 4.485,37
Variation BIC	(-) 4.485,33

Lorsqu'on remplace l'actuelle zone par la nouvelle, la déviance AIC et BIC s'améliorent substantiellement. Finalement, le nouveau zonier est plus performant et permet ainsi un meilleur ajustement aux données.

Comparaison graphique

Au niveau de l'estimation des coefficients :

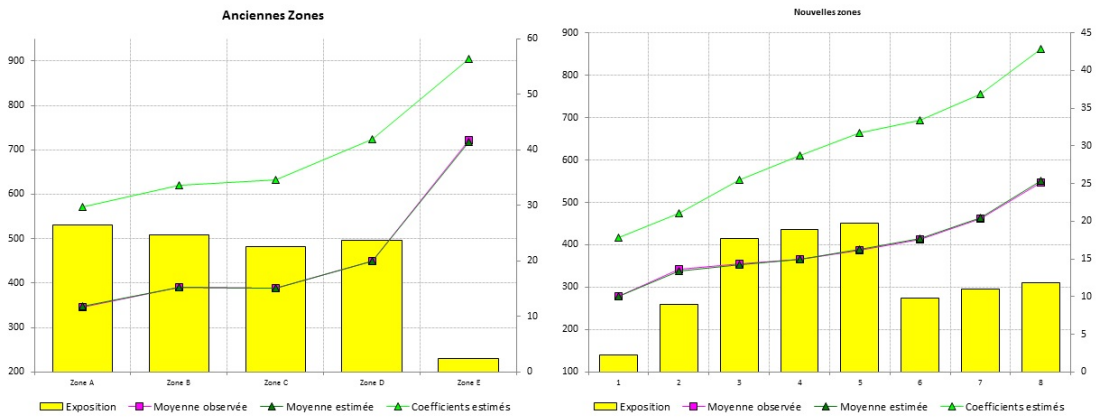


FIGURE 4.29 – Effet marginal des anciennes zones VS Effet marginal nouvelles zones sur la prime pure globale

Le nombre des zones de l'actuel zonier est plus réduit que celui du nouveau zonier. Si globalement les anciennes zones croissent à mesure que le risque augmente, ce qui est plus ou moins cohérent avec la sinistralité étudiée, les coefficients obtenus restent assez proches

entre eux et on constate que les zones B et C présentent un risque sensiblement identique. Au contraire, les coefficients des nouvelles zones sont plus éloignés les uns des autres, elles présentent en conséquence des amplitudes plus significatives qui amènent vers une meilleur segmentation de la tarif en fonction de la zone.

Au niveau de la distribution à l'intérieur de chaque zone :

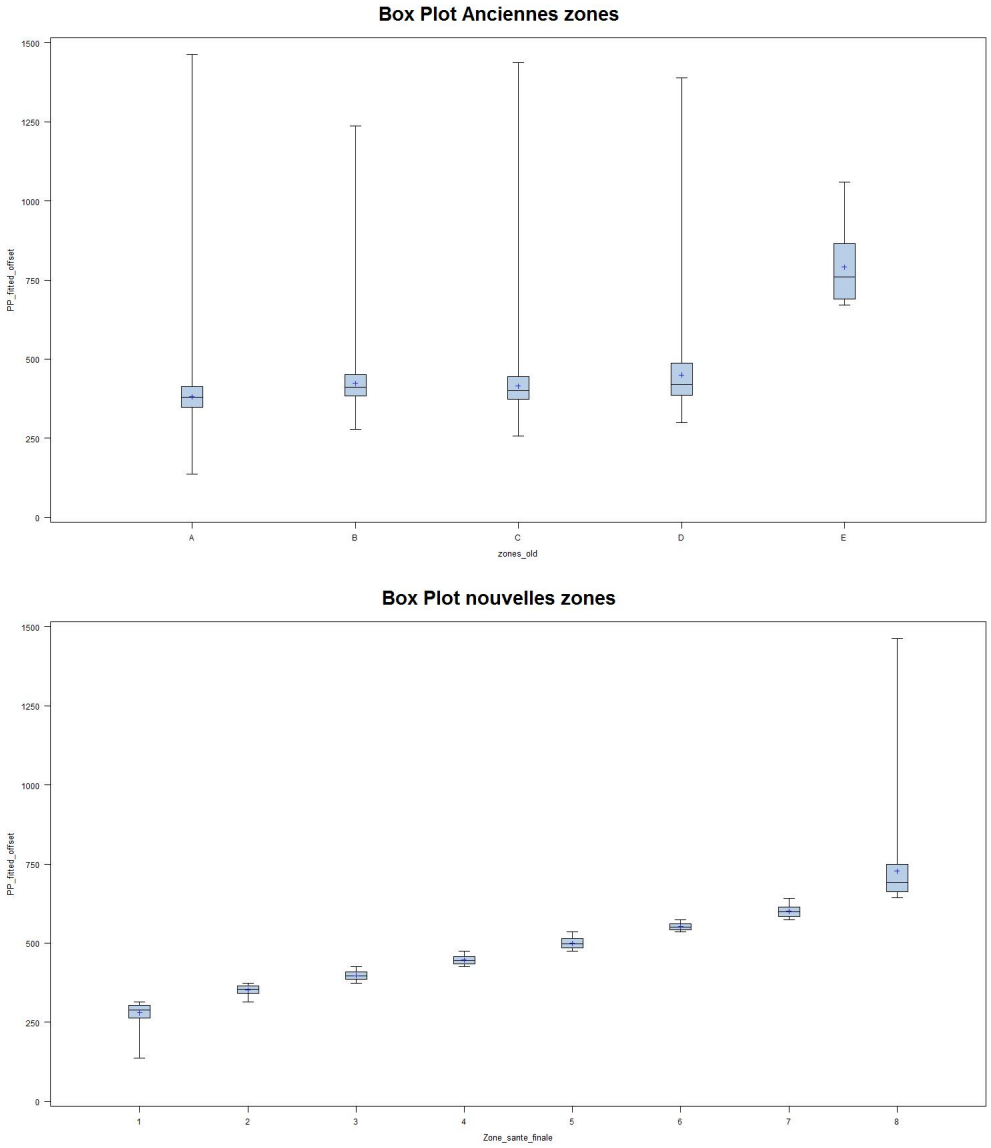


FIGURE 4.30 – Boite à moustache Actuel Zonier VS Boite à moustache Nouveau Zonier

On constate qu'à l'intérieur de chaque ancienne zone les variations sont plus importantes, ce qui entraîne des erreurs potentielles au niveau du classement. En revanche les nouvelles zones présentent une distribution plus homogène donnant ainsi une meilleure répartition et en conséquence une classification plus fiable.

Au niveau de la répartition dans la France :

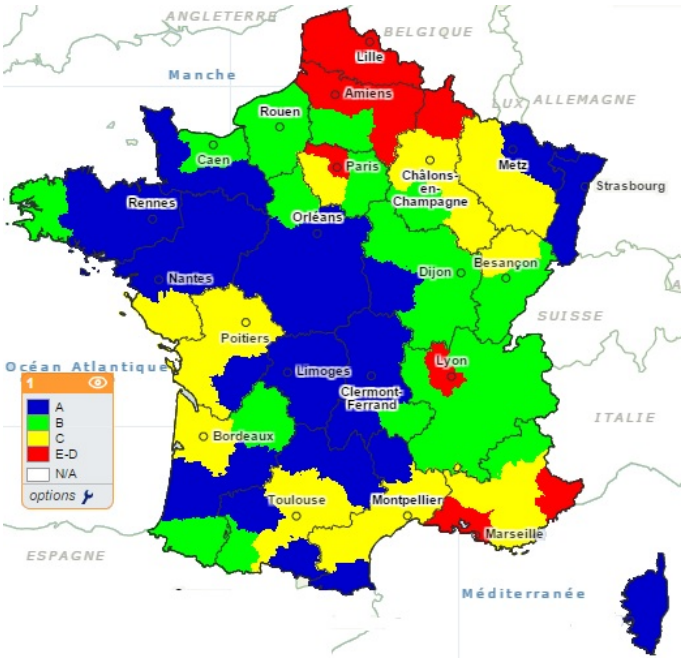


FIGURE 4.31 – Carte de l'Actuel Zonier

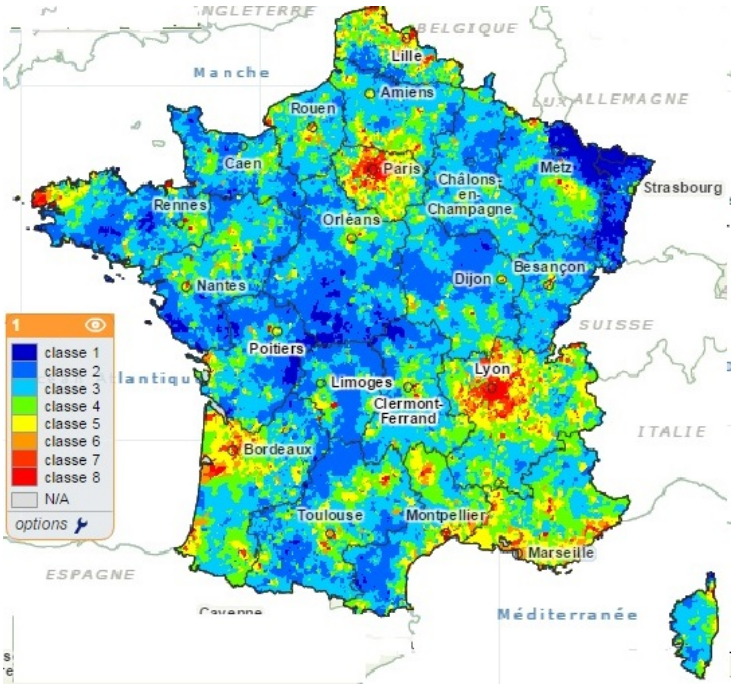


FIGURE 4.32 – Carte du nouveau Zonier

On peut constater grâce aux deux illustrations précédentes, que le nouveau zonier reste assez cohérent avec l'ancien. Néanmoins, grâce à une segmentation plus performante et plus approfondie, il nous livre une identification de groupes plus homogènes, à partir de segmentations plus performantes issues d'une méthode dont le niveau de précision supplémentaire est visible sur l'ensemble de la carte de France.

Chapitre 5

Conclusion

Dans ce mémoire, nous avons construit un nouveau zonier à l'aide des informations disponibles du portefeuille (variables tarifaires) et des informations supplémentaires (variables externes).

Dans un premier temps, nous avons modélisé la prime pure de chaque poste, en sélectionnant les variables les plus significatives pour expliquer au mieux le risque propre au poste. Le type de modélisation a été adapté à la nature des postes : modélisation directe de la prime pure par une loi Tweedie, ou bien modélisation de la fréquence multipliée par le coût moyen, lorsqu'on démontre que le lien entre ces deux quantités est suffisamment faible pour accepter l'hypothèse d'indépendance. Les postes Pharmacie, Dentaire et Optique ont révélé une liaison significative entre les coûts moyens et les fréquences associés à leur consommation respective, donc la prime pure a dû être estimée directement à partir d'un modèle Tweedie où nous avons ajusté une loi composée Poisson-Gamma. Une indépendance entre le nombre d'occurrences et la sévérité moyenne associés à la consommation Soins Courants a pu être raisonnablement acceptée, en conséquence l'estimation de la prime pure de ce poste à été obtenue à partir de la multiplication du modèle de fréquence par le modèle de coût moyen. Bien que les études réalisées sur le poste Hospitalisation aient donné comme résultat un très faible niveau de corrélation entre la fréquence et le coût moyen, l'indépendance n'a pour autant pas pu être démontrée de manière sûre, et les deux approches de modélisation ont donc dû être testées. La comparaison des résultats a permis de conclure que la modélisation Tweedie est la plus performante pour le poste Hospitalisation.

Pour la construction du zonier, il est nécessaire de bien distinguer le risque tarifaire du risque géographique, ce dernier étant en premier lieu représenté par les variables externes ajoutées au modèle de prime pure. Cet apport permet de résumer un effet géographique temporaire qui stabilise les estimations du modèle, en l'ajustant de manière à ce que les variables tarifaires soient le moins corrélées à un effet géographique. Après construction et réintroduction du zonier dans le modèle de prime pure, l'ensemble des coefficients estimés est ainsi moins perturbé par l'échange des variables externes par le zonier. La décorrélation de l'information tarifaire et de l'information géographique a également été renforcée par le forçage de l'estimation de L'Alsace-Moselle présente a priori dans la variable tarifaire Régime. Ce travail a ainsi permis la modélisation de la première partie de l'effet géographique synthétisé par les variables externes, en plus d'une estimation fiable de l'impact des variables tarifaires dans le modèle de prime pure.

Dans un second temps, l'analyse des résidus a mis en relief l'existence d'une structure

géographique perturbée par un effet aléatoire. La suite du travail a donc consisté à extraire cet effet géographique résiduel en lissant les résidus du modèle de prime pure. Deux méthodes de lissage ont pu être testées, la première nommée « Adjacency »qui lisse le risque d'une commune par simulations successives liées aux communes directement voisines. La seconde « Distance »qui lisse de manière déterministe avec une moyenne pondérée par les risques des communes voisines, où l'apport des communes dans la moyenne est contrôlé par un paramètre de distance. Les lissages optimaux ont été déterminés pour chaque méthode, ce qui a ensuite permis de les comparer. La méthode Adjacency a fourni un effet géographique résiduel plus prédictif, plus proche de l'effet réel moyen recherché. Enfin, cet effet a été ajouté à l'effet géographique externe pour constituer l'effet géographique total dont la répartition en classes par la méthode de Ward a abouti au zonier final. Par une iso-méthode de classement, la répartition des zones ainsi que la distribution de l'effet géographique à l'intérieur de ces zones (paliers d'effet moyen plus réguliers et variance inter-zone plus réduite) ont démontré que la méthode Adjacency est préfèrable à la méthode Distance. Dans l'optique de la détermination de la méthode optimale, l'hypothèse d'utilisation des variables externes a également été vérifiée, un zonier hors variables externes ayant été testé et s'étant révélé moins performant.

Enfin, on a ajouté le zonier au modèle de prime pure en remplacement des variables externes, le zonier devenant le facteur explicatif unique du risque géographique intégré dans la structure tarifaire. Ce travail est présenté en détail pour le poste Hospitalisation, mais en parallèle il a été réalisé pour tous les autre postes, permettant la construction d'un zonier représentant le risque géographique des postes Hospitalisation, Soins Courants, Pharmacie, Dentaire et Optique. La combinaison de ces zoniers a finalement permis d'obtenir un zonier global synthétisant la totalité de la consommation santé. On a vérifié que ce zonier global est bien représentatif du risque géographique à l'intérieur de chaque poste, en comparant son impact à celui du zonier propre au poste. Toutefois, le besoin d'une structure tarifaire simple demandée dans l'entreprise limite en pratique l'utilisation effective d'un zonier par poste, et mène plutôt à une estimation directe de la consommation moyenne tous postes confondus.

Grâce à son niveau de détail communal et à la méthode Adjacency appliquée sur une sinistralité plus récente, le nouveau zonier se révèle plus performant que le zonier actuel. Les risques sont ciblés de manière plus précise, avec une segmentation du risque géographique réel plus graduelle entre les zones et plus homogène à l'intérieur de ses zones. En insérant ce nouveau zonier dans le modèle de prime pure, cette meilleure estimation du risque géographique entraîne en définitive une meilleure performance du modèle de risque de santé global.

Annexe A

Complément de l'étude

A.1 Analyse descriptive par poste

Analyse de l'observé :

Pour mieux comprendre les comportements des différents postes, nous allons définir les coefficients suivants :

La **Prime Pure** représente la consommation moyenne annuelle par assuré,

$$PP = \frac{Sommes\ des\ charges}{Sommes\ des\ expositions}.$$

Le **Coût Moyen** représente l'intensité d'un sinistre, donnant un indicateur unitaire comparable,

$$CM = \frac{Sommes\ des\ charges}{Nombre\ de\ sinistres}.$$

La **Fréquence** représente le nombre de fois que se reproduit un sinistre dans un période de temps donné,

$$Freq = \frac{Nombre\ de\ sinistres}{Sommes\ des\ expositions}.$$

Fréquence par poste et formule					
	Hospitalisation	Soins courants	Pharmacie	Dentaire	Optique
F1	0,4	5,9	5,7	0,3	0,1
F2	0,5	7,4	6,4	0,4	0,2
F3	0,5	6,9	6,5	0,5	0,2
F4	0,5	7,5	6,9	0,4	0,2
F5	0,5	7,7	6,9	0,4	0,3
F6	0,4	7,0	6,5	0,5	0,4

FIGURE A.1 – Répartition de la Fréquence observée par poste et formule

Coût Moyen par poste et formule					
	Hospitalisation	Soins courants	Pharmacie	Dentaire	Optique
F1	137,1	12,5	13,0	26,2	26,2
F2	196,7	12,9	13,1	42,5	42,5
F3	312,7	16,8	12,3	65,3	65,3
F4	269,2	18,7	12,2	85,8	85,8
F5	379,9	23,2	12,4	128,1	128,1
F6	309,2	28,2	12,3	172,8	172,8

FIGURE A.2 – Répartition du Coût Moyen observée par poste et formule

Prime Pure par poste et formule					
	Hospitalisation	Soins courants	Pharmacie	Dentaire	Optique
F1	59,2	74,1	74,2	8,4	12,4
F2	100,3	95,5	83,9	17,5	27,3
F3	145,8	115,9	79,1	31,2	32,8
F4	125,6	141,0	83,7	31,0	39,3
F5	184,6	178,1	85,4	48,4	64,5
F6	127,3	196,7	79,9	79,0	112,8

FIGURE A.3 – Répartition de la Prime Pure observée par poste et formule

A.2 Étude croisée de la couverture complémentaire par de niveau garantie et par poste

Pour analyser les niveaux de couverture des garanties, notons :

- **C S.S.** : Couverture Sécurité Social sur les montants de frais réels. On a,

$$\textbf{C S.S.} = \frac{\sum \textit{Remboursement Sécurité social}}{\sum \textit{Montants frais réels}}.$$

- **CG** : Couverture complémentaire Santé Generali sur les montants de frais réels. On a,

$$\textbf{CG} = \frac{\sum \textit{Prestations}}{\sum \textit{Montants frais réels}}.$$

- **RC** : Proportion de reste à charge à l'assuré. On a,

$$\textbf{RC} = 1 - \textbf{CG} - \textbf{C S.S.}$$

Afin de bien identifier les impacts par poste pour les différents niveaux de garantie, les tableaux suivants, nous indiquent les résultats empiriques en pourcentage obtenus sur chaque poste.

Hospitalisation			
	C S.S.	CG	RC
Entrée de gamme	59,2%	37,2%	3,6%
Milieu de gamme	49,4%	43,4%	7,2%
Haut de gamme	45%	46,2%	8,8%

FIGURE A.4 – Répartition des dépenses du poste Hospitalisation

Soins Courants			
	C S.S.	CG	% RC
Entrée de gamme	62,2%	31,8%	6,1%
Milieu de gamme	58,1%	35,8%	6,1%
Haut de gamme	53,4%	40,0%	6,6%

FIGURE A.5 – Répartition des dépenses du poste Soins Courants

Pharmacie			
	C S.S.	CG	RC
Entrée de gamme	57,4%	42,3%	0,3%
Milieu de gamme	57,4%	42,0%	0,6%
Haut de gamme	57,4%	41,8%	0.8%

FIGURE A.6 – Répartition des dépenses du poste Pharmacie

Dentaire			
	C S.S.	CG	RC
Entrée de gamme	34,8%	23,2%	42%
Milieu de gamme	28,4%	34,6%	37%
Haut de gamme	22,7%	42,6%	34.7%

FIGURE A.7 – Répartition des dépenses du poste Dentaire

Optique			
	C S.S.	CG	RC
Entrée de gamme	3,7%	45,2%	51%
Milieu de gamme	3,1%	58,5%	38,4%
Haut de gamme	2,5%	63,9%	33,6%

FIGURE A.8 – Répartition des dépenses du poste Optique

Où,

— « Entrée de Gamme »correspond aux formules F1 et F2.

— « Milieu de Gamme »correspond aux formules F3 et F4.

— « Haut de Gamme »correspond aux formules F5 et F6.

Les tableaux nous indiquent que pour les mêmes niveau de couverture, les garanties proposées ne sont pas forcement homogènes.

A.3 Tableau des garanties

Garanties[1]	Formule 1	Formule 2	Formule 3	Formule 4	Formule 5	Formule 6 réservée aux TNS multiéquipés
Hospitalisation (y compris maternité)						
Frais de séjour en secteur conventionné	100 %	200 %	FR[2]	FR	FR	FR
Frais de séjour en secteur non conventionné	100 %	100 %	150 %	200 %	300 %	300 %
Chirurgie en secteur conventionné (hors chirurgie dentaire)	100 %	150 %	FR	FR	FR	FR
Chirurgie en secteur non conventionné (hors chirurgie dentaire)	100 %	150 %	200 %	300 %	400 %	400 %
Transport du malade	100 %	150 %	200 %	300 %	FR	FR
Forfait hospitalier[3]	FR	FR	FR	FR	FR	FR
Maison de repos et de convalescence (suite à hospitalisation)/an[4]	100 % limité à 30 J	100 % limité à 30 J	100 % limité à 30 J	100 % limité à 60 J	100 % limité à 60 J	100 % limité à 60 J
Hospitalisation à domicile, limitée à 90 jours/bénéficiaire/an	100 %	125 %	150 %	200 %	300 %	300 %
Garanties « Hospitalisation + » (y compris maternité)						
Chambre particulière en secteur conventionné[5]	-	50 €/jour	80 €/jour			100 €/jour
Bonus Fidélité : à partir de la 3ème année - forfait porté à :	-	60 €/jour	90 €/jour			120 €/jour
Lit d'accompagnant enfant de moins de 16 ans	-	25 €/jour limité à 30 j	50 €/jour limité à 30 j			50 €/jour limité à 30 j
Chambre particulière en secteur non conventionné	-	-	-	50 €/jour limité à 30 j		50 €/jour limité à 30 j
Forfait téléphone et télévision	-	-	-	4 €/jour limité à 30 j		4 €/jour limité à 30 j
Soins courants						
Consultations et visites de généralistes (par acte)	100 %	100 % + 5 €	100 % + 10 €	100 % + 20 €	100 % + 35 €	100 % + 35 €
Consultations et visites de spécialistes (par acte)	100 %	100 % + 10 €	100 % + 15 €	100 % + 25 €	100 % + 40 €	100 % + 50 €
Analyses, auxiliaires médicaux, imagerie, radiologie, échographie	100 %	125 %	150 %	200 %	250 %	300 %
Actes de spécialité : orthopédie, orthophonie, orthoptie	100 %	125 %	150 %	200 %	250 %	300 %
Appareillage, prothèses auditives	100 %	125 %	150 %	200 %	250 %	300 %
Pharmacie						
Médicaments et homéopathie pris en charge par le RO	TM[6]	FR	FR	FR	FR	FR
Optique						
Verres et montures pris en charge par le RO, lentilles prises en charge ou non par le RO/bénéficiaire/an	60 €	100 €	150 €	200 €	300 €	400 €
Bonus Fidélité : à partir de la 3ème année - forfait annuel porté à :	75 €	125 €	175 €	250 €	400 €	500 €
à partir de la 5ème année - forfait annuel porté à :	100 €	150 €	200 €	300 €	500 €	500 €
Chirurgie réfractive/par œil/bénéficiaire/an	100 €	150 €	200 €	300 €	500 €	600 €
Dentaire						
Soins (y compris chirurgie dentaire)	100 %	100 % + 10 €	100 % + 20 €	100 % + 40 €	100 % + 60 €	100 % + 60 €
Prothèses dentaires, implantologie (implant, support de prothèse, prothèse sur implant), orthodontie et inlay-core pris en charge par le RO/bénéficiaire/acte	60 €	120 €	150 €	200 €	300 €	350 €
Bonus Fidélité : à partir de la 3ème année - forfait porté à :	60 €	120 €	175 €	225 €	325 €	400 €
à partir de la 5ème année - forfait porté à :	60 €	120 €	200 €	250 €	350 €	400 €
Prothèses dentaires, implantologie (implant, support de prothèse, prothèse sur implant) non prises en charge par le RO/bénéficiaire/an	-	50 €	100 €	150 €	200 €	200 €
Inlays/onlays pris en charge par le RO/bénéficiaire/acte	50 €	75 €	100 €	150 €	200 €	200 €
Parodontologie prise en charge par le RO/bénéficiaire/an	50 €	75 €	100 €	150 €	200 €	200 €
Plafond de remboursement dentaire (hors soins)/bénéficiaire/an	400 €	600 €	800 €	1 100 €	1 500 €	1 500 €
Bonus Fidélité : à partir de la 3ème année - plafond annuel porté à :	500 €	700 €	950 €	1 300 €	2 500 €	2 500 €
à partir de la 5ème année - plafond annuel porté à :	600 €	800 €	1 100 €	1 500 €	3 000 €	3 000 €

FIGURE A.9 – Tableau des garanties

A.4 Ajustement des données à une lois théorique

Modèle de Fréquence

Comparaison des indicateurs des moments : Espérance et variance sur les données de fréquence pour les postes Hospitalisation et Soins courants.

— Si $\mathbb{E}[Y] \approx \mathbb{V}(Y)$: la loi de Poisson est adaptée.

— Si $\mathbb{E}[Y] < \mathbb{V}(Y)$: la loi Binomiale Négative peut aussi être adaptée

— Si $\mathbb{E}[Y] > \mathbb{V}(Y)$: la loi Binomiale peut aussi être adaptée.

	Hospitalisation	Soins Courants
Moyenne	0,414	6,336
Variance	1,495	6,926

TABLE A.1 – Moyenne et Variance pour les postes Hospitalisation et Soins Courants

La moyenne et la variance ici sont très proches pour le poste Soins Courants ; ce résultat nous indique qu'un ajustement par la loi de Poisson semble cohérent. Néanmoins, ils sont significativement éloignés pour le poste Hospitalisation, l'ajustement par une lois Poisson peut amener des erreurs plus importants dans la modélisation .

Modèle de Coût Moyen

Illustration pour le poste Hospitalisation :

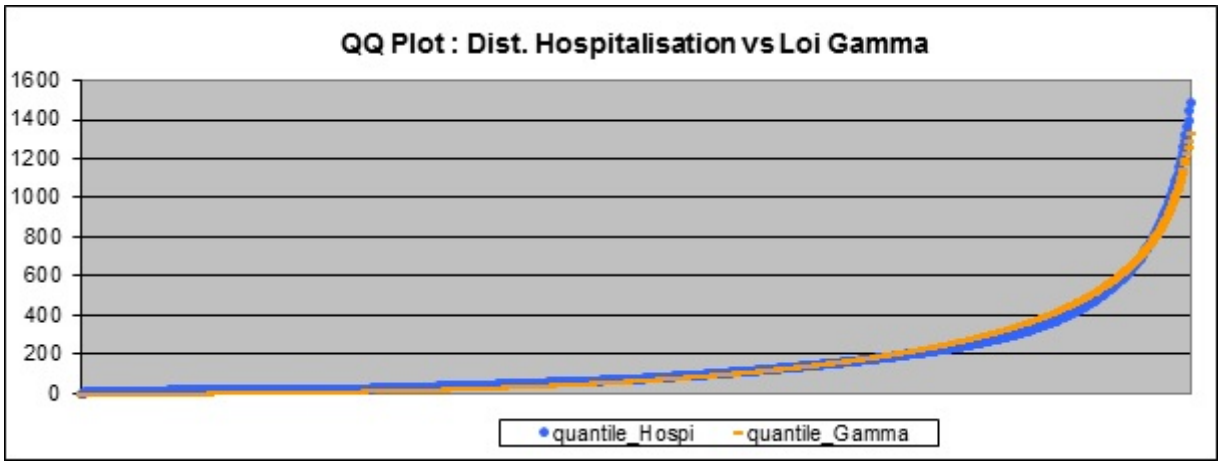


FIGURE A.10 – QQplot Distribution empirique VS Distribution d'une loi Gamma

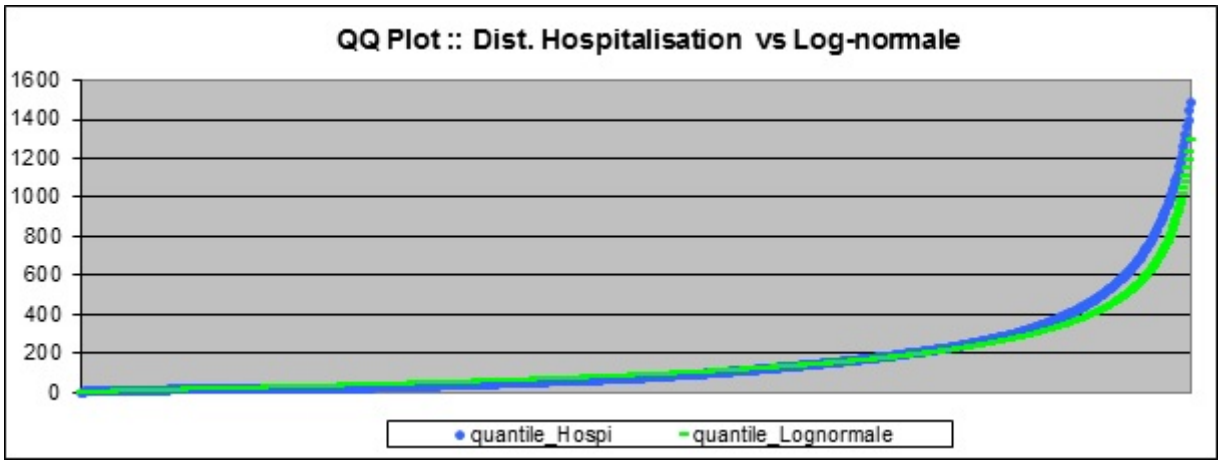


FIGURE A.11 – QQplot Distribution empirique VS Distribution d'une loi LogNormal

Modèle de la Consommation moyenne

Le modèle Tweedie appartient à la classe des modèles de dispersion exponentielle, célèbres pour leur rôle dans les modèles linéaires généralisés. C'est une famille de distributions de probabilité qui comprend des distributions continues telles que la distribution Normale et Gamma, la distribution de Poisson exclusivement discrète, et la classe des distributions composées mixtes Poisson-Gamma qui ont une quantité importante de zéros. C'est sur cette dernière distribution qu'on s'appuie pour la modélisation directe de la prime pure.

Les modèles Tweedie peuvent être vus comme des modèles Poisson composés. Nous supposons que $Y = \sum_{k=0}^{N} Z_k$ où Z_k sont indépendates et identiquement distribuées. Nous pourrons supposer que ces variables suivent une loi Gamma $G(\alpha, \beta)$ indépendamment de N suivant une loi de Poisson $P(\lambda)$. Alors,

$$\begin{aligned} E(Y) =& E(N)E(Z_k) \\ =& \lambda \frac{\alpha}{\beta} = \mu, \end{aligned}$$

et,

$$\begin{aligned} V(Y) &= E(N)E(Z_k^2) + V(N)E(Z_k)^2 \\ &= \lambda \left[\frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} \right]. \end{aligned}$$

Supposons qu'il existe $p \in]1, 2[$ avec $\phi > 0$ tels que :

$$\alpha = \frac{2-p}{p-1}, \beta = \frac{1}{\psi(p-1)\mu^{p-1}}, \lambda = \frac{\mu^{2-p}}{\phi(2-p)}.$$

Alors on peut montrer que la loi de Y appartient à la famille exponentielle avec,

$$\begin{aligned} E(Y) &= \mu, \\ V(Y) &= \phi \mu^p, \end{aligned}$$

où ϕ est un paramètre de dispersion ; la fonction variance est alors $V(\mu) = \mu^p$.

$$\mu = \frac{\lambda\alpha}{\beta},$$
$$p = \frac{\alpha+2}{\alpha+1},$$
$$\phi = \frac{\lambda^{1-p}}{2-p}\left(\frac{\alpha}{\beta}\right)^{2-p}.$$

On constate l'impossibilité de dissocier les paramètres pour son estimation, et donc que la distribution est entièrement définie par leur paramètre p. Dans des nombreux logiciels (SAS, R et Emblem) il existe différentes procédures (PROC severity, tweedie.profile) qui permettent d'obtenir l'estimation la valeur du paramètre p, elles sont basées sur l'article **([18])**.

Dans le cadre de ce mémoire, on a utilisé le logiciel « emblem »qui teste une plage de valeurs des p compris entre $(1, 2)$ et retourne la valeur du p qui fait que l'algorithme converge, correspondant au meilleur estimateur.

Poste	« Best Estimate »du paramètre p
Hospitalisation	1.672084972159
Pharmacie	1.57919059047082
Dentaire	1.63500701165103
Optique	1.24353775866055

TABLE A.2 – Tableau des valeurs de p obtenus.

A.5 Zonier final par poste

Donnons finalement le Zonier associé à chaque poste.

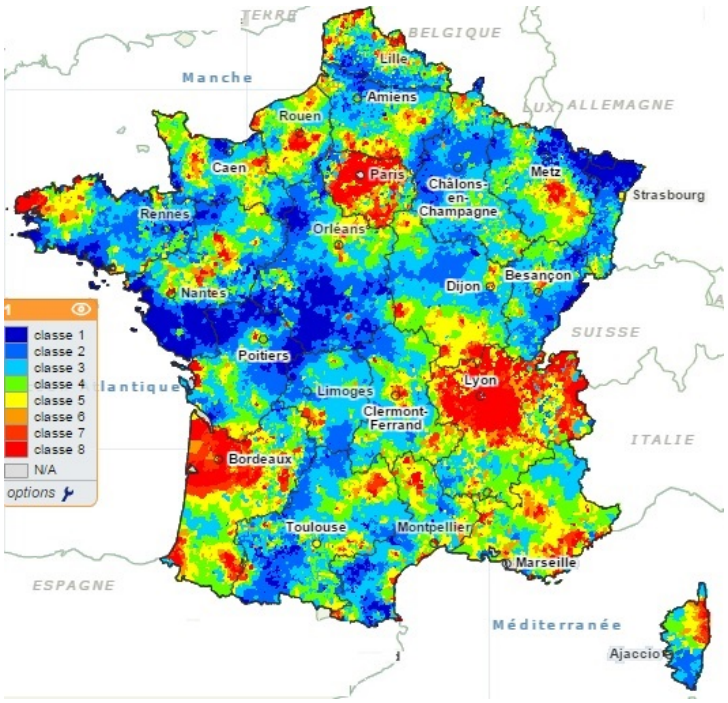


FIGURE A.12 – Zonier du poste Hospitalisation

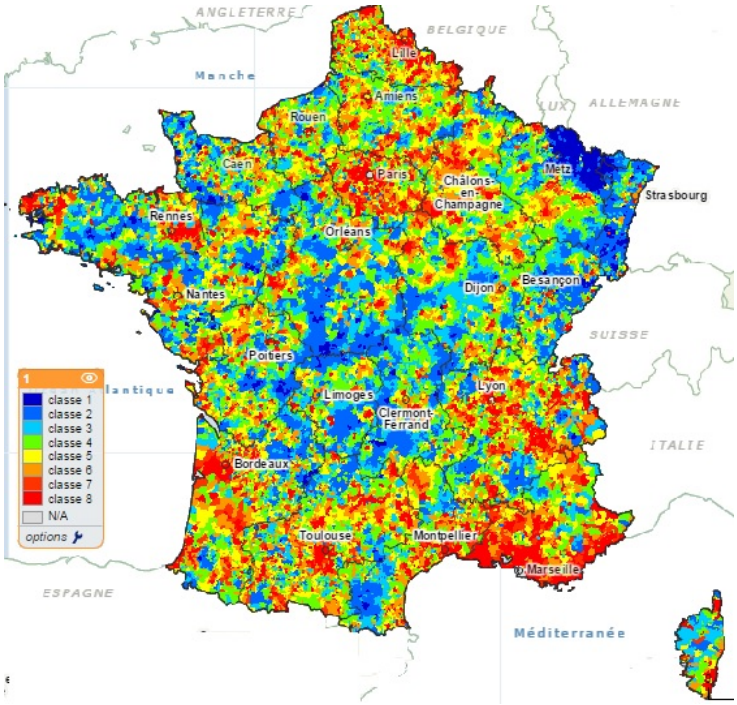


FIGURE A.13 – Zonier du poste Soins Courants

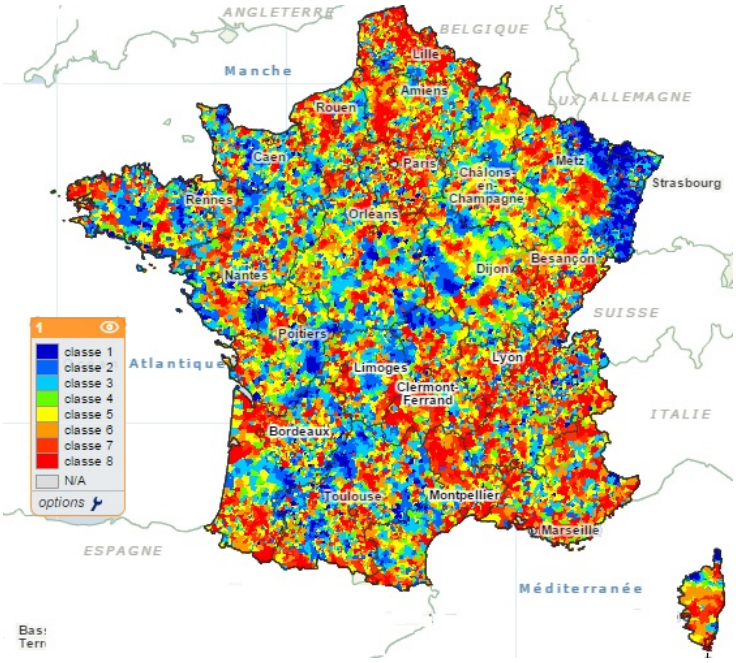


FIGURE A.14 – Zonier du poste Pharmacie

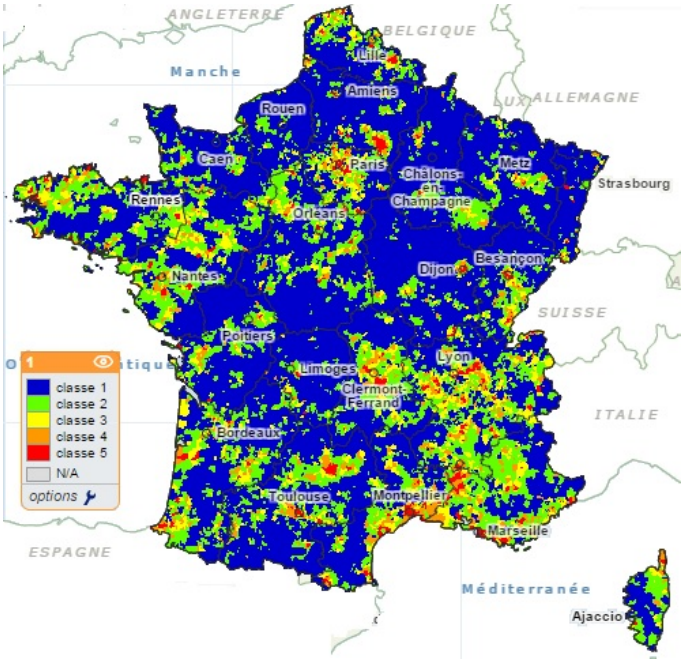


FIGURE A.15 – Zonier du poste Dentaire

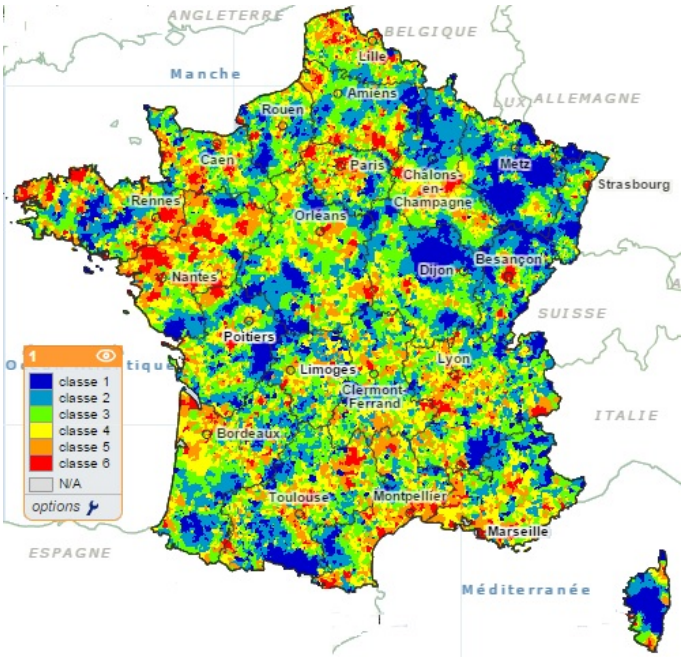


FIGURE A.16 – Zonier du poste Optique

A.6 Analyse descriptives des zones finales par département Méthode Adjacency vs Distance

Département	Min Zone Adjacency	Max Zone Adjacency	MinZone Dis	Max Zone Distance	Ecart Zones Adjacency	Ecart Zones Distance
1	2	12	2	14	10	12
2	1	9	1	12	8	11
3	1	8	1	12	7	11
4	2	9	1	12	7	11
5	2	8	1	12	6	11
6	2	10	1	12	8	11
7	2	9	1	10	7	9
8	1	8	1	12	7	11
9	1	6	1	10	5	9
10	1	9	1	9	8	8
11	1	10	1	11	9	10
12	2	10	1	14	8	13
13	2	10	1	9	8	8
14	1	9	1	13	8	12
15	2	7	1	8	5	7
16	1	9	1	12	8	11
17	1	11	1	15	10	14
18	1	10	1	12	9	11
19	1	10	1	12	9	11
21	1	10	1	14	9	13
22	1	9	1	13	8	12
23	2	7	1	11	5	10
24	1	10	1	13	9	12
25	1	10	1	14	9	13
26	2	9	1	11	7	10
27	1	12	1	15	11	14
28	1	7	1	11	6	10
29	1	12	1	13	11	12

FIGURE A.17 – Zones Méthode Adjacency vs Distance

Département	Min Zone Adjacency	Max Zone Adjacency	MinZone Dis	Max Zone Distance	Ecart Zones Adjacency	Ecart Zones Distance
30	1	10	1	13	9	12
31	1	9	1	12	8	11
32	1	9	1	13	8	12
33	1	12	1	14	11	13
34	1	10	1	14	9	13
35	1	8	1	11	7	10
36	1	7	1	12	6	11
37	1	9	1	14	8	13
38	2	12	1	15	10	14
39	1	8	1	12	7	11
40	2	11	1	11	9	10
41	1	5	1	11	4	10
42	3	11	2	12	8	10
43	2	8	2	9	6	7
44	1	10	1	15	9	14
45	2	11	1	13	9	12
46	1	5	1	9	4	8
47	2	13	1	15	11	14
48	2	9	2	11	7	9
49	1	10	1	13	9	12
50	1	12	1	12	11	11
51	1	9	1	14	8	13
52	2	8	2	11	6	9
53	1	10	1	15	9	14
54	1	11	1	14	10	13
55	1	9	1	12	8	11
56	1	8	1	13	7	12
57	1	7	1	11	6	10
58	2	10	1	12	8	11
59	1	12	1	12	11	11
60	2	10	1	13	8	12
61	1	9	1	10	8	9
62	1	10	1	14	9	13
63	1	12	1	15	11	14
64	2	11	1	13	9	12
65	1	10	1	13	9	12
66	1	9	1	11	8	10
67	1	6	1	10	5	9
68	1	7	1	10	6	9
69	5	15	4	15	10	11
70	1	9	1	11	8	10
71	2	10	2	13	8	11
72	1	9	1	13	8	12
73	2	9	1	12	7	11
74	2	12	2	14	10	12
75	9	13	4	11	4	7
76	2	12	1	14	10	13
77	2	12	1	14	10	13
78	1	13	1	15	12	14
79	1	7	1	11	6	10
80	1	8	1	10	7	9
81	1	10	1	13	9	12
82	2	8	1	14	6	13
83	2	10	1	12	8	11
84	2	9	1	11	7	10
85	1	4	1	8	3	7
86	1	10	1	12	9	11
87	1	8	1	9	7	8
88	1	7	1	11	6	10
89	1	9	1	12	8	11
90	1	6	1	9	5	8
91	3	12	1	14	9	13
92	7	14	2	14	7	12
93	9	13	3	12	4	9
94	8	13	3	12	5	9
95	1	12	2	14	11	12
2A	2	5	1	4	3	3
2B	2	11	1	12	9	11

FIGURE A.18 – Zones Méthode Adjacency vs Distance

Bibliographie

[1] Au service des professionnels de santé. `www.caducee.net`.

[2] Dress. `www.drees.sante.gouv.fr`.

[3] Insee. `www.insee.fr`.

[4] Irdes. `www.irdes.fr`.

[5] L'assurance maladie en ligne. `www.ameli.fr`.

[6] Le portail du service public de la sécurité sociale. `www.securite-sociale.fr`.

[7] Observatoire des territoires. `http://carto.observatoire-des-territoires.gouv.fr/`.

[8] Valérie Albouy, Bretin Emmanuel, Nicolas Carnot, and Muriel Deprez. Les dépenses de santé en france : déterminants et impact du vieillissement à l'horizon 2050, ministère de santé. *Documents de Travail de la DGTPE*, 2009.

[9] M. Boskov and R.J Verrall. Premium rating by geographic area using spatial models. *ASTIN Bulletin*, 24 :131–143, 1994.

[10] Comptes de la santé. Le ralentissement de la progression de la consommation de soins et biens médicaux se confirme. *Comptes nationaux de la santé*, 2013.

[11] DRESS. Recueil d'indicateurs régionaux ; offre de soins et état de santé. *Comptes nationaux de la santé*, 2014.

[12] Arul Earnest, Geoff Morgan, Kerrie Mengersen, Louise Ryan, Richard Summerhayes, and John Beard. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (car) models. *International Journal of Health Geographics*, 6, 2007.

[13] Myriam El Jerdy. Tarification des groupes en assurance santé. *Mémoire ISFA*, 2008.

[14] Patrik Emanuelsson. Construction of rating territories for water-damage claims. *Master Thesis in Mathematical Statistics*, 2011.

[15] Jose Garrido and Jun Zhou. Credibililty theory for generalized linear and mixed models. *Department of Mathematics and Statistics. Technical Report No. 5/06. Concordia University*, 2006.

[16] Charlotte Geay and Grégoire Lagasnerie. Projection des dépenses de santé à l'horizon 2060, le modèle promede. *Documents de Travail de la DGTPE*, 2009.

[17] Peter J. Green and Sylvia Richardson. Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.*, 97(460) :1055–1070, 2002.

[18] Bent Jørgensen and Marta C. Paes de Souza. Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.*, (1) :69–93, 1994.

[19] Julien Mathis. Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations mcmc. *Mémoire ULP*, 2009.

[20] J.A. Nelder and R.J Verrall. Credibility theory and generalized linear models. *ASTIN Bulletin*, 27 :71–82, 1997.

[21] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*, volume 104 of *EAA Series*. Springer-Verlag Berlin Heidelberg, 2010.

[22] G.M. Philip and D.F. Watson. A precise method for determining contoured surfaces. *Australian Petroleum Exploration Association Journal*, 22 :205–212, 1982.

[23] G.M. Philip and D.F. Watson. A refinement of inverse distance weighted interpolation. *Geoprocessing*, 2 :315–327, 1985.

[24] Fabien Poubennec. Refonte de la tarification d'un produit assurance santé modulaire avec implantation d'un lissage spatial. *Mémoire EURIA*, 2013.

[25] Xacur Quijano and Alberto Oscar. Property and casualty premiums based on tweedie families of generalized linear models. *Masters thesis, Concordia University*, 2011.

[26] Christian Robert. *Le choix bayésien : Principes et pratique*. Statistique et probabilités appliqués. Springer-Verlag Paris, 2006.

[27] STATISS. Statistiques et indicateurs de la santé et du social. *Comptes nationaux de la santé*, 2014.

[28] Frédéric Tallec. Mobilité géographique des patients en soins hospitaliers de court séjour : comment la répartition de l'offre structure-t-elle le territoire ? *Comptes nationaux de la santé*, 2010.

[29] Cm. Theobald, Firat, and R. Thompson. Gibbs sampling, adaptive rejection sampling and robustness to prior specification for a mixed linear model. *Genet Sel Evol - Genetics Selection Evolution*, 29(1) :57–72, 1997.

[30] Pierre Thérond. Théorie de la credibililté. *Cours ISFA*, 2013.

[31] Ildikó Vitéz. Location as risk factor spatial effect in insurance. In *Department of Probability Theory and Statistics Eötvös Loránd University*. 2007.

[32] Olivier Wintenberger. Théorie de la crédibilité. *Cours ISUP, UPMC*, 2013-2014.

[33] Jun Zhou. Theory and applications of generalized linear models in insurance. *PhD thesis, Concordia University*, 2011.