Institut des ACTUAIRES

ISUP

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

Par : Madame / ~~Monsieur~~ Fatouma Yéli SANASSA

Titre du mémoire : Mise en place d'un zonier dans la mesure de performance d'un régime santé

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

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

Membres présents du jury de la filière : Signature :

Entreprise : CREDIT AGRICOLE ASSURANCES

Nom : SALAH ADEL

Signature :

Directeur de mémoire en entreprise

Membres présents du jury de l'Institut des Actuaires : Signature :

Nom : SALAH ADEL

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 :

Master Actuariat
Mémoire actuariat

Mémoire actuariat : Mise en place d'un zonier dans la mesure de performance d'un régime santé

Étudiante :
Fatouma Yéli SAMASSA

Tuteur:
Adel SALAH

Résumé

Ce mémoire présente une stratégie de conseil visant pour se démarquer de la concurrence et à développer de l'activité de la complémentaire santé collective au sein du Crédit Agricole Assurance. Cette approche consiste à évaluer l'efficacité d'une couverture santé en analysant le reste à charge moyen résultant des prestations, en tenant compte la situation géographique par un zonier. Le zonier est défini à l'aide d'un modèle linéaire généralisé, comparé avec deux procédés de machine learning basés sur des arbres de régression : la forêt aléatoire et le gradient boosting.

Chacune de ces méthodes est utilisée pour estimer le cout moyen à partir des paramètres tarifaires conventionnels du portefeuille et des variables externes visant à partiellement expliquer l'impact géographique. La méthode offrant la meilleure estimation minimise le bruit résiduel et ainsi isole le mieux l'effet géographique non expliqué par les variables du modèle.

L'effet géographique total est obtenu en combinant les résidus avec les effets géographiques des variables externes. Une classification hiérarchique ascendante est utilisée pour regrouper les effets géographiques en un découpage zonal. Cependant, l'utilisation de données départementales peut induire une perte de fiabilité dans certains départements en raison de faibles effectifs.

Pour pallier ce problème, le coût moyen des prestations dans les départements moins fiables est estimé en utilisant une méthode de lissage spatial basée sur la théorie de crédibilité. Cette approche pondère les coûts moyens des départements voisins en fonction de leurs effectifs respectifs.

Le zonier obtenu a été intégré dans un outil d'aide au conseil présenté dans ce mémoire avec un cas pratique pour illustrer une de ses utilisations.

Mots clés : Assurance santé collective, notation, grille de garanties, variables externes, modèles linéaires généralisés, forêts aléatoires, bagging, gradient boosting, classification hiérarchique ascendante, zonier.

Abstract

This paper presents a consulting strategy aimed at distinguishing itself from competition and fostering the development of collective supplementary health insurance within Crédit Agricole Assurance. This approach involves assessing the effectiveness of health coverage by analyzing the average out-of-pocket expenses resulting from benefits, taking into account the geographical situation through a zoning system. The zoning is defined using a generalized linear model, compared with two machine learning techniques based on regression trees : random forest and gradient boosting.

Each of these methods is used to estimate the average cost based on conventional tariff parameters of the portfolio and external variables aimed at partially explaining the geographical impact. The method offering the best estimate minimizes residual noise and thus isolates the geographical effect not explained by the model's variables.

The total geographical effect is obtained by combining the residuals with the geographical effects of external variables. An ascending hierarchical classification is used to group the geographical effects into a zoning structure. However, the use of departmental data may lead to a loss of reliability in certain departments due to low sample sizes.

To address this issue, the average cost of benefits in less reliable departments is estimated using a spatial smoothing method based on credibility theory. This approach weighs the average costs of neighboring departments based on their respective sample sizes.

The obtained zoning structure has been integrated into a consulting support tool presented in this paper, along with a practical case to illustrate one of its applications.

Keywords : collective health insurance, rating, benefit grid, external variables, generalized linear models, random forests, bagging, gradient boosting, ascending hierarchical clustering, zoning.

Remerciements

Je souhaiterais adresser mes sincères remerciements à Adel SALAH mon tuteur en entreprise qui m'a soutenue tout au long de ce mémoire. Je te remercie pour ton écoute, ta disponibilité, tes conseils avisés et surtout l'engagement auquel tu as su faire preuve. J'ai beaucoup appris et apprécié de travailler avec toi.

Je tiens à remercier l'ensemble des collaborateurs des équipes techniques des assurances collectives de chez Crédit Agricole, pour leur accueil tout au long de mon apprentissage. Durant mon expérience parmi eux, j'ai énormément appris et vécu une belle expérience humaine.

Je souhaite exprimer ma gratitude envers le corps enseignant de l'Institut de Statistique de l'Université de Paris (ISUP) pour la qualité de son enseignement, qui a rendu ces deux dernières années extrêmement enrichissantes.

Je tiens également à remercier toutes les personnes qui m'ont soutenu tout au long de ma scolarité.

Enfin, je remercie infiniment mes proches, pour leur soutien infaillible durant toutes ces années.

Synthèse

Le marché de la complémentaire santé en France est très concurrentiel, particulièrement le marché de la complémentaire santé collective, qui en plus d'être très compétitif, est également très réglementé et évolue constamment. Les employeurs ont l'obligation de proposer une complémentaire santé à leurs salariés, ce qui représente un marché important pour les compagnies d'assurance, les mutuelles et les institutions de prévoyance. Peu de nouveaux acteurs entrent sur le marché, tandis que d'autres fusionnent ou quittent le marché.

Dans un contexte tendu, les acteurs de ce marché doivent faire preuve d'innovation pour se démarquer de leurs concurrents. C'est pourquoi le Crédit Agricole profite de son statut de bancassureur pour privilégier l'approche-conseil pour se démarquer de la concurrence. Cette démarche passe par le développement et l'automatisation de la procédure accompagnant les clients dans le choix et le suivi de leurs couvertures santé, à travers un outil d'aide au conseil permettant d'auditer, mais également d'évaluer la performance d'un régime santé par les garanties, les cotisations ainsi que le profil démographique.

Dans l'audit d'un régime santé, le calcul de l'indicateur mesurant la performance des garanties est basée sur un zonier établit à partir des montants de frais de santé couvert en portefeuille.

L'objectif de ce mémoire est le développement de l'outil d'aide au conseil, principalement au travers :

— Une nouvelle métrique : la métrique actuelle attribue une note entre 0 et 5 pour un ensemble restreint d'actes de soins présentant des enjeux sur la performance d'une couverture. De plus, cette dernière comporte un aspect subjectif et perd du sens hors de son contexte. Elle est remplacée par le taux de couverture moyen calculé sur un ensemble plus large d'actes de soins. Cet indicateur est plus facile à interpréter pour le client.

— La refonte du zonier : le zonier actuel a été construit sur l'expérience, c'est-à-dire la connaissance du marché de la santé sur la maille régionale. Il est remplacé par un nouveau zonage construit par le procédé classique du modèle linéaire généralisé.

Le zonier de remplacement est construit sur la base des régions à partir d'une méthode classique donnée par la modèle linéaire généralisé du coût moyen des prestations, estimé par les variables tarifaires internes non géographiques que nous avons complétés avec des variables externes qui en revanche apportent de l'information géographique. Enfin, le zonier est donné par le regroupement par classification des résidus de l'estimation des coûts moyens de prestations, isolant l'effet géographique contenu dans ces derniers additionnés à l'information géographique des données.

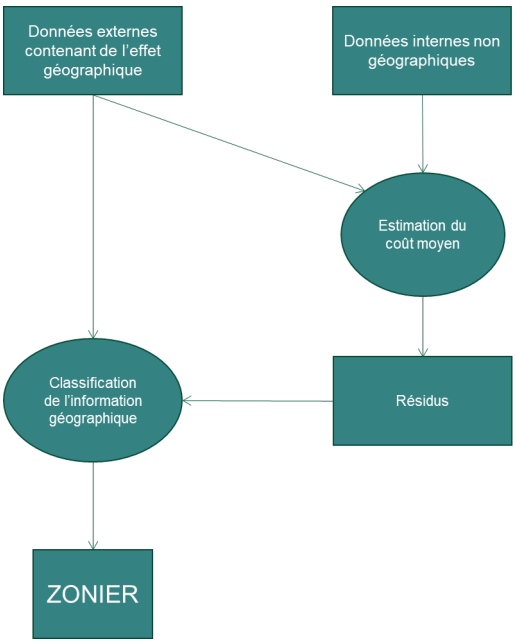


FIGURE 1 – Les étapes de l'élaboration d'un zonier

La base des régions est un choix donnant des variables fiables, néanmoins ces variables résument trop d'information. c'est pourquoi, nous avons repris le procédé de zonage avec la maille départementale permettant de récupérer plus de précisions. Cependant, la précision récupérée n'est pas robuste pour les départements avec peu d'assurés. Toutefois, ce manque de fiabilité causé par des effectifs faibles dans certains départements est corrigé en estimant les coûts moyens de prestations avec un lissage spatial basé sur la théorie de crédibilité.

De plus, les estimations coûts moyens de prestations sur la base des départements sont comparées avec celles données par deux méthodes de machine learning utilisant des arbres de décisions : la forêt aléatoire et le gradient boosting.

À l'application de chaque méthode, nous avons ajusté au mieux chaque algorithme pour obtenir une meilleure estimation du coût moyen de prestations, et ainsi mieux isoler l'effet géographique contenu dans ce dernier par les résidus.

	MSE
GLM	11.97178
GLM optimisé	9.855609
RandomForest	17.4192
RandomForest optimisé	13.97998
GBM	19.23063
GBM optimisé	9.25108

FIGURE 2 – Les performances de chaque méthode

Nous pouvons remarquer que les ajustements de modèle de machine learning diminuent significativement les erreurs de prédictions. Il faut noter que le gradient boosting ajusté (GBM optimisé) propose une qualité de prédiction au-dessus des deux autres modèles. Néanmoins, la paramétrisation du modèle est assez lourde, une grande quantité de calculs est effectuée ce qui gêne fortement la traçabilité et la vérification du résultat.

Pour construire le zonier, nous avons retenu les résidus donnés par le gradient boosting. En effet, en termes d'erreur de prédiction ce dernier donne l'erreur quadratique la plus petite, nous considérons donc qu'il isole mieux l'effet géographique contenu dans le coût moyen de prestation.

À ces résidus, nous avons ajouté l'information géographique des variables externes pour la classification des départements. Ainsi, nous avons obtenu de zonier regroupant les départements en quatre zones figurant sur la figure ci-dessous. Ce nouveau regroupement présente plusieurs similitudes avec le zonier actuel.

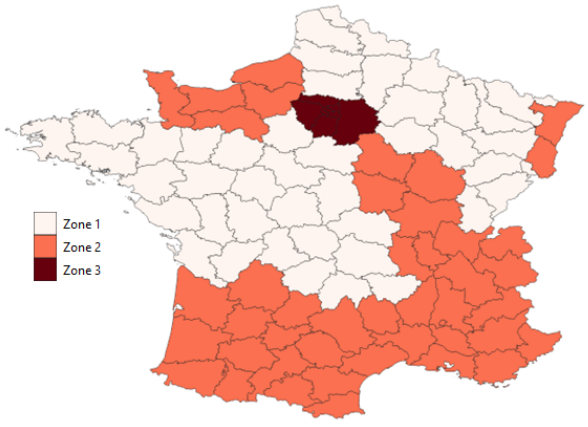


FIGURE 3 – Zonier actuel

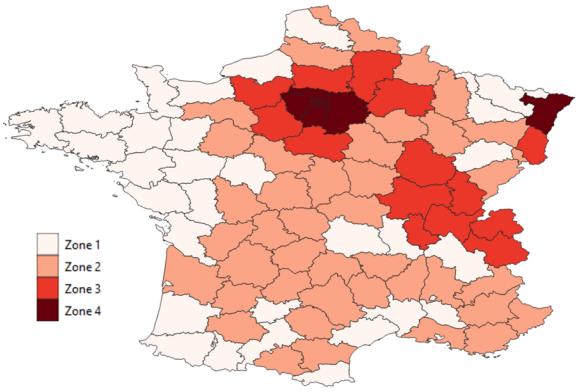


FIGURE 4 – Nouveau zonier

Synthesis

The French supplementary health market is highly competitive, particularly the market for collective supplementary health, which is not only very competitive but also highly regulated and constantly evolving. Employers are obligated to offer supplementary health coverage to their employees, which represents a significant market for insurance companies, mutuals, and welfare institutions. Few new players enter the market, while others merge or exit.

In a tense context, the actors in this market must show innovation to stand out from their competitors. That is why Crédit Agricole takes advantage of its status as a "bancassureur" and favors a consultancy approach to differentiate itself from the competition. This approach involves the development and industrialization of the procedure accompanying clients in the selection and monitoring of their coverage, through a tool for consulting that can audit and evaluate the performance of coverage by guarantees, contributions, and demographic profile.

In the audit of a health plan, the calculation of the indicator measuring the performance of guarantees is based on a zoning established from the amounts of healthcare expenses covered in the portfolio.

The objective of this dissertation is the development of the tool for consulting, mainly through :

— A new metric : The current metric assigns a score between 0 and 5 for a limited set of medical procedures that have an impact on the performance of a coverage. Moreover, it has a subjective aspect and loses its meaning outside of its context. It is replaced by the average coverage rate calculated on a broader set of medical procedures. This indicator is easier to interpret, and its meaning is clearer for the client.

— The revamping of the zoning system : The current zoning system was built on experience, i.e., knowledge of the health market at the regional level. It is replaced by a zoning system at the regional level established by the classic process of developing a zoning system : the GLM.

The replacement zoning is built based on regions using a classical method given by the generalized linear model of the average cost of services, estimated by internal pricing variables that do not contain geographic information, which we supplemented with external variables that provide geographic information. Finally, the zoning is obtained by grouping through classification the residuals from the estimation of the average service costs, isolating the geographic effect contained therein and adding the geographic information from the data.

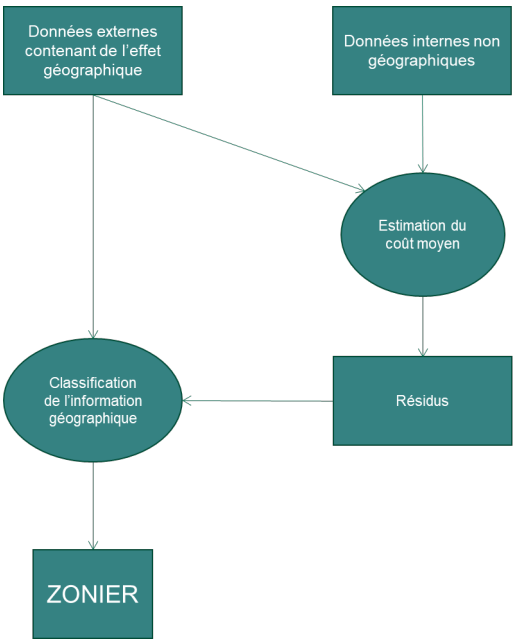


FIGURE 5 – The steps for developing a zoning plan

The region-based approach is a choice that provides reliable variables, but these variables summarize too much information. Therefore, we used the zoning process with the departmental level to obtain more precision. However, the recovered precision is not robust for departments with few insured individuals. Nevertheless, this lack of reliability caused by small sample sizes in some departments is corrected by estimating the average service costs using spatial smoothing based on credibility theory.

In addition, the average service cost estimates based on departments are compared with those given by two machine learning methods using decision trees : random forest and gradient boosting.

For each method, we adjusted the algorithm to obtain a better estimation of the average cost of services and thus better isolate the geographic effect contained in the residuals.

	MSE
GLM	11.97178
GLM optimisé	9.855609
RandomForest	17.4192
RandomForest optimisé	13.97998
GBM	19.23063
GBM optimisé	9.25108

FIGURE 6 – The performance of each method

We can observe that the adjustments made to the machine learning model significantly reduce prediction errors. It should be noted that the optimized gradient boosting model (GBM) offers higher prediction quality than the other two models. However, the model parameterization is quite complex, requiring a large amount of computation, which greatly hinders traceability and result verification.

To construct the zoning, we selected the residuals given by the gradient. Indeed, in terms of prediction error, the gradient produces the smallest quadratic error, which allows us to better isolate the geographic effect contained in the average cost of services.

To these residuals, we added the geographic information from external variables for the classification of departments. Thus, we obtained a zoning that groups departments into four zones, as shown in the figure below. This new grouping presents several similarities with the current zoning.

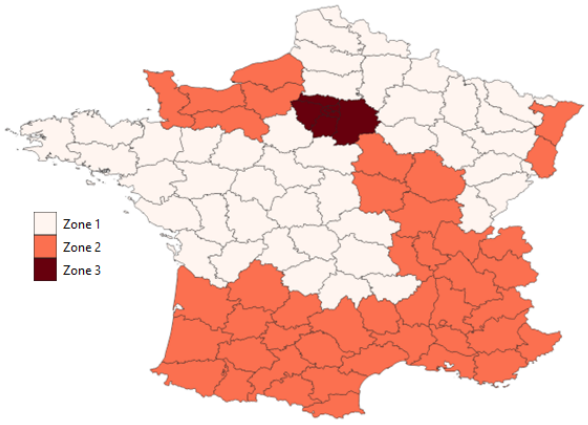


FIGURE 7 – Zonier actuel

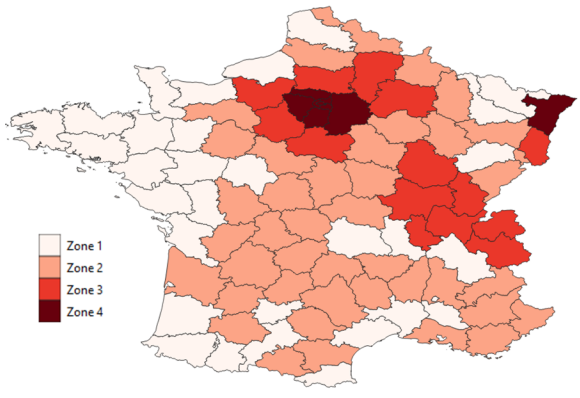


FIGURE 8 – Nouveau zonier

Introduction

Depuis plusieurs décennies, les dépenses en santé ne cessent d'augmenter, cette pente croissante s'est établie face aux progrès médicaux qui demandent plus de moyens et également face au vieillissement de la population qui consomme davantage de soins médicaux. Pratiquement tous les ans, une dérive de plus ou moins 2% est observée sur les prestations de soins par les assureurs. Malgré les prix limites de vente fixés par l'État sur certains soins et biens médicaux, cette tendance va encore poursuivre.

C'est pourquoi, il devient primordial de souscrire à une complémentaire santé individuelle ou collective afin de limiter ou d'annuler dans le meilleur des cas le restant à charge. En particulier, depuis 2016, la loi l'Accord National Interprofessionnel (ANI) n'impose plus de proposer, mais oblige les entreprises à couvrir leurs salariés pour des besoins santé par le biais des assurances collectives avec un financement d'au moins 50%.

La complémentaire santé est initialement une structure très concurrentielle et de nombreuses réformes et réglementations en faveur des bénéficiaires viennent intensifier cette concurrence et contraignent les acteurs de la complémentaire santé à tarifer avec une marge technique très faible, voire nulle.

De nos jours, en assurance collective, proposer le meilleur tarif ne suffit plus pour conserver sa clientèle et développer son portefeuille, c'est pourquoi Crédit Agricole, en tant que bancassureur, a développé une stratégie par le conseil afin de déployer son activité d'assurance à partir de son réseau client banque installé depuis de nombreuses décennies. Cette approche consiste à accompagner le client dans la recherche ou la construction d'une meilleure couverture santé en s'appuyant sur des benchmarks construits sur le portefeuille solide déjà établi tout en intégrant leurs besoins.

Dans ce mémoire, nous allons présenter l'amélioration et l'automatisation de cette approche de conseil et d'accompagnement par le biais d'un outil d'aide au conseil qui mesure et compare la performance de plusieurs régimes santé à partir des coûts réels de soins observés, de la position géographique et du secteur d'activité.

Dans un premier temps, nous présenterons le périmètre de l'assurance santé collective en France et son marché, ensuite, nous apporterons les éléments de construction de l'outil d'audit permettant d'établir les benchmarks et mesurer la performance des couvertures. Dans une troisième partie, nous aborderons les différents modèles utilisés pour intégrer l'aspect géographique dans l'outil d'aide au conseil à travers un zonier. Nous poursuivrons sur l'application et l'analyse des modèles présentés. Et, enfin, nous illustrerons avec un cas pratique une application concrète de l'outil.

Table des matières

1 Fonctionnement de la complémentaire santé collective et son marché

De nos jours, la complémentaire santé est un élément fondamental pour l'accès aux soins des ménages. Les offres sont très variées en termes de garanties et de tarifs, afin de répondre aux besoins des particuliers, des travailleurs indépendants et des entreprises. Le marché de la complémentaire santé est également soumis à une réglementation stricte pour garantir la qualité et l'accessibilité de l'offre.

1.1 L'assurance complémentaire santé en France

En France, une partie des prestations de soins est prise en charge par l'État par le biais de la Sécurité sociale et peut être complétée par un organisme assureur, ce qui réduit fortement la part du bénéficiaire.

1.1.1 La Sécurité sociale

La Sécurité sociale est un organisme social public de l'État français. Elle est destinée à assurer, pour tout individu résidant sur le territoire français, une couverture santé dite « de base » pour des risques sociaux tels que la santé, incapacité de travail ou invalidité, décès, chômage, et bien d'autres.

Cette protection s'exerce par l'affiliation des individus et de leurs ayants droit à l'un des différents régimes dépendant de l'activité professionnelle des assurés.

La Sécurité sociale regroupe trois régimes principaux :

- Le régime général couvre les salariés et les travailleurs assimilés à des salariés ainsi que leurs ayants droit, ce qui représente 80% de la population. Ainsi, il génère plus de la moitié des dépenses de la Sécurité sociale.
- Le régime agricole couvre les exploitants et salariés agricoles au sein de la Mutualité Sociale Agricole (MSA).
- Le régime social d'indépendants couvrant les artisans, les commerçants et les salariés du secteur industriel et professions libérales.

En plus de ces principaux régimes, il existe d'autres régimes spéciaux fonctionnant sous la base de solidarité restreinte à une profession ou à une entreprise : les régimes de la SNCF, de la RATP, des marins, des étudiants, etc. En particulier, le régime d'Alsace-Moselle qui est spécifiquement destiné aux salariés des entreprises implantées dans les départements de la Moselle, du Haut-Rhin et du Bas-Rhin. Les bénéficiaires de ces régimes spéciaux disposent d'une couverture assurance maladie obligatoire plus avantageuse, en particulier les montants des remboursements y sont plus importants.

Chaque régime est géré de manière indépendante et organisé en cinq branches d'activités qui couvrent plusieurs risques (la maladie, les accidents du travail et maladies professionnelles, la retraite, la famille et le recouvrement).

Les ressources financières des différents régimes de la Sécurité sociale se repartissent en trois catégories :

— les cotisations sociales payées par les employeurs et personnes salariées ;

— la Contribution Sociale Généralisée (CSG) prélevée sur l'ensemble des revenus ;

— les autres taxes et les impôts de toute nature comme la contribution sociale de solidarité des sociétés, la TVA brute sur le tabac et autre.

Au fil des années, les ressources financières se sont révélées insuffisantes, ce qui a incité la Sécurité sociale à prendre des mesures pour contrôler les remboursements de frais de santé, tels que le respect de l'Objectif National de Dépenses d'Assurance Maladie.

En outre, l'État se désengage progressivement en transférant certaines dépenses de soins, notamment celles liées à l'optique, aux organismes complémentaires (mutuelles, institutions de prévoyance, assureurs), à travers diverses réformes et réglementations.

1.1.2 Les différents organismes complémentaires collectifs

En France, selon les statistiques de la Mutualité française en 2021, environ 95% de la population est couverte par une assurance complémentaire santé, afin de compléter le régime de base qu'offre l'assurance maladie en souscrivant à une assurance individuelle ou collective, auprès de l'un des trois différents organismes de complémentaire santé.

Les contrats individuels sont souscrits par des particuliers, tandis que les contrats collectifs sont souscrits dans la plupart des cas par une collectivité, une personne morale ou un chef d'entreprise en vue de couvrir les salariés au près d'un organisme assureur.

L'objectif des organismes complémentaires est donc d'assurer une partie ou la totalité des dépenses de santé non prises en charge par l'Assurance Maladie. Ils sont regroupés en trois groupes régis par des législations différentes et ont des modes de fonctionnement qui leur sont propres :

— Les mutuelles, régies par le Code de la mutualité, sont des sociétés de personnes à but non lucratif appartenant à leurs assurés. Elles sont surtout actives sur le marché de l'assurance individuelle, qui représente environ 70% de leurs bénéficiaires et de leur chiffre d'affaires, comme nous pouvons le constater en figure 9.

— Les Sociétés d'assurance sont des organismes assureurs relevant du Code des assurances et à but lucratif.

— Les institutions de prévoyance, à but non lucratif, sont soumises au Code de la Sécurité sociale, qui leur permet d'exercer des activités uniquement sur le champ des risques sociaux. Elles sont spécialisées dans la couverture des salariés des entreprises ou des branches professionnelles et sont des organismes dits paritaires : leurs conseils d'administration comportent, à égalité, des représentants des salariés

et des employeurs des entreprises ou branches souscriptrices.

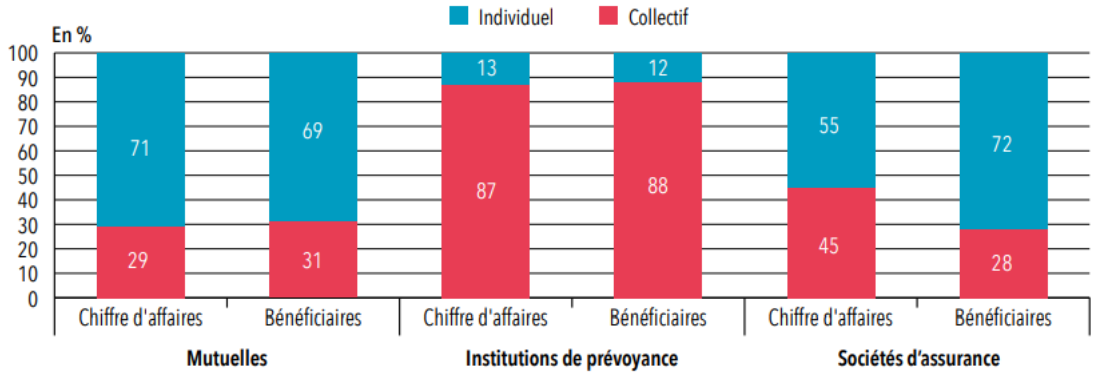


FIGURE 9 – L'activité des différents types d'organismes de complémentaire santé ; Source :Drees

Parmi les trois types d'acteurs, les mutuelles dominent sur le marché de la santé complémentaire, leurs chiffres d'affaires représentent 48,5% (voir figure 10) du marché et proviennent majoritairement des contrats individuels, les institutions de prévoyance avec 16,3% du marché et enfin, les sociétés d'assurance 35,3% du marché) ont une activité assez semblable entre contrats individuels et contrats collectifs, mais leur chiffre d'affaires dans le domaine de la santé reste très minoritaire.

Sur le marché de l'assurance santé collective, les écarts sur la répartition des chiffres d'affaires sont moins importants, les sociétés d'assurance sont majoritaires sur ce marché avec 39%, suivi par les mutuelles avec 33% de part de marché et enfin les institutions de prévoyance avec 29%.

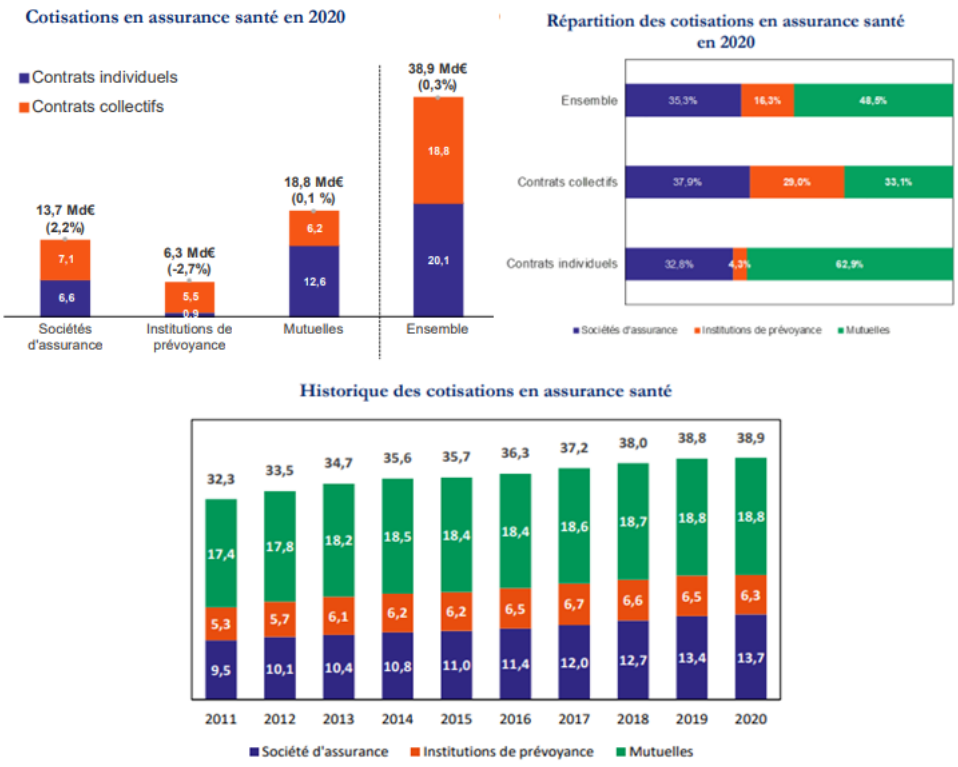


FIGURE 10 – Les organismes complémentaires santé, source : France Assureurs

Toutefois, avec la généralisation de la complémentaire Santé ANI, la part des sociétés d'assurance collective a tendance à s'accroître au profit des salariés du secteur privé.

Depuis vingt-cinq ans, le nombre d'organismes pratiquant des activités d'assurance est en baisse constante. En 2020, 683 organismes assureur de toute nature sont présents sur le marché selon l'Autorité de Contrôle Prudentiel et de Résolution (ACPR). En santé collective, nous retrouvons 369 organismes de mutuelle, 281 des sociétés d'assurance ou de réassurance et 33 institutions de prévoyance.

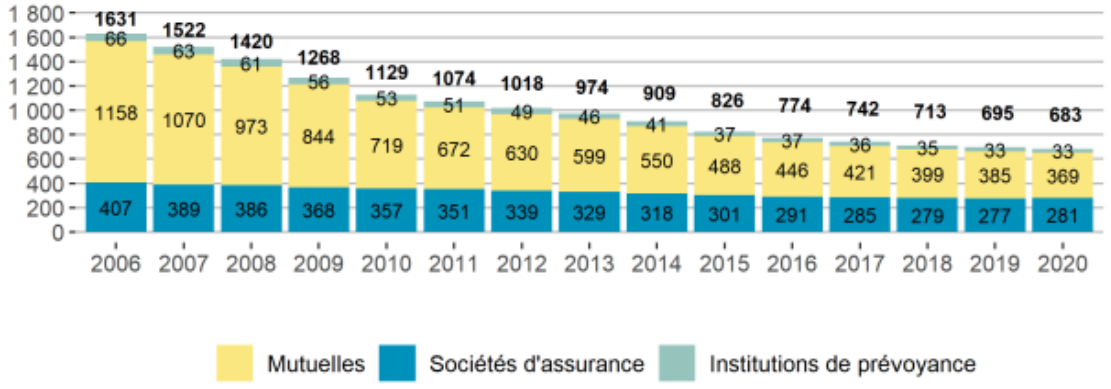


FIGURE 11 – Nombre d'organismes ; Source : ACPR

De 2006 à 2020, le nombre d'organismes a ainsi diminué de 67% chez les mutuelles, de 48% chez les institutions de prévoyance et de 32% chez les sociétés d'assurance, principalement par effet, de concentration. Le nouveau régime Solvabilité II a accentué cette tendance depuis 2013. La généralisation de la complémentaire santé d'entreprise au 1er janvier 2016 a également pu conduire à des regroupements ou à la création d'alliances et de partenariats sur le marché du collectif.

1.1.3 Mécanisme de remboursement

Pour un remboursement de frais de santé, il a trois intervenants : la Sécurité sociale, un organisme d'assurance complémentaire et l'assuré.

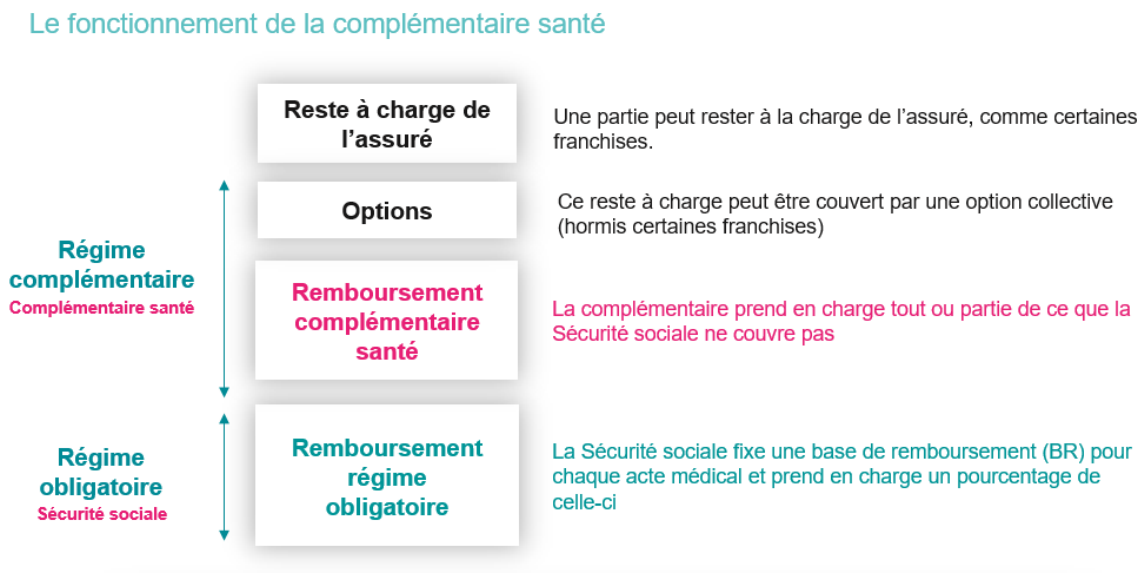


FIGURE 12 – Fonctionnement d'une couverture santé

Une partie des frais de santé est remboursée par la Sécurité sociale qui se base sur un tarif de référence appelé base de remboursement, celui-ci varie selon les actes médicaux. En effet, la Sécurité sociale estime un pourcentage appelé taux de remboursement de cette base de remboursement qui reviendra à sa charge. Ainsi, tout affilié à un régime de la Sécurité sociale bénéficie de cette prise en charge de base.

La part non remboursée de la base de remboursement correspond au ticket modérateur (TM). Le ticket modérateur revient à la charge de l'individu, de même que tout dépassement d'honoraires éventuel. Cette zone correspond au champ d'intervention des organismes complémentaires sous réserve que l'individu ait souscrit à un contrat d'assurance.
De plus, dans le but de responsabiliser les adhérents sur certains actes, une participation forfaitaire à la charge de l'assuré a été instaurée. Ce montant est obligatoirement déduit du remboursement de la Sécurité Sociale.

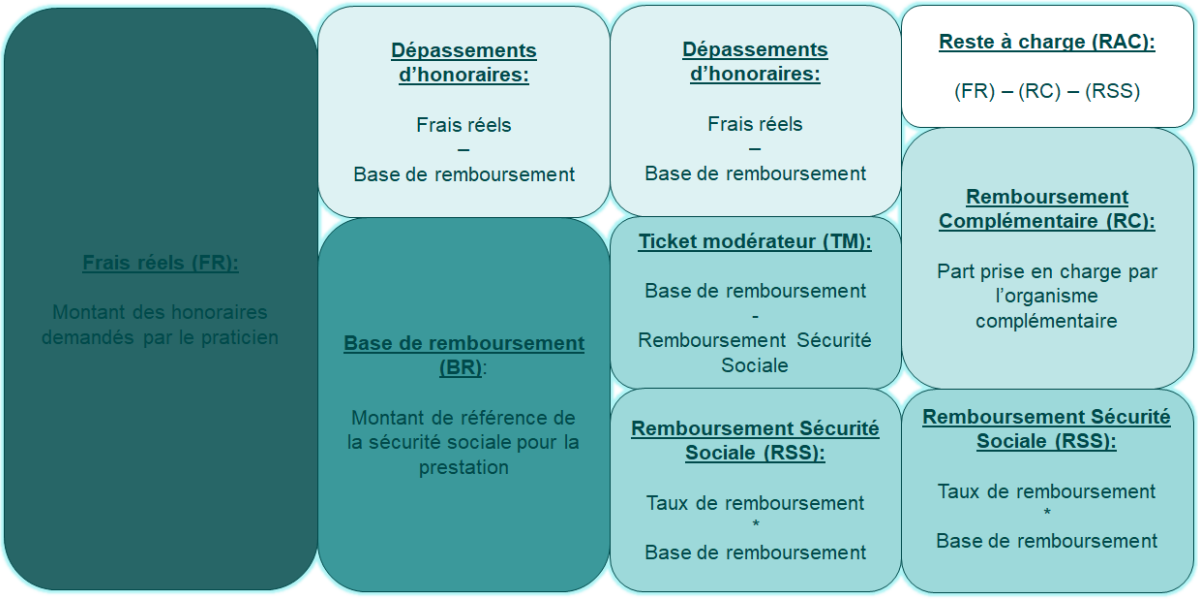


FIGURE 13 – décomposition des frais de santé

L'assurance complémentaire santé intervient 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 exemple, les forfaits de médecine douce, les cures thermales.

Illustrons cela avec un exemple :
Un contrat prévoit une garantie « Consultation spécialiste » de 120% de la base de remboursement (BR) sous déduction des prestations versées par la Sécurité sociale. Un individu a consulté le médecin (adhérant à OPTAM) spécialiste.
La base de remboursement de cette consultation spécialiste est de 28 euros. Les prestations versées par la Sécurité sociale et par les organismes assureurs sont les suivantes :

- Participation forfaitaire de 1€.
- Sécurité sociale rembourse 70% BR auquel il faut soustraire la participation forfaitaire, soit 18,6 €
- Ticket Modérateur correspond à 30% BR, soit 8,4€
- Dépassements honoraires est la différence entre le coût de la prestation et la BR, soit 12€
- Remboursement de la complémentaire santé est la somme du ticket modérateur et des dépassements honoraires plafonnée à 50% BRSS, soit min(20,4€ ; 14€)

1.2 La réglementation en assurance santé collective

1.2.1 L'Accord National Interprofessionnel (ANI)

L'accord national interprofessionnel (ANI) est un accord négocié et signé par les différents partenaires sociaux au niveau national et qui s'applique à l'ensemble des secteurs d'activités sur le territoire national.

Les partenaires sociaux sont constitués des représentants des principaux syndicats de salariés et d'employeurs. Ils sont à la base du dialogue social et agissent notamment dans la mise en place de garanties collectives de protection sociale au travers des conventions collectives et des accords de branche.

L'ANI porte notamment sur les conditions de travail et les garanties sociales des salariés. L'accord du 11 janvier 2013 a, par exemple, imposé à toutes les entreprises de proposer une couverture complémentaire santé collective à leurs salariés à compter du 1er janvier 2016.

1.2.2 Les contrats responsables

Un contrat est dit non responsable lorsqu'il ne remplit pas le cahier des charges redéfini par le décret du 18 novembre 2014. Cela correspond à un minimum et un maximum pour chaque garantie à respecter pour le régime base obligatoire mis à disposition par l'employeur.

Les cotisations sont taxées à 13,27% pour le régime général et à 6,27% pour le régime agricole si les contrats sont responsables, tout dépassement est taxé à 20,27% c'est-à-dire lorsque le contrat est non responsable. En plus de ces contraintes, la Convention Collective Nationale impose un minimum pour chaque garantie.

1.2.3 Le 100% santé

Dans l'objectif de diminuer le renoncement aux soins de la population française pour des raisons financières, le président de la République Emmanuel Macron a mis en place la reforme 100%. Cette mesure consiste à définir des paniers de soins sur les équipements optiques, dentaires et audioprothèses dont le reste à charge est important afin que la qualité soit encadrée par des normes, à des prix de vente limités, sans reste à charge pour les assurés, elle permet également d'améliorer la lisibilité des garanties.

- En optique : depuis 2020, les prix de vente ont été limités et si l'assuré choisit une monture et des verres de l'offre 100% Santé, il n'a plus rien à payer. La paire de lunettes lui est entièrement remboursée par l'Assurance Maladie et sa complémentaire santé.
- En dentaire : Il existe 3 paniers pour les couronnes, les bridges et les dentiers mis en place en 2020 et 2021.

1. le panier 100% Santé : les couronnes, les bridges et les dentiers sont intégralement remboursés si le patient bénéficie d'une complémentaire santé qui le prévoit (« contrat responsable ») ;
2. le panier aux tarifs maîtrisés : il intègre des couronnes, des bridges et des dentiers des prothèses dentaires dont les prix sont plafonnés. Selon les conditions de son contrat santé, l'assuré peut avoir un reste à charge à payer, mais modéré ;
3. le panier aux tarifs libres : le reste à charge peut être plus important pour l'assuré, selon son contrat de mutuelle.

— En auditif : depuis le 1er janvier 2021 avec la réforme 100% Santé, les remboursements de l'Assurance Maladie et de la complémentaire santé ont augmenté et l'assuré n'a plus rien à payer, il n'y a plus de frais à sa charge.

1.2.4 Autres textes

La loi de sécurisation de l'emploi

La loi de sécurisation de l'emploi du 14 juin 2013, oblige les entreprises à proposer une complémentaire santé à tous leurs salariés depuis le 1er janvier 2016.

Le financement employeur

La loi (art. L911-7 Code de la Sécurité sociale) impose aux employeurs de participer au financement de la complémentaire santé à hauteur de 50% minimum sur la couverture de base obligatoire choisie par l'employeur.

— Périmètre de prise en charge : La prise en charge par l'employeur s'applique sur la couverture obligatoire choisie par l'employeur.
— La participation éventuelle du CE n'entre pas dans l'appréciation de participation à 50% mais peut intervenir en plus de la participation employeur.
— La participation n'est pas substituable à un élément de rémunération, elle intervient en plus du salaire.

La couverture des populations périphériques

Loi Evin article 4 :

L'assureur doit prévoir les modalités et les conditions tarifaires du maintien des garanties santé au profit d'anciens salariés (retraités, bénéficiaires d'une rente d'incapacité ou d'invalidité, licenciés) et des ayants droit d'assurés décédés pendant une durée minimale de douze mois à compter du décès, qui en font la demande dans les 6 mois qui suivent la rupture du contrat de travail ou la date du décès de l'assuré, ou l'expiration de la couverture de la portabilité.

— Garanties : L'assureur propose un contrat proposant exactement les mêmes garanties que celui des actifs au jour de la rupture du contrat de travail, sans condition de période probatoire ni d'examens médicaux.

— Conjoint : L'obligation de maintien de garantie ne s'applique pas au conjoint.

— Tarif : Le tarif est identique la 1 année, puis +25% la 2de année et +50% à compter de la 3 année. Au-delà, l'évolution des cotisations selon les dispositions de l'article 6 de la loi Evin.

— Durée : Le maintien de la couverture est sans condition de durée.

Portabilité des droits :
Tout salarié perdant son emploi hors faute lourde, qui bénéficie d'une couverture complémentaire de santé et de prévoyance (décès, incapacité, invalidité) obligatoire ou facultative (option) au sein de son ancienne entreprise peut continuer à en bénéficier pour une durée égale à celle du ou des derniers contrats de travail consécutifs, exprimée en mois entier, dans la limite de 12 mois.

— Conditions à respecter :
Adhésion aux régimes de l'entreprise à la date de rupture du contrat de travail.
Rupture ou cessation du contrat de travail, hors licenciement pour faute lourde.
Prise en charge par l'assurance chômage.

— Tarif : Maintien des garanties à titre gratuit (mutualisé avec le tarif des actifs majorés en conséquence).

— Changement d'assureur : En cas de changement d'assureur, le nouvel assureur reprend la couverture des salariés en portabilité.

1.3 Contexte marché de l'assurance collective

1.3.1 Un contexte tendu

Le monde de l'assurance santé complémentaire est régulièrement soumis à des changements, en particulier pour les assurances collectives qui sont structurellement très concurrentielles, ce qui conduit les assureurs à constamment adapter leurs offres et à renforcer les enjeux de tarification et de surveillance de portefeuille.

Les évolutions réglementaires en faveur des assurés rendent le marché de l'assurance santé plus concurrentiel qui ne l'est déjà. Dans ce contexte, les assureurs doivent proposer des garanties adaptées à la sinistralité des individus et les tarifs les plus fins. Ce qui entraîne des tarifs tirés vers le bas et une baisse des marges techniques.

C'est pourquoi, certains acteurs n'hésitent pas à opérer de larges stratégies de rapprochements. Depuis plus d'une dizaine d'années, le nombre d'organismes a ainsi fortement diminué : il a baissé de 68% chez les mutuelles, de 50% chez les institutions de prévoyance et de 31% chez les sociétés d'assurance de 2006 à 2020 selon l'ACPR (voir figure 11), principalement par fusions ou absorptions avec transferts de portefeuille de contrats avec les droits et obligations qui s'y rattachent.

Cependant, la diminution du nombre d'organismes assurant une activité de complémentaire santé rejoint plus généralement la baisse de celui des organismes exerçant une activité d'assurance.

De plus, ces dernières années, les dépenses en santé ne cessent d'augmenter, cette hausse est en partie liée au vieillissement de la population, l'accroissement des malades chroniques et les progrès médicaux entraînant une augmentation du coût des traitements et des dérives des prestations.

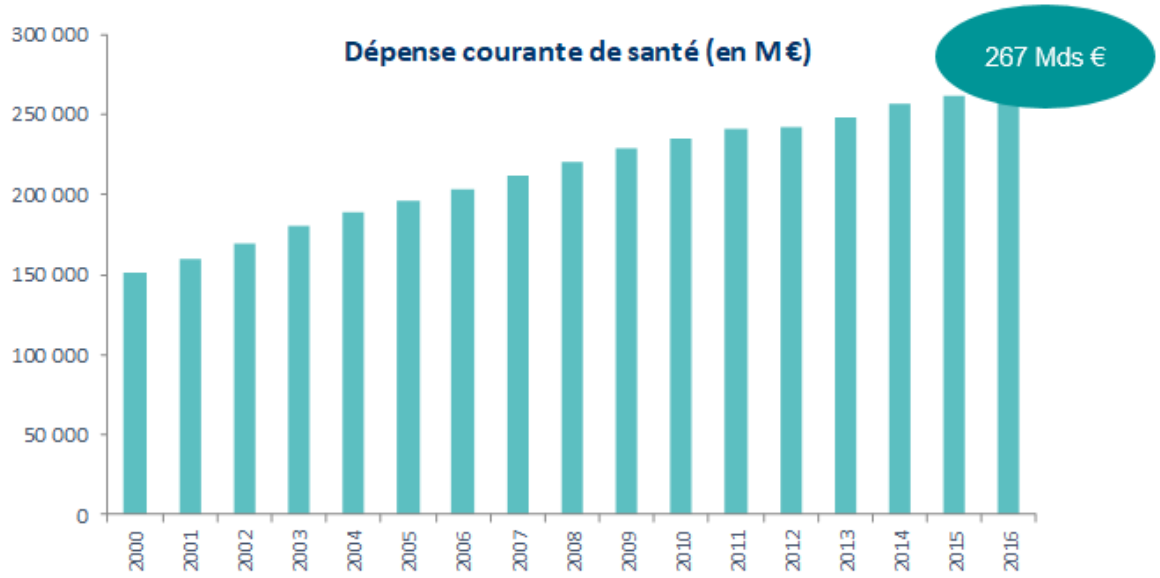


FIGURE 14 – Les dépenses en santé

À tout cela s'ajoute le désengagement progressif de l'assurance publique en France, ce qui entraîne une intervention plus importante des complémentaires qui voient augmenter leurs prestations versées aux assurés.

Face à ces problématiques, les assureurs doivent trouver de nouvelles méthodes pour piloter leur risque. En particulier, suite à ces dernières années exceptionnelles, la pertinence des modèles traditionnels fondés sur des données historiques est remise en question.

Dans la course à l'innovation, les assurés doivent tirer profit des progrès technologiques et inévitablement en faire une force. Deux leviers au Crédit Agricole Assurances sont activés pour gagner en efficacité et en performance : la digitalisation et l'approche-conseil.

1.3.2 Le groupe Crédit Agricole : rôle de banque/assureur

Dans cet environnement réglementaire devenu plus strict, les intermédiaires ont su tirer leur épingle du jeu. En effet, fort de leur proximité et de leur rôle de conseil, les agents et courtiers se sont, en effet, imposés comme des interlocuteurs privilégiés pour les assurés.

Cette posture d'accompagnement et conseil, crée de la valeur ajoutée dans le déploiement et le pilotage des assurances collectives. Les bancassureurs ont très bien compris ce modèle développement et ils sont également bien avantagés avec la base client banque, en particulier dans le groupe Crédit Agricole.

Le groupe Crédit agricole

Le Crédit agricole est le plus grand réseau de banques coopératives et mutualistes au monde. C'est le premier bancassureur en Europe, le premier gestionnaire d'actifs en Europe et le premier financeur de l'économie française. En France, il est composé des 39 caisses régionales de Crédit agricole.

Le Crédit Agricole Assurances

Le Crédit Agricole Assurances est une filiale du groupe crédit Agricole S.A. Il est composé de plusieurs entités : Crédit Agricole Assurances Solution, Predica qui est le 2ème assureur en France, Spirica, CACI 2ème assureur emprunteurs en France, Pacifica premier bancassureur de France, et La Médicale. Il y a également sept filiales internationales présentes dans six pays différents.

Le mode de distribution repose essentiellement sur le modèle de bancassurance, c'est-à-dire vie les réseaux bancaires du groupe Crédit Agricole. Le groupe est organisé en Business Units réparties en trois pôles :

— Le Pôle Vie France regroupant les Business Units Prévoyance Emprunteur, Épargne Retraite, Collectives, Spirica/ULP.

— Le pôle Non Vie France avec Pacifica/Dommages et La Médicale.

— Le Pôle International avec l'Assurances à l'international.

Ces Business Units sont appuyés par des services units, ceux des fonctions clés secrétariat Général, Fonction actuarielle, Risques et Audit et d'autres services units comme IT Achats, Finances, Investissements, RH, Innovation, RSE et Communication et les Moyens généraux et sécurité.

Le Crédit Agricole Assurance s'est lancé dans l'assurance collective en 2015, proposant des produits de complémentaire santé, de prévoyance et de retraite à destination des entreprises.

Ces activités sont concentrées au sein de la BU Collective qui se distingue en deux directions :

- La direction gestion, service client et marketing : cette direction s'attache au développement de l'offre et accompagne le client de l'appel d'offre au suivi du contrat en passant par la contractualisions.
- La direction développement et technique : elle définit les caractéristiques des régimes proposés aux entreprises, en particulier les ETI. Elle est composée de 4 équipes : grands comptes, animation des réseaux (caisses régionales et LCL), souscription, comptes clients.

L'ADN de l'assurance collective au sein du groupe crédit agricole s'articule autour de trois grands piliers : conseil client, la digitalisation et le service.

Le conseil client est un atout majeur dans le déploiement de l'activité, en effet, elle a permis la conquête d'un certain nombre de clients de taille intermédiaires et grands comptes.
Dans le cadre de l'accélération du développement commercial, en maintenant cette approche, l'industrialisation de nos méthodes et procédures devient un enjeu crucial, pour notamment conquérir de nouvelles entreprises de taille intermédiaire.

Au sein de la direction développement et technique, l'équipe chargé de grands comptes travaille en étroite collaboration avec les services marketing et commerciale afin de développer l'activité à travers les réseaux de clients des caisses régionales. L'objectif est

Les assureurs grignotent des parts de marché au détriment des mutuelles et des groupes de protection sociale. Parmi eux, les bancassureurs sont ceux qui tirent le plus leur épingle du jeu, même si les disparités sont fortes d'une entité à l'autre.

1.3.3 Une approche-conseil testée, validée et à industrialiser

Le contexte des assurances complémentaires santé étant très concurrencé et très contraint par les réglementations, les assureurs doivent ainsi établir de nouvelles stratégies dans le déploiement et le pilotage de leurs activités. Le crédit agricole assurance a très bien assimilé l'environnement tendu de la complémentaire santé collective et profite de son statut de bancassureur pour développer son activité à travers son réseau clients des 39 caisses régionales.

En effet, les assurances collectives du groupe ont opté pour une approche-conseil qui s'avère être très efficace dans le développement commercial de l'activité. Cette démarche consiste à réaliser une étude complète du régime et des besoins d'un groupe afin d'en

déterminer les axes d'améliorations en termes de couverture, de tarif ou encore de gestion.

L'accompagnement des clients dans le choix de la protection sociale proposée à leurs salariés apporte une forte valeur ajoutée d'un point de vue client. Les entreprises ont l'obligation de proposer une couverture santé à l'ensemble des salariés, celle-ci présente un atout pour conserver leurs salariées.

En plus de fidéliser le client, cette démarche permet également de mieux piloter le portefeuille puisqu'elle nous engage à construire un régime sur mesure adapté aux besoins du client avec un tarif également adapté.

L'approche-conseil précédente mise en place

Le système actuel en place pour cette approche-conseil consiste pour les groupes ciblés à évaluer la performance de la couverture santé établie à partir d'un système de notation construit sur une base de prestations du portefeuille qui prend en compte la situation géographique.

De nombreux facteurs interviennent dans la mesure de la qualité d'une couverture santé, la situation géographique est l'un des facteurs les plus importants. En effet, le tarif qu'une prestation de soins peut beaucoup varier d'un spécialiste à un autre, mais également avec la localisation. Pour des raisons particulières, il est connu que le coût de vie est souvent plus important dans les grandes villes urbaines et moindre dans les zones rurales.

L'intégration du facteur géographique commence par le regroupement des régions françaises en trois zones par rapport au coût moyen des frais de soins afin de distinguer les régions qui ont des frais moindres et ceux de qui ont des frais plus importants. Le partitionnement des régions dans l'approche actuelle est construit uniquement sur l'expérience.

La répartition des régions est la suivante :

— La zone 1 regroupe les régions ayant les coûts de prestation les plus faibles : l'Auvergne, La Bretagne, Centre, la Champagne-Ardenne, la Franche compte, le Limousin, la Lorraine, le Nord pas de Calais, le Pays de la Loire, la Picardie, le Poitou-Charentes.

— La zone 2 se compose principalement des régions provinciales comportant une ou plusieurs villes urbaines, on y retrouve les régions du sud de la France : l'Aquitaine, le Languedoc-Roussillon, les Midi-Pyrénées, la Provence Alpes Côte d'Azur, ainsi que l'Alsace, la Haute-Normandie et la Bourgogne.

— La zone 3 isole l'Île-de-France des autres régions et est définie comme la zone la plus chère en termes de prestations. En effet, le coût des prestations comprend souvent des dépassements d'honoraires.

Ce zonier est-il en phase avec la base de prestations du portefeuille ?

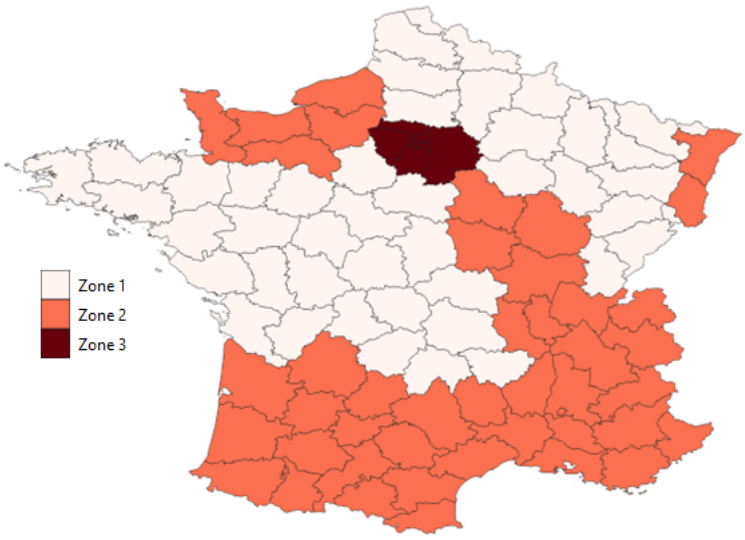


FIGURE 15 – Zonier actuel

Pour répondre à cette interrogation, comparons le zonier actuel aux figures 16 et 17 sur lesquelles on retrouve les quantiles des coûts moyens par régions du portefeuille. Si nous confrontons les trois zones du zonier actuel aux trois zones construites à l'aide de trois quantiles, on constate plusieurs similitudes. En effet, des régions placées en zone 1 se retrouvent dans le premier quantile Q1 (figure 16) notamment la Bretagne ou encore la Franche-Compté. Cependant, nous y retrouvons également des régions n'ayant pas la même position dans ces deux partitionnements. Pour des zones ordonnées, nous distinguons deux différences :

— La différence d'un niveau, c'est-à-dire passer d'une zone k à la $k+1$ ou de la zone k à la $k-1$ dans ce cas la différence peut s'expliquer par la césure entre deux zones adjacentes. Par exemple, la région Centre ou les Midi-Pyrènes se situent en zone 1 dans le zonier actuel, mais se retrouvent dans le 2e quantile Q2 des coûts moyens par régions.

— La différence de deux niveaux, c'est-à-dire passer d'une zone k à $k+2$ ou de la zone k à $k-2$, c'est le cas de la Picardie. Cette différence peut être d'une position non correcte dans le zonier actuel, mais également dû à plusieurs facteurs que nous aborderons dans la suite.

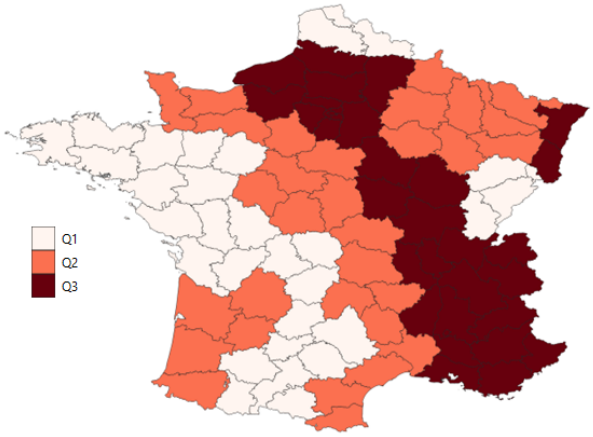


FIGURE 16 – Quantiles des coûts moyens par région

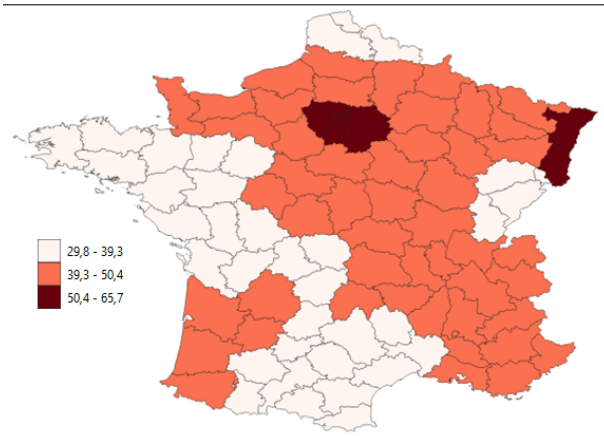


FIGURE 17 – Jenks des coûts moyens par région

En revanche, le partitionnement de régions donné en figure 17 utilise le principe Jenks, cette méthode regroupe les régions en fonction de la distribution des couts moyens afin de réduire la variance au sein des éléments des groupes et à maximiser la variance entre les groupes. Le résultat engendré par cette méthode isole l'Île-de-France et l'Alsace comme les régions aux frais de soins les plus chers, ce qui confirme la position de l'Île-de-France du zonier actuel.

Le cas de l'Alsace est à argumenter, il a été évoqué dans la partie précédente l'existence de régimes particuliers, la région Alsace Moselle possède un régime particulier dans lequel la part du remboursement de la Sécurité sociale est plus importante que dans le régime général conduisant à un tarif de la complémentaire santé plus faible, ce qui entraîne une couverture plus importante et ainsi des coûts moyens des prestations plus forts.

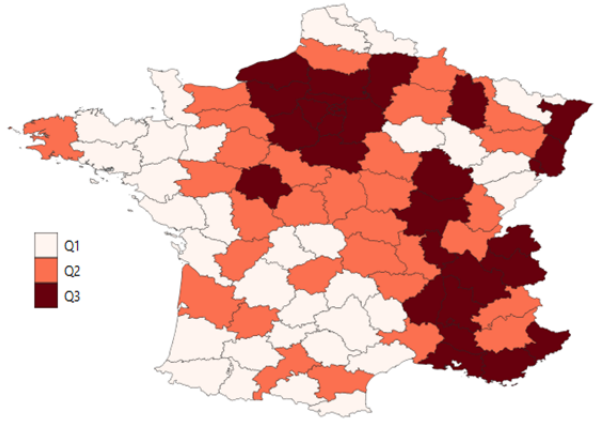


FIGURE 18 – Quantiles des coûts moyens par département

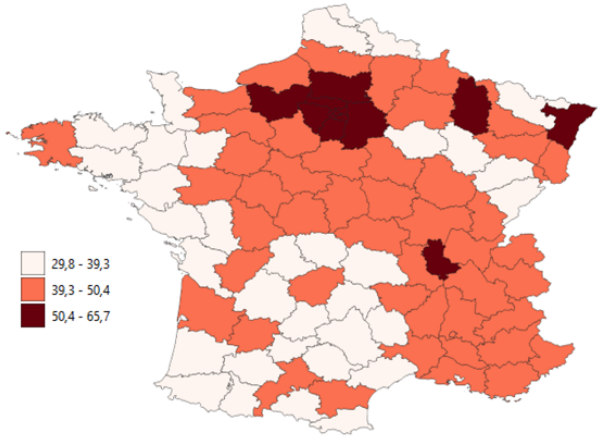


FIGURE 19 – Jenks des coûts moyens par département

À présent, nous allons examiner les mêmes types de répartition des coûts moyens, mais cette fois à l'échelle départementale. Cette approche permet de fournir des informations supplémentaires et une précision plus fine que la méthode de regroupement par région. Les résultats confirment la position de l'Île-de-France en tant que région la plus coûteuse,

avec chacun de ses départements ayant également une position élevée en termes de coûts. En outre, les départements de l'Oise, de l'Eure et du Rhône se distinguent parmi les départements les plus coûteux selon la méthode Jenks. La Meuse est également identifiée comme un département coûteux, mais cela s'explique par des effets d'effectifs qui rendent les indicateurs moins robustes.

Pour chacune de ces trois zones, une grille de notation a été construite en suivant la procédure suivante : Pour chaque acte de soins, cinq niveaux de coûts sont définis à partir des quantiles 25, 50, 75, 90 et 95 provenant d'une base de frais de soins enregistrés en portefeuille pour l'acte considéré. Prenons l'exemple du tableau ci-dessous pour l'acte de prothèse dentaire. La note 1 est associée au quantile 25 qui s'interprète comme suivant pour la zone 1 : 25% des prestations de prothèses dentaires couverts en portefeuille coûtent moins de 250% de la base de remboursement (BR).

Note	Quantile	Zone 1	Zone 2	Zone 3
1	25	250%	300%	350%
2	50	280%	350%	410%
3	75	330%	410%	470%
4	90	390%	480%	530%
5	95	450%	540%	590%

La note d'une garantie intermédiaire (entre deux garanties associées à des notes entières) est définie par l'interpolation linéaire de l'intervalle des notes à laquelle cette dernière appartient. Pour reprendre l'exemple du tableau, la note une garantie G pour des soins de prothèses dentaires réalisés dans la zone Z est définie comme suivant :

$$Note(Z,G) := \begin{cases} \frac{G}{G^Z_{q25}} \text{ pour } G \in I^Z_1 =]G^Z_{q0}; G^Z_{q25}[\\ 1 + \frac{G-G^Z_{q25}}{G^Z_{q50}-G^Z_{q25}} \text{ pour } G \in I^Z_2 =]G^Z_{q25}; G^Z_{q50}[\\ 2 + \frac{G-G^Z_{q50}}{G^Z_{q75}-G^Z_{q50}} \text{ pour } G \in I^Z_3 =]G^Z_{q50}; G^Z_{q75}[\\ 3 + \frac{G-G^Z_{q75}}{G^Z_{q90}-G^Z_{q75}} \text{ pour } G \in I^Z_3 =]G^Z_{q90}; G^Z_{q75}[\\ 4 + \frac{G-G^Z_{q25}}{G^Z_{q50}-G^Z_{q25}} \text{ pour } G \in I^Z_2 =]G^Z_{q25}; G^Z_{q50}[\end{cases}$$

Et $Note(Z, G^Z_{q0}) = 0$, $Note(Z, G^Z_{q25}) = 1$, $Note(Z, G^Z_{q50}) = 2$, $Note(Z, G^Z_{q75}) = 3$, $Note(Z, G^Z_{q50}) = 4$, $Note(Z, G^Z_{q95}) = 5$ et pour $G > G^Z_{q95}$ $Note(Z, G) = 5$.

Ainsi, pour une couverture de 275% BR, on obtient pour les zones 1, 2 et 3 les notes respectives de 1.83, 0.92, 0.79.

En utilisant cette méthode, la note d'un poste de soins est déterminée en calculant la moyenne des notes des différents actes de soins qui le composent, pondérée par leurs poids en termes de prestations dans la base de données qui a permis d'établir les quantiles. Enfin, la note d'un régime de santé est calculée en faisant la moyenne des notes des cinq postes pondérés par leurs poids en prestations dans la base d'étude.

Contrairement à la façon dont est construite la note d'une couverture santé, l'analyse de celle-ci commence par le niveau le plus général pour ensuite descendre dans les détails. L'analyse commence par la note de la grille de garanties, puis celles des postes de soins et enfin termine avec les notes des actes de soins. Ainsi est évaluée la performance d'une couverture santé dans un processus d'audit ciblé sur des groupes sélectionnés en amont de la tarification.

Ce processus présente plusieurs avantages. Tout d'abord, il permet non seulement de comparer une couverture santé à d'autres couvertures dans le même secteur d'activité, tout en garantissant l'anonymat. Ensuite, il permet d'analyser la couverture pour identifier les lacunes et les points forts, détecter les garanties qui sont excessivement élevées et qui pourraient être évitées. Enfin, il contribue dans l'harmonisation des différents régimes de couverture pour un client donné.

Cette posture de conseiller vise à accompagner le client dans le processus de fournir à ses employés une couverture de santé optimale avec un point de vue critique sur sa situation actuelle.

Une approche-conseil validée

Le conseil client s'est avéré être un atout majeur dans le déploiement de l'activité, en effet, il a permis la conquête d'un certain nombre d'entreprises de taille intermédiaire (ETI) en particulier des groupes avec une masse salariale importante.
La figure 20 ci-dessous présente les appels d'offres traités de 2021 dans lesquels nous distinguons les offres ayant été auditées à l'aide du processus décrit plutôt et celles uniquement tarifées.

Nous pouvons constater que le taux d'appels d'offre gagnés en portefeuille est plus fort lorsqu'une étude complète du régime a été réalisée. Cependant, il est à noter que la démarche optée dans l'approche-conseil est un processus qui nécessite du temps, elle est ainsi proposée à des groupes cibles.

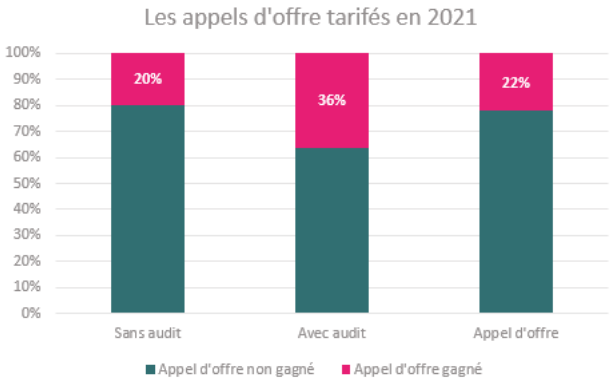


FIGURE 20 – Répartitions et résultats des appels d'offre tarifés

Les études de couverture sont davantage réalisées sur les groupes (voir figure 22), les groupes ayant une masse salariale plus importante sont les plus prisées, en effet, elles apportent un chiffre d'affaires plus conséquent.

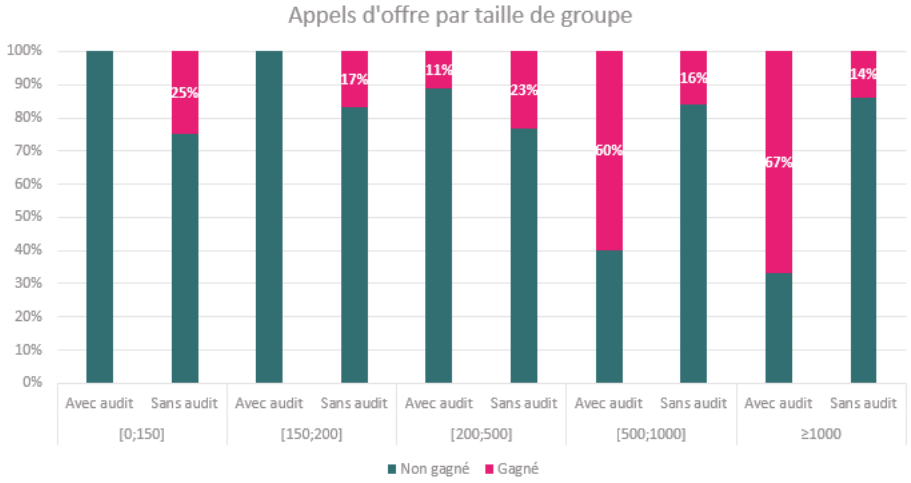


FIGURE 21 – Répartitions et résultats des appels d'offre tarifer

L'objet de ce présent mémoire est tout d'abord la refonte d'un zonier sur une base de données plus récente, mais également d'apporter de la précision en travaillant sur une maille plus fine : la maille départementale. Cependant, en travaillant sur une maille plus fine, nous récupérons plus d'informations, mais nous perdons de la robustesse sur l'information récupérée. Nous proposerons donc un procédé de lissage spatial basé sur la théorie de crédibilité qui fiabilisera l'information récupérée.

L'objectif second de ce mémoire est le développement et l'amélioration de l'outil d'audit de régimes santé selon deux axes principaux.

— L'amélioration de la métrique permettant d'analyser la performance d'un régime. La métrique actuelle est jugée trop subjective et perd du sens si celle-ci n'est pas contextualisée. Elle sera remplacée par un indicateur plus facilement interprétable d'un point de vue client.

— L'amélioration de l'outil s'entend également l'industrialisation de la procédure d'audit à travers l'automatisation de la construction des différents éléments d'analyse de l'outil sous VBA. Cette démarche vise à étendre le processus de benchmarking et de mesure de performance à un plus grand nombre de potentiels clients qui sont aujourd'hui ciblés.

2 Un outil d'aide au conseil pour se démarquer dans un marché tendu

Au sein des assurances collectives, l'approche-conseil et d'accompagnement auprès des entreprises couvre l'ensemble de l'univers des besoins en protection social notamment en santé, en prévoyance et également retraite. Cependant, dans ce mémoire, nous traitons uniquement le cas de la santé.

L'objectif de cette partie est de mettre en lumière tous les éléments caractérisant un régime et les étapes du processus d'une étude de performance d'une couverture santé.

2.1 Principes et démarches de l'approche-conseil

La stratégie mise en place par l'assurance collective de Crédit Agricole Assurances, afin de déployer et de mieux la piloter son activité, vise à accompagner les clients dans la protection sociale de leurs salariés. En santé, l'enjeu est de proposer un régime de couverture de soins médicaux optimal tout en limitant le reste à charge des bénéficiaires avec un budget très souvent limité. Cet accompagnement sur-mesure se traduit plusieurs étapes clés :

1. Tout d'abord, les besoins du client en termes de niveau de garanties et de budgets sont déterminés. Ces besoins peuvent notamment concerner une simple étude de la cohérence de leur couverture ou dans le cas de plusieurs régimes, une harmonisation de ces derniers.
2. Ensuite, nous procédons à une étude de benchmark de positionnement dans le secteur d'activité par rapport à la localisation. Cette étape consiste à comparer le régime actuel du client à d'autres couvertures du même secteur d'activité détenues en portefeuille à l'aide de métriques. Ainsi, nous pouvons déterminer la performance de celui-ci et la positionner sur leur marché au vu des améliorations que nous pourrons leur proposer, en particulier si leur régime actuel n'est pas en accord avec les besoins actuels.

L'étude mise en place pour évaluer la performance de régimes santé se ventile en trois grands volets :

— Les garanties : les grilles de garanties sont confrontées aux coûts des prestations couverts en portefeuille en prenant en compte l'architecture des garanties et la situation géographique.

— Les cotisations : Les budgets employeur et salarié sont analysés à partir des taux et des structures de cotisation.

— Le profil démographique.

En parallèle de la mesure de performance d'un régime santé, celui-ci est comparé à des couvertures santé d'autres entreprises du même secteur d'activité afin de déterminer la position par rapport aux performances du secteur. Ce processus de benchmarcking s'appuie sur les mêmes métriques évaluant les couvertures.

2.1.1 La métrique de comparaison et d'évaluation

Pour mesurer la performance d'un régime santé, nous considérons le taux de couverture moyen de chaque acte ayant un enjeu sur reste à charge d'un adhérent à ce régime. Celui-ci représente la prise en charge totale moyenne des frais relatifs à un acte de soins, à un poste de soins ou encore à l'ensemble des garanties d'une couverture. Le taux est construit à partir d'une base de prestations fiable.

Contrairement au système de notation décrit plus tôt, le taux de couverture est facilement interprétable pour un client et moins subjectif pour déterminer la performance d'une couverture santé. En effet, du point de vue du consommateur, il est unanime que la quantité de reste à charge est un très bon indicateur de performance.

Un régime complémentaire santé est caractérisé par une grille de garanties, dans laquelle le niveau de prise en charge de chaque type de prestation de soins est définie et ces actes de soins sont regroupés en cinq postes de soins. Pour construire le taux de couverture moyen d'une grille de garanties, il est nécessaire de définir le taux de couverture moyen d'un acte de soins ensuite celui d'un poste de soins.

Taux de couverture d'un acte de soins

Notons A un type d'acte de soins et $\mathcal{A}$ l'ensemble des actes de soins de type A couvert en portefeuille sur un périmètre d'étude fixé. Considérons un niveau de garantie g_A pour l'acte de soins de type A dont le montant maximal de remboursement, y compris la sécurité social, est p. Ainsi, nous définissons le taux de couverture $\mathcal{T}_c$

$$\mathcal{T}_c(g_A) = \sum_{a \in \mathcal{A}} n_a min\left(\frac{p}{p_a}, 1\right)$$

Où n_a est la proportion d'acte a de la base de prestations dont le montant de la prestation est de p_a

Le taux de couverture de g_A correspond à la moyenne des ratio de prise en charge par la garantie g_a sur les coûts correspondant aux actes de type A enregistrés en portefeuille.

Illustrons nos propos avec le cas d'une couverture sur les montures. Nous considérons une garantie remboursant à hauteur de 100 euros (y compris la charge de la Sécurité sociale).

	Couts d'une monture c_i	Nombre d'actes en % n_i	Nombre d'actes cumulés en %
1	55 €	8%	8%
2	60 €	6%	14%
3	75 €	20%	34%
4	80 €	5%	39%
5	85 €	7%	46%
6	90 €	15%	61%
7	95 €	35%	96%
8	105 €	2%	98%
9	110 €	1%	99%
10	120 €	1%	100%

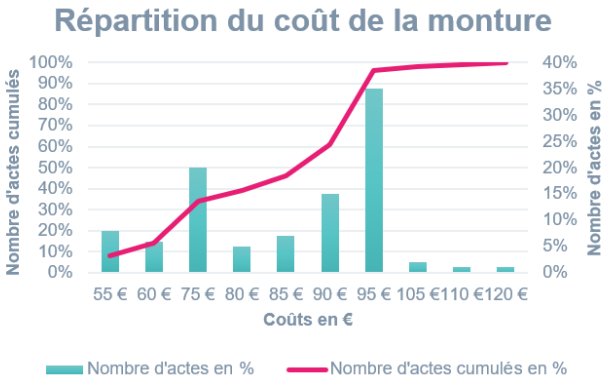


FIGURE 22 – Illustration du taux de couverture d'un acte de soins

Dans le tableau en figure 22, nous avons les différents montants pour une prestation de monture. Déterminons le taux de couverture sur cette base de prestations pour une garantie remboursant jusqu'à 100 euros. 96% des montures sont entièrement prises en charge par la couverture et le taux de couverture moyen.

$$\mathcal{T}_c(g_{monture}) = \sum_{i=1}^{10} n_i \min(\frac{100}{c_i}; 1) = 99,6\%$$

Une couverture remboursant jusqu'à 100€ pour monture rembourse en moyenne $99,6\%$ des prestations de l'exemple d'illustration.

Taux de couverture d'un poste de soins

Suite à la définition du taux de couverture d'un acte de soins, nous pouvons introduire celui d'un poste de soins. Pour un poste donné, le taux couverture correspond à la moyenne des taux de couverture des actes qui la composent, pondérés par leurs poids en prestation.

Plus formellement, pour un ensemble de garanties fixé g_P pour un poste de soins P et notons $\mathcal{T}_c(g_{A_1}), ..., \mathcal{T}_c(g_{A_{np}})$ les taux de couverture moyens des n_P actes soins composant le poste de soins, nous définissons le taux de couverture $\mathcal{T}_c(g_P)$.

$$\mathcal{T}_c(g_P) = \sum_{i=1}^{n_P} w_{A_i} \mathcal{T}_c(g_{A_i})$$

Où w_{A_i} est le poids en prestations de l'acte A_i dans le poste P de la base d'étude.

Postes	Actes	Poids de l'acte / poste
Hospitalisation	Honoraires en établissements non optam	
	Honoraires en établissements optam	
	Frais de séjour	
	Chambre particulière	
Soins courants	Honoraires généralistes non optam	
	Honoraires généralistes optam	
	Honoraires spécialistes non optam	
	Honoraires spécialistes optam	
	Honoraires actes techniques et radiologie non optam	
	Honoraires actes techniques et radiologie optam	
	Honoraires paramédicaux	
	Analyses et examens médicaux	
Dentaire	Prothèses dentaires maitrisées	
	Soins dentaires	
	Inlay onlay	
	Orthodontie	
	Implantologie	
Optique	Monture	
	Verre simple	
	Verre complexe	
	Chirurgie de l'œil	
	Lentilles adultes	
Autre	Prothèses et appareillages	
	Prothèses auditives	
	Consultations médecine douce	

FIGURE 23 – Les actes des soins par poste de soins

Il faut remarquer que certains biens et actes de soins standards tels que la pharmacie, la cure thermale ou encore les actes de maternité ne sont pas présents dans la figure 23. Ces prestations sont toujours couvertes, mais elles n'interviennent pas dans la construction de nos métriques pour des raisons que nous évoquerons plus loin dans ce mémoire.

Taux de couverture d'un régime santé

Après avoir défini le taux de couverture d'un poste de soins, nous pouvons construire le taux couverture global d'un régime correspondant à la moyenne des taux de couverture des postes pondérés par leurs poids en prestation. Notons g la grille de garanties d'un régime de complémentaire santé et pour $i = 1, ..., 5$, w_{P_i} le poids en pretations du poste de soins P_i.. Ainsi, nous définissons le taux de couverture global du régime g par $\mathcal{T}_c(g)$:

$$\mathcal{T}_c(g) = \sum_{i=1}^{5} w_{P_1} \mathcal{T}_c(g_{P_1})$$

Nous retrouvons en figure 24 le poids en prestations de chaque poste de soins de notre base d'étude.

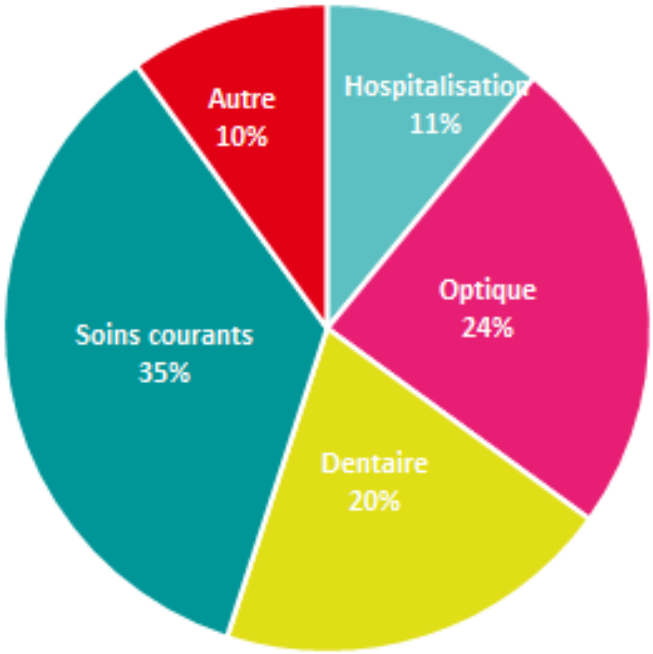


FIGURE 24 – Réparation des prestations par poste

La limite du taux de couverture moyen

Le taux de couverture moyen estime le reste charge moyen dans la consommation de biens de soins médicaux pour une couverture santé. Cette métrique évalue bien la performance d'un d'un régime santé. Néanmoins, il ne permet pas d'identifier les « sur-garantis ». c'est-à-dire les niveaux de garanties excessivement élevés pour des prestations qui n'atteindront rarement voir jamais ce coût. C'est pourquoi, le taux de couverture sera accompagné d'un ratio entre la garantie et le maximum observé dans la zone.

2.2 Évaluation et analyse des cotisations et financement

Le volet cotisation présente un enjeu majeur pour toutes les parties concernées, les clients veulent proposer à leurs salariés une bonne couverture limitée par un budget afin que les assurés paient le meilleur prix, que la participation financière de l'employeur reste modérée et l'assureur veut un tarif équilibré afin d'assurer ses engagements sans perte.

Les cotisations matérialisent donc l'engagement des assurés envers l'organisme d'assurance, afin de bénéficier d'une couverture santé. la partie obligatoire du régime est directement payée par l'employeur puisqu'il a l'obligation de financer ces cotisations au moins à 50%.

De plus, le choix de la structure de cotisation est un enjeu majeur pour les budgets. En effet, puisque l'employeur participe au minimum à hauteur de 50%, la structure de cotisation a un impact direct sur les budgets employeur et salarié.

2.2.1 Le financement

L'employeur a le choix de financer plus, voire la totalité de la couverture santé dans le but d'avantager leurs salariés. Ce taux de financement peut également être défini par la convention collective auquel est rattachée l'entreprise ou encore négocié par le comité d'entreprise. Il arrive souvent que le taux de financement varie lorsqu'une couverture est dissociée par catégorie socioprofessionnelle, dans ce cas, les salariés cadres sont souvent les plus avantagés avec des garanties sur le socle obligatoire plus élevées ou un taux de financement plus élevé.

2.2.2 Assiette de cotisation

Il existe plusieurs formats de cotisation que nous retrouvons en figure 25, le plus courant et privilégié par les organismes assureur est la cotisation en pourcentage du PMSS. Le PMSS est Plafond Mensuel de la Sécurité sociale utilisé comme base de calcul de certaines prestations sociales. Il est réévalué chaque année par les pouvoirs publics via un décret publié au journal officiel en fonction de l'évolution des salaires bruts des Français pour une date d'effet au 1 er janvier.

Le PMSS permet de prendre en compte des dérives quasi annuelles des prestations de soins.

Format de cotisation	Avantages	Inconvénients
en Euros	Cotisations figées	Pas d'indexation des cotisations en rapport avec l'évolution des dépenses de santé
en % du PMSS	Indexation automatique qui suit l'indexation de certaines prestations	Pas d'indexation des cotisations en rapport avec l'évolution des dépenses de santé
en % du salaire	- Cotisation qui varie selon les statuts et donc les salaires	- Cotisations décorrélées des prestations - Des cotisations différentes pour des prestations équivalentes
mixtes (Euros + % du salaire)	- Cotisation qui varie selon les statuts mais de façon limitée pour les salaires les plus importants - Permet d'instaurer une cotisation « plancher »	- Complexité de gestion et de paie

FIGURE 25 – Les différents formats de cotisation

2.2.3 La structure de cotisation

Les salariés peuvent couvrir leurs ayants droit (enfant à charge ou conjoint), il est imposé à l'employeur de proposer un régime permettant aux assurés de couvrir leurs ayants droit. Ainsi, il existe différentes structures de cotisation allant du plus solidaire avec un tarif uniforme au moins solidaire, répertorié dans la figure 26.

Structure	Tarif	Salarié	Conjoint	Enfant
Famille	Unique, quel que soit la situation	Obligatoire	Obligatoire	Obligatoire
Isolé / famille (famille : obligatoire)	Différencié, que l'on soit salarié seul ou avec ayants droit	Obligatoire	Obligatoire	Obligatoire
Salarié+ Enfant(s) / conjoint fac	Majoration en cas d'ajout du conjoint	Obligatoire	Facultatif	Obligatoire
Isolé / duo / famille	Dépend du nombre d'ayants droit affiliés	Obligatoire	Facultatif	Facultatif
Salarié / Conjoint / Enfant	Le tarif dépend du type d'affilié	Obligatoire	Facultatif	Facultatif

+ Solidaire

- Solidaire

FIGURE 26 – Les principales structures de cotisation

Lorsque l'adhésion des ayants droit est obligatoire, il faut noter que l'employeur participe également à leurs cotisations au même titre qu'un salarié. Un client qui veut avantager ses salariés aura plutôt tendance à se diriger vers une structure solidaire.

Depuis l'apparition de l'accord national interprofessionnel de 2016, le conjoint est en général couvert par le régime de son entreprise. Ainsi, afin d'éviter une double couverture du conjoint, le format de cotisation « Salarié + enfant(s) / conjoint facultatif » est de plus en plus répandu.

Structure de cotisation	Avantages	Inconvénients
Uniforme	-Cotisation identique quelle que soit la situation familiale -Permet une mutualisation (favorable aux familles qui paient la même cotisation qu'un célibataire)	Les célibataires (notamment les jeunes) paient la même cotisation que les familles
Salarié + enfant(s) / Conjoint fac	Permet de dissocier le conjoint qui peut être couvert par ailleurs par son employeur	Les salariés seuls paient la même cotisation que les parents célibataires
Isolé / Famille	Permet aux assurés de payer un tarif plus adapté à leur situation familiale et à leur consommation	Les familles composées de 2 assurés paient la même cotisation que les familles nombreuses
Isolé / Duo / Famille	Permet aux assurés de payer un tarif plus adapté à leur situation familiale (distinction couple avec ou sans enfant)	Les cotisations sont identiques quel que soit le nombre d'enfants de la famille
Adulte / Enfant	Permet de cotiser en fonction de sa situation familiale au plus juste	Pas d'effet de mutualisation pour les familles nombreuses

Le taux de financement employeur permet de compenser partiellement ou intégralement ces écarts de solidarité.

2.3 Évaluation et analyse des garanties

Le volet Garanties regroupe les garanties de la couverture santé, qui présentent les niveaux de prise en charge des soins médicaux après l'intervention de la Sécurité sociale, mais également l'architecture de couverture.

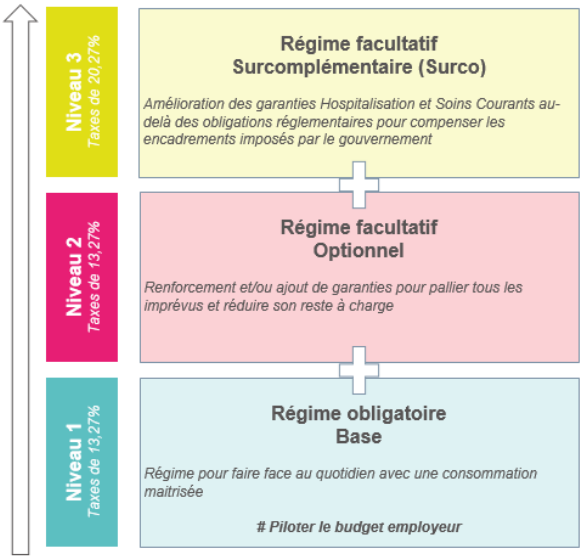
2.3.1 L'architecture d'un régime santé

L'architecture correspond aux différents niveaux de couches facultatives renforçant le régime base obligatoire si celui-ci n'est pas suffisamment élevé.
Ces options sont distinguées en deux catégories :

— L'option responsable vient compléter les garanties de la base.

— L'option non responsable également appelée sur-complémentaire, en général, elle est mise en place pour des clients se situant dans les régions où des dépassements importants sont pratiqués.

Le nombre de couches n'est pas limité, cependant il est assez restreint, il dépend généralement du niveau de garantie de la base, du secteur d'activité ou encore de la catégorie socioprofessionnel.



Nous constatons également que l'architecture mise en place par les entreprises dépend du secteur d'activité de l'entreprise et des négociations historiques menées.

2.3.2 Les garanties

Tous les actes de soins sont répartis en cinq postes que nous retrouvons dans la figure 27. Certains actes présentent une part importante de reste à charge, en particulier sur les postes optique, dentaire ou encore en hospitalisation avec des actes présentant des dépassements d'honoraires.

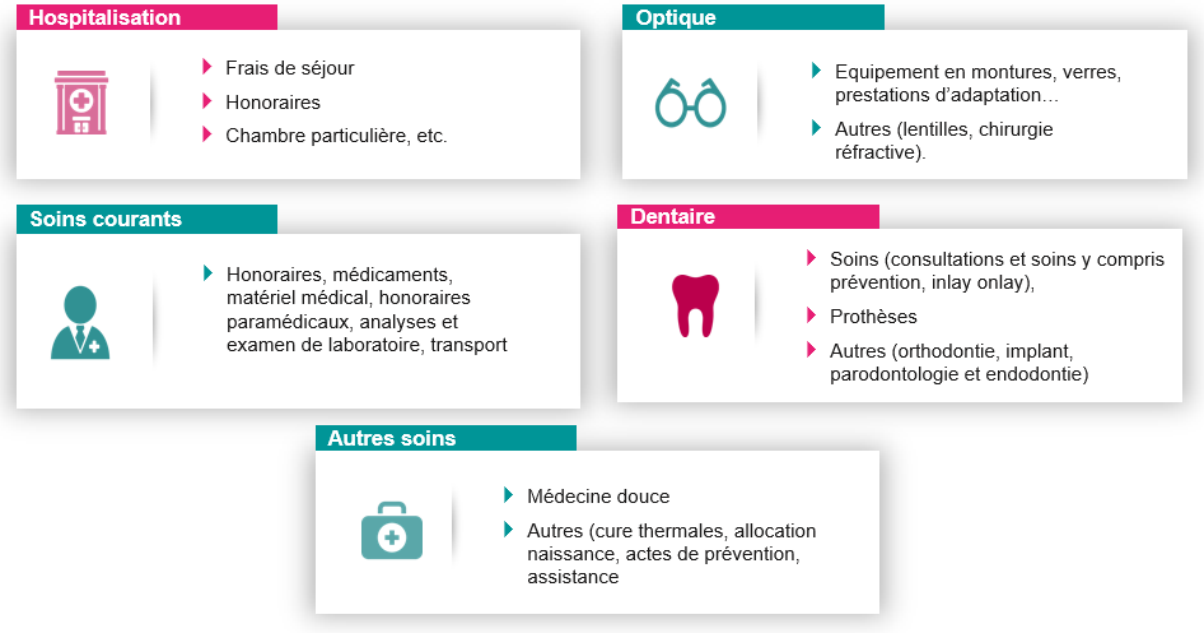


FIGURE 27 – Les postes de soins

La réforme 100% santé abordée dans la première partie introduit des paniers 100% qui regroupent des prestations sans reste à charge, cependant, ces paniers ne sont pas toujours favorisés par les assurés. En effet, en optique, le panier sans reste à charge est très peu utilisé.

Dans l'étude de performance d'une couverture, il est nécessaire de sélectionner uniquement les actes de soins qui présentent un enjeu dans notre contexte. En effet, certains types de prestations ne contribuent pas à l'évaluation, en particulier les actes sans reste à charge.

Dans un premier temps, nous identifions l'ensemble des actes qui contribuent à l'évaluation de la performance d'une couverture, en particulier les prestations courantes avec un gros reste à charge. Pour cela, il nous faut déterminer dans chaque poste les prestations ayant un enjeu pour évaluer une couverture de santé

La figure 28 représentant les remboursements moyens des différents acteurs de la complémentaire santé effectués par poste.

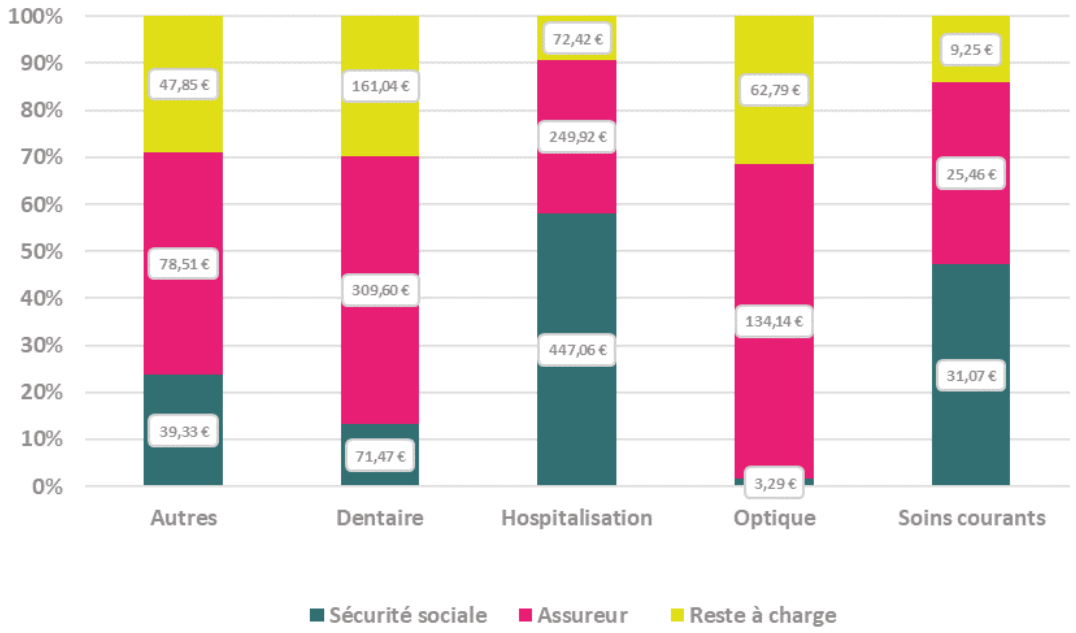


FIGURE 28 – Remboursement moyen par poste

Environ 60% des frais moyens d'hospitalisation sont couverts par la Sécurité sociale, tandis que l'assureur prend en charge environ 30%. Même si ces deux organismes prennent en charge 90% des frais, le coût restant pour l'assuré est élevé en raison des coûts importants de ce type de traitement. En moyenne, l'assuré doit débourser 72€.
En optique, le remboursement de la Sécurité sociale est quasiment inexistante, ainsi l'assureur est le principal contributeur avec près de 70% des frais pris en charge, tandis que l'assuré doit payer 30%, soit en moyenne 63€.
En dentaire, les frais restants à la charge de l'assuré sont les plus élevés, avec une moyenne de 161€, soit environ 30% du coût. C'est également le deuxième poste de dépense où l'assureur joue un rôle prépondérant, couvrant près de 60% des frais. Le poste soins courants présente le restant à charge le plus faible pour l'assuré avec en moyenne 9€, cependant, il comporte les actes fréquemment utilisés.

L'ensemble des postes de soins présente un enjeu sur le restant à charge non seulement en termes de coût avec plus particulièrement le dentaire, mais également avec la fréquence pour le poste de soins courants.

À présent, passons à l'identification des actes d'intérêt ayant du restant à charge important ou un restant à charge moindre avec une fréquence élevée pour chaque poste de soins.

Le poste Hospitalisation

L'hospitalisation est un risque de pointe, en général, un coût important sur ce poste est principalement dû à une dérive exceptionnelle par un ou plusieurs bénéficiaires. C'est également le poste le moins privilégié par les clients tandis que des restes à charge très importants peuvent-être constatés sur ce poste. il s'agit du poste dans lequel la part de remboursement de la Sécurité sociale est la plus élevée qui est en moyenne de 60% des frais réels et un reste à charge moyen de 63 euros ce qui représente près de 10% du coût réel.

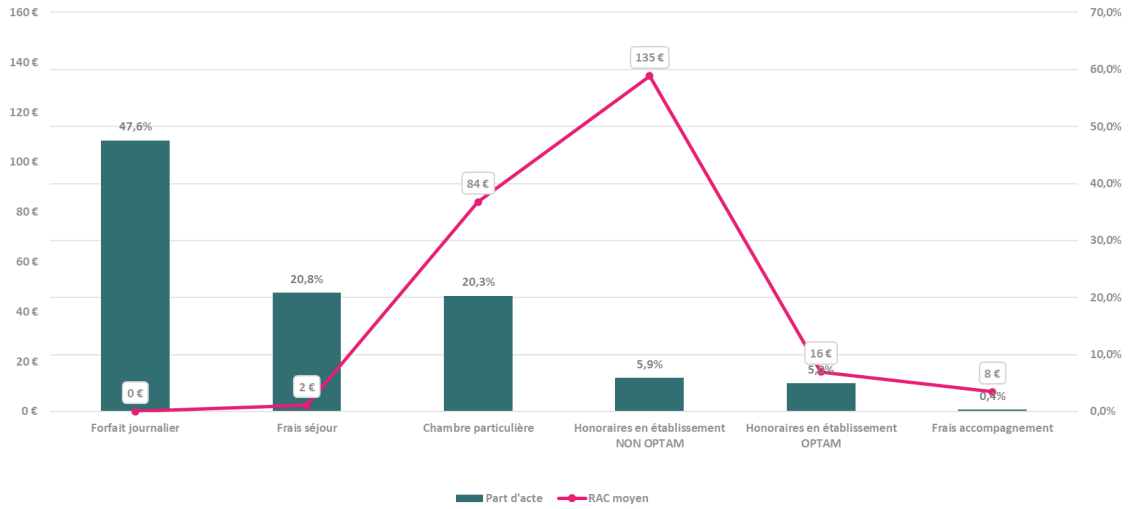


FIGURE 29 – Reste à charge moyen par acte en Hospitalisation

De plus, il présente un certain nombre d'actes comportant des dépassements d'honoraires, ce qui est générateur de reste à charge.

Les actes d'honoraires en établissement non optam (non adhérant à l'option de pratique tarifaire maîtrisée) et de chambre particulière présentent le plus d'enjeux sur ce poste, nous y retrouvons des restes à charge importants.

Le forfait journalier est très fréquent, mais il est pris en charge à 100% par la complémentaire santé. Le frais d'accompagnement est peu fréquent avec peu de reste à charge.

Ainsi, pour le poste hospitalisation, nos études seront établies sur les actes suivants :

- la chambre particulière qui est un acte fréquent avec un reste à charge important,
- l'honoraire en établissement non optam est un acte moyennement fréquent avec un reste à charge important,
- l'honoraire en établissement optam est un acte moyennement fréquent avec un reste à charge,
- "les frais de séjour" est un acte très fréquent avec peu de reste à charge.

Le poste Dentaire

Le dentaire est l'un des postes présentant le plus de restants à charge en moyenne 161 euros, ce qui représente 30% du coût moyen (en figure 28). C'est également l'un des postes les plus privilégiés par les clients pour offrir une meilleure couverture santé à leurs salariés.

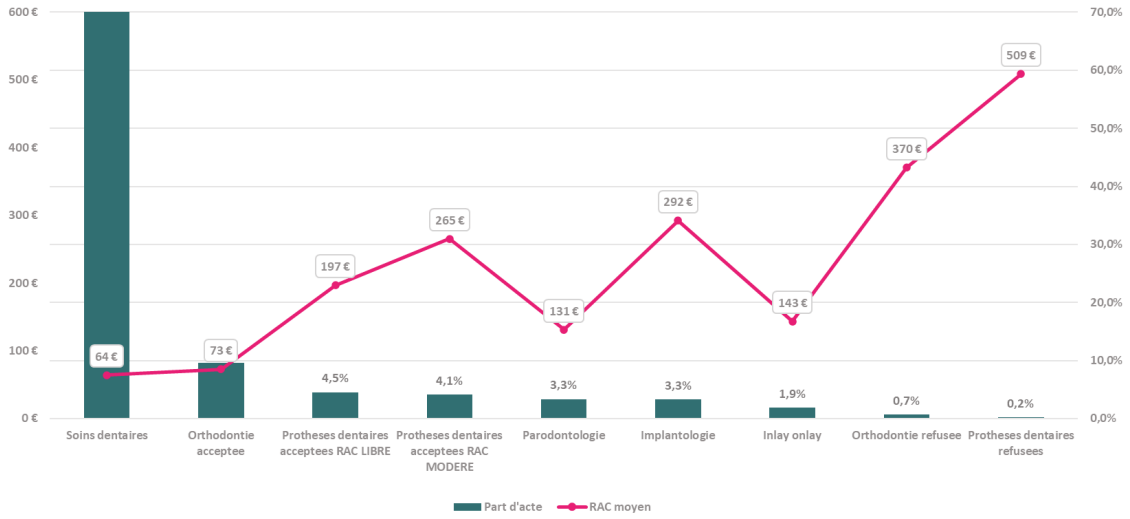


FIGURE 30 – Reste à charge moyen par acte en Dentaire

Les soins d'orthodontie et les prothèses dentaires qui ne sont pas couverts par la Sécurité sociale peuvent entraîner des frais considérables pour le patient, mais ils sont rares et ne sont donc pas pertinents pour notre étude.

Les actes restants seront introduits dans la base d'étude puisque les soins dentaires sont plus ou moins fréquents, et ils présentent des restes à charge moyens non négligeables.

Le poste Optique

Le poste optique inclut les dépenses liées aux lunettes, aux lentilles et à la chirurgie des yeux. Les lunettes sont subdivisées en plusieurs types de remboursements, tels que la monture, les verres simples, les verres complexes et les verres hypercomplexes. En outre, les actes varient selon l'âge, avec des prestations destinées aux enfants et d'autres réservées aux adultes.

Toutefois, pour évaluer la qualité des garanties relatives à ce poste, nous avons choisi arbitrairement de nous concentrer uniquement sur les prestations concernant les adultes. Ce choix est fait de façon arbitraire afin de limiter la base de calcul et de se restreindre aux actes les plus courants et plus privilégiés par les clients.

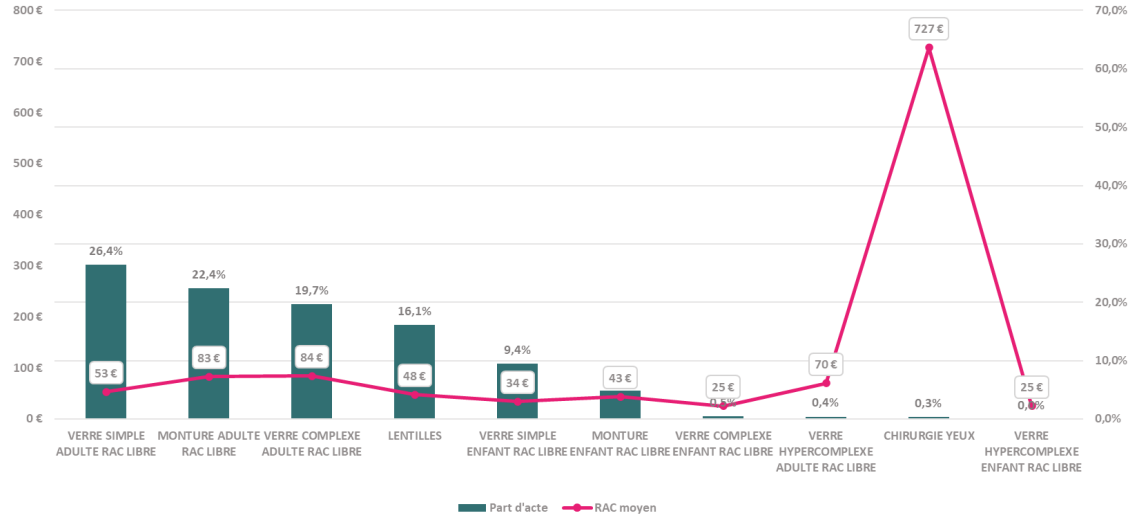


FIGURE 31 – Reste à charge moyen par acte en Optique

Ainsi, en optique, nous étudierons unique les performances sur les actes suivants :

— La lunetterie adulte restreinte à la monture du panier libre, au verre simple et complexe du panier libre

— Les lentilles sont beaucoup consommées avec un reste à charge non négligeable.

— la chirurgie de l'œil est une prestation très peu fréquente, mais qui génère un reste à charge très élevé.

Les verres hypercomplexes sont très peu consommés que soient chez les enfants ou les adultes.

Le poste Soins courants

Bien que les soins courants tels que les consultations chez un médecin généraliste, les examens médicaux, les médicaments prescrits et les actes de radiologie génèrent un reste à charge moindre pour l'assuré, leurs fréquences de consommation en font un élément clé dans l'évaluation de la performance d'une couverture santé.

En effet, les soins courants sont les prestations de santé les plus consommées par les assurés et le reste à charge cumulé peut représenter une somme significative sur une année. Par conséquent, une couverture santé performante devrait prendre en compte les coûts liés à ces prestations et proposer des garanties adaptées à la fréquence de consommation de ces soins.

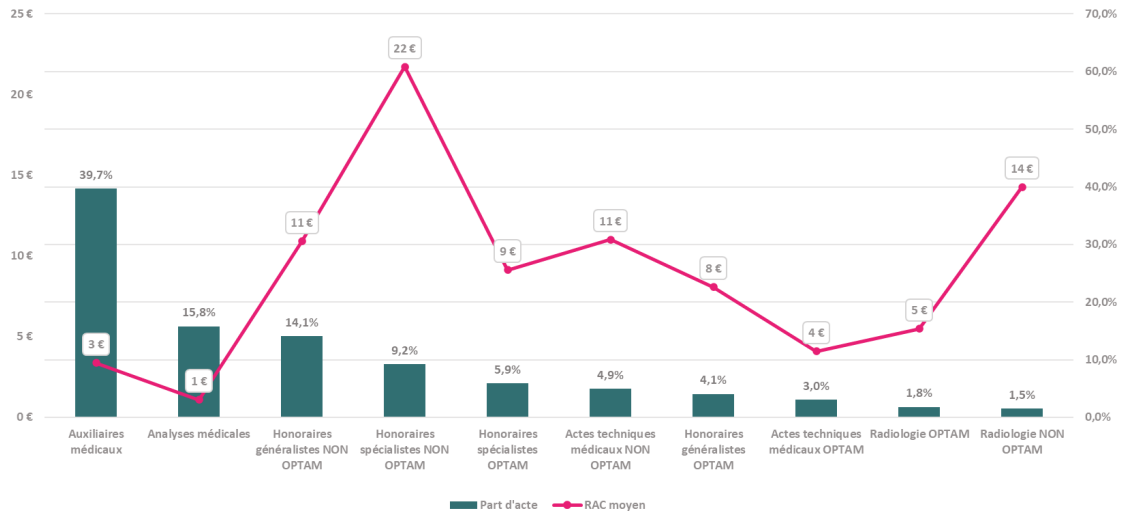


FIGURE 32 – Reste à charge moyen par acte en Soins courants

C'est pourquoi, l'ensemble de ces soins sera intégré dans l'étude.

Le poste Autres

Concernant le poste "autres", nous devons d'abord nous concentrer sur les actes de santé les plus consommés par les assurés. Nous avons identifié trois actes principaux :

— La médecine douce : elle inclut des actes tels que l'ostéopathie, la chiropraxie, la consultation chez un diététicien ou un pédicure-podologue. Bien que ces actes ne soient pas remboursés par la Sécurité sociale, ils sont souvent plébiscités par les patients en raison de leur caractère préventif.

— L'appareillage : il comprend des actes tels que les orthèses et les appareils respiratoires. Ces prestations sont souvent nécessaires pour améliorer la qualité de vie des patients souffrant de troubles respiratoires ou de problèmes de mobilité. Bien que certains de ces actes soient remboursés par la Sécurité sociale, le reste à charge peut être important.

— Les prothèses auditives : bien que les prothèses auditives soient remboursées par la Sécurité sociale, le reste à charge peut être élevé. Les coûts liés à l'achat et à l'entretien de ces appareils peuvent également être significatifs.

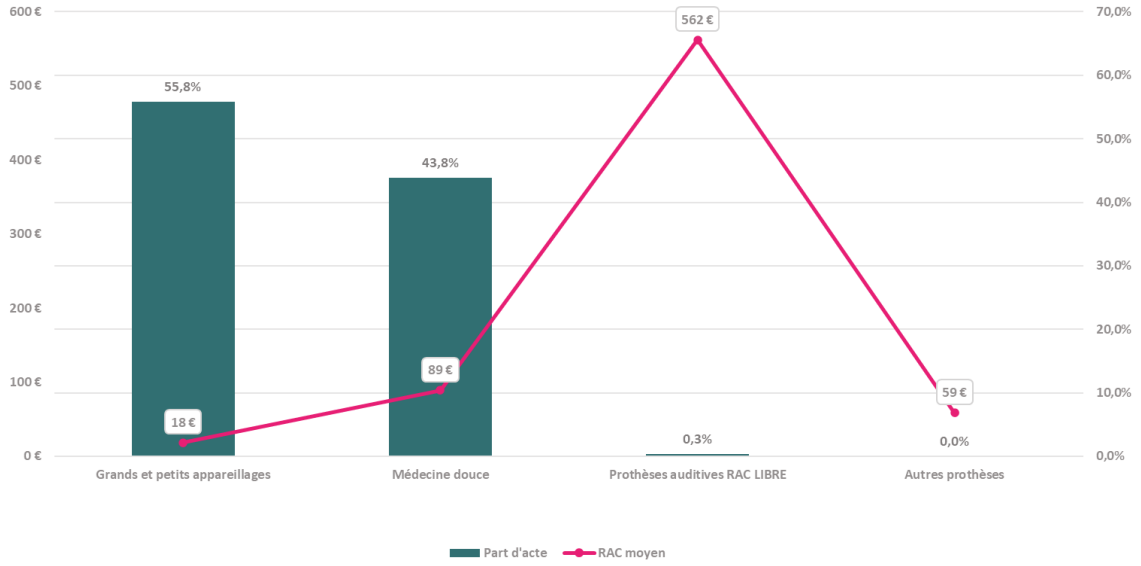


FIGURE 33 – Reste à charge moyen par acte en Autres

2.4 Le risque à modéliser

Le but principal de ce mémoire est de modéliser le risque géographique en matière de santé. Nous partons du postulat que la zone géographique dans laquelle vit une personne a une incidence directe sur les coûts de ses dépenses de santé, soit en raison des tarifs imposés par l'offre et la demande, soit en fonction de la qualité de vie de la zone. Bien que l'État intervienne pour réglementer certains niveaux de dépassement, les acteurs de soins de santé ne sont pas obligés d'y adhérer. Ainsi, les niveaux de dépenses ne sont pas définis uniquement par les caractéristiques individuelles de chaque assuré.

Pour mesurer l'effet réel du risque lié à la zone géographique, il est essentiel de disposer d'une connaissance aussi large que possible de chaque département. En assurance santé, la situation géographique est l'un des principaux éléments de tarification, mais également de mesure de la performance d'un régime complémentaire de santé. Il est donc nécessaire d'introduire la notion de zone géographique. En effet, la comparaison du niveau de garantie et de la localisation est essentielle pour évaluer et comparer les couvertures de santé.

Pour être plus concret, prenons l'exemple de la couverture des honoraires hospitaliers suivants.

Fixons la base de remboursement pour les honoraires hospitaliers à 30€ et considérons deux individus ayant une situation géographique différente. l'individu A réside à Paris tandis que l'individu B habite Angers.

Base de remboursement (BR) pour des honoraires hospitalier: 30€

	Individu A	Individu B
Localisation	Paris	Angers
Montant des honoraires	95 €	55 €
Couverture y compris la part Sécurité Sociale	250% BR	200% BR
Maximum pris en charge par la couverture	250% x 30€ = 75€	200% x 30€ = 60€
Restant à charge	20 €	0 €

Bien que l'individu A dispose d'une garantie supérieure à celle de l'individu B, la présence d'un restant à charge plus élevé pour l'individu A peut réduire l'efficacité globale de sa couverture.

3 Présentation de la base de modélisation

La construction d'une base de données fiable est une étape cruciale dans la modélisation. La qualité et la précision de cette base de données peuvent grandement influencer les résultats de l'étude. Par conséquent, il est important de s'assurer de la qualité des données avant de procéder à toute analyse ou modélisation. L'objectif de cette section sera de présenter nos données de modélisation.

3.1 Le périmètre d'étude

Le risque santé est un risque court, chaque contrat d'assurance complémentaire santé est établi pour une année de survenance. De plus, ce marché est régulièrement soumis à des réglementations et réformes qui peuvent faire évoluer le risque dans le temps. Ce changement dans le temps peut également être engendré par une évolution du portefeuille qui entraîne une évolution de la population assurée. Autrement dit, le portefeuille des premières années d'exercice d'une activité d'assurance ne peut être comparé ou étudié de la même manière cinq ans plus tard. Ainsi, le choix du périmètre d'étude est nécessaire dans le cadre d'une modélisation afin de limiter un biais dû à des évolutions du portefeuille ou du risque dans le temps.

Nous avons restreint notre étude à une base de données établie sur le risque santé regroupant l'ensemble des prestations réalisées sur le portefeuille des gestionnaires direct du Crédit Agricole Assurances durant les années de survenance 2019 et 2021 arrêtés à fin mars 2022. Nous avons pris deux années de survenance afin d'avoir une base d'étude assez conséquente, et nous avons arbitrairement fait le choix d'écart les données de l'exercice 2020 en raison du contexte particulier de la pandémie de la covid.

Ainsi, le périmètre concerne l'ensemble des prestations santé de tous les bénéficiaires des contrats standards de branche et des contrats sur-mesures gérés par les gestionnaires directs de Crédit Agricole Assurance, observés en 2019 et 2021.

3.2 Construction de la base

3.2.1 Données internes

Pour notre étude, nous sommes partis de trois bases de données :

— la base de prestations répertoriant tous les actes de soins et de biens médicaux des survenances 2019 et 2021 ;

— la base des effectifs dans laquelle nous retrouvons les informations concernant les bénéficiaires de contrats en portefeuille ;

— la base des entreprises ayant souscrit à une couverture santé complémentaire à destination de leurs salariés.

Certaines de ces bases ont nécessité un prétraitement avant la fusion.

Le prétraitement des données de prestations

La base de prestations répertorie tous les soins et biens médicaux perçus par les bénéficiaires du périmètre considéré. Celle-ci contient des régulations, ces prestations négatives ne sont pas prises en compte et n'ont pas d'intérêt étant donné que nous souhaitons modéliser les coûts moyens réels des prestations.

De plus, dans cette même table, une prestation est caractérisée par un libellé d'acte pouvant regrouper plusieurs prestations différentes. Toutefois, l'outil d'aide au conseil que nous développons manipule des prestations aux libellés d'actes plus fins, nous avons donc procédé à la construction d'une table de correspondance afin que chaque garantie traitée dans notre étude puisse être associée à l'ensemble des actes correspondants à cette garantie de notre base d'étude et nous avons gardé uniquement les actes d'intérêt sélectionnés dans la partie précédente que nous retrouvons en figure 23.

Chaque ligne de cette table représente les caractéristiques des prestations d'un acte particulier perçu par un bénéficiaire un jour donné. Dans ce cas, une ligne peut représenter plusieurs prestations du même type réalisées par un même bénéficiaire la même journée. Les caractéristiques d'intérêt de notre étude sont les suivantes :

- nb_actes correspond au nombre de prestations perçues le même par le même bénéficiaire ;
- $mt\_fr\_unitaire$ est le rapport entre les frais réels des prestations mt_fr et nb_actes, ce qui représente le coût moyen de l'acte réalisé ;
- lib_actes_final correspond au libellé de l'acte réalisé ;
- lib_poste_final correspond au poste auquel appartient l'acte réalisé ;
- $annee\_surv$ correspond à l'année durant laquelle l'acte a été réalisé ;
- $type\_benef$ est le type de bénéficiaire, il peut s'agir de l'adhérent (le salarié/ assuré) ou de l'un de ses ayants droit (conjoint ou enfant) ;
- rg_benef le rang bénéficiaire est un chiffre allant de 1 au nombre d'enfants + 1 que possède l'adhérent, il permet de différencier les bénéficiaires en particulier lorsqu'il y a plusieurs enfants ;
- $id\_entreprise$ correspond à l'identifiant de l'entreprise
- id_assure correspond à l'identifiant de l'adhérent et de ses ayants droit

Afin de pouvoir fusionner avec les autres tables, nous avons constitué une clé unique pouvant relier des prestations à son bénéficiaire. cette clé est composée de la concaténation des variables suivantes : année de survenance, l'identifiant de l'assuré, l'identifiant de l'entreprise et enfin le rang du bénéficiaire.

À présent, nous souhaitons fusionner la base de prestation avec la base des effectifs afin de récupérer les informations concernant les bénéficiaires. Toutefois, cette base de données ne peut pas être exploitée telle quelle, elle nécessite une étape de prétraitement.

Le prétraitement des effectifs bénéficiaires

La base de données des effectifs répertorie les informations concernant l'ensemble des bénéficiaires ayant été couverts au moins un mois sur une année de survenance. En effet, chaque ligne indique la couverture du bénéficiaire avec le mois de l'année indiqué en base.

Ainsi, il faut agréger la base de telle sorte que chaque ligne correspond à la présence du bénéficiaire en question à l'année de survenance indiquée. Nous devons introduire la notion d'exposition qui est une valeur entre 0 et 1 représentant la période de couverture sur une année d'exercice.

Les éléments d'informations concernant les bénéficiaires récupérés sont les suivants :

— l'âge du bénéficiaire ;
— le sexe du bénéficiaire ;
— le type de bénéficiaire (l'adhérent, le conjoint ou l'enfant)
— l'exposition ;
— la catégorie socioprofessionnel de l'adhérent auquel le bénéficiaire est rattaché ;
— le département résidence du bénéficiaire.

Certains bénéficiaires n'ont pas le département de résidence indiqué, ils seront rattachés au département de résidence de l'adhérent. Nous avons pris une hypothèse plus conséquente, les adhérents dont le département de résidence n'est pas indiqué seront associés au département de l'entreprise par lequel ils sont couverts.
La fusion des deux tables constitue la base des données internes. et afin de fiabiliser cette base, nous avons comparé la réparation des prestations avec celle donnée lors de l'inventaire.

3.2.2 Données externes

Comme mentionné précédemment, l'objectif de ce mémoire est de modéliser le risque géographique en utilisant les coûts moyens des prestations de santé observées dans différents départements.

Cependant, les données de portefeuille disponibles pour cette étude ne sont pas suffisamment exhaustives et ne contiennent pas toutes les informations nécessaires pour caractériser les départements ou les régions françaises. Par conséquent, nous avons décidé d'inclure des variables externes dans notre base de modélisation pour compléter ces informations.

La quête des variables externes repose sur les résultats issus de diverses publications d'études gouvernementales portant sur les dépenses de santé, notamment sur *Les dépenses de santé en France : déterminants et impact du vieillissement à l'horizon 2050, ministère de Santé*, ou encore *Projection des dépenses de santé à l'horizon 2060,le modèle*

PROMEDE, ministère de Santé. Ces études visent à anticiper la croissance des dépenses de santé globales en corrélation avec l'évolution de divers indicateurs socio-économiques reflétant les projections démographiques ou le niveau de richesse d'un pays.

Les variables démographiques

- La densité de population au km^2 ;
- Les effectifs de la population par tranche d'âge et par sexe ;
- Les espérances de vie par sexe ;
- L'indice de vieillissements (le nombre de séniors pour 100 jeunes).

Les variables sociales

- Le salaire médian par foyer ;
- Le taux de chômage ;
- Le taux de pauvreté ;
- Les salaires moyens par sexe, par catégorie socioprofessionnelle (cadre, intermédiaire, employer et ouvrier) ;
- Le taux d'actifs.

Les variables des professions médicales

- Le nombre de médecins généralistes ;
- Le nombre de médecins Spécialiste : cardiologie, dermatologie, gynécologie, gastro-entérologie, psychiatrie, et autres ;
- Le nombre de spécialistes en ophtalmologie ;
- Le nombre de chirurgiens dentistes ;
- Le nombre de sagefemmes, infirmiers ;
- Le nombre d'auxiliaires médicaux.

Les variables des établissements médicaux

- Le nombre d'établissements hospitaliers ;
- Le nombre d'établissements psychiatriques ;
- Le nombre d'urgences ;
- Le nombre de maternités ;
- Le nombre de maisons de santé pluridisciplinaire ;
- Le nombre de pharmacies ;
- Le nombre de laboratoires d'analyses médicales.

Ces données proviennent de différentes sources.

Traitement des variables externes

Les variables externes sont nombreuses et corrélées entre elles, afin de réduire le nombre tout en conservant les informations fournies par ces dernières et également d'annuler les dépendances, nous allons procéder à une Analyse en Composantes Principales (ACP).
Cette méthode permet de réduire la dimension d'un ensemble de données correspondant à un grand nombre de variables quantitatives corrélées entre elles, tout en conservant un maximum de l'information présente dans les données.
Cet outil d'analyse de données va permettre de représenter les variables externes dans un sous-espace en explicitant les liaisons initiales entre ces variables, tout en réduisant le nombre de variables.
L'Analyse en Composantes Principales à partir des métriques sélectionnées, fournit un nouveau jeu de variables, les composantes principales, décorrélées entre elles et ordonnées de manière à ce que les premières contiennent la plus grande partie de l'information contenue dans l'ensemble des variables initiales.
Les composantes principales contiennent les coordonnées des observations sur les axes factoriels qui correspondent aux directions de l'espace qui expliquent au mieux la variance de l'échantillon. Ces axes factoriels sont les vecteurs propres de la matrice de covariance des variables. La valeur propre de chaque axe factoriel représente le pourcentage de variance expliquée par l'axe.

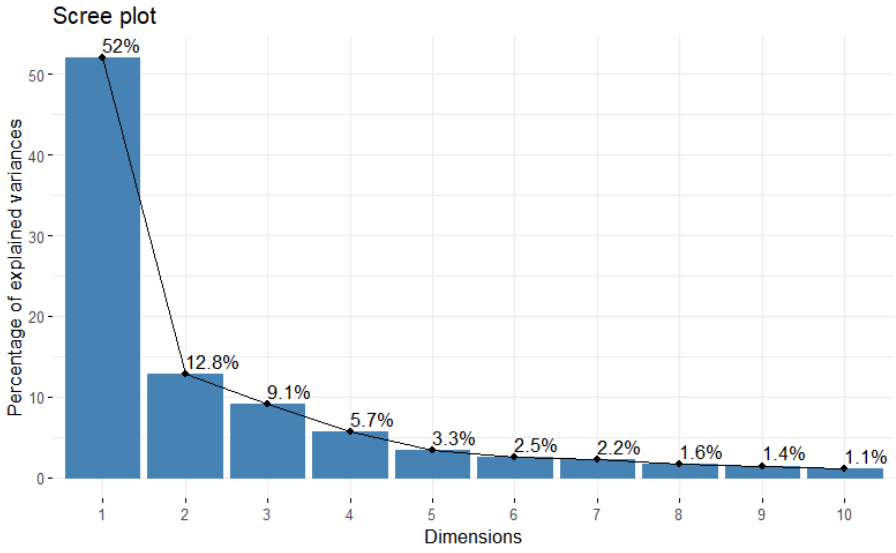


FIGURE 34 – Nouvelles variables de l'ACP

Nous constatons que les trois premières variables suffisent pour résumer 75% des variables externes. Nous les avons intégrés parmi les variables explicatives pour estimer le coût moyen.

3.3 Statistiques descriptives

La consommation médicale d'un individu en matière de santé dépend de plusieurs facteurs, ce qui rend indispensable l'intégration de ces caractéristiques lors de la tarification d'une assurance santé individuelle ou collective pour la population à couvrir. Les principaux facteurs à considérer sont l'âge, le sexe, la catégorie socioprofessionnelle et le secteur d'activité. Examinons à présent les interdépendances entre certains de ces facteurs.

3.3.1 Étude du coût moyen par rapport aux variables internes

La consommation médicale varie significativement entre les sexes, avec les femmes ayant tendance à consommer plus de soins que les hommes. Cependant, depuis 2012, la tarification basée sur le sexe est considérée comme discriminatoire par la Cour de justice de l'Union européenne. Les assureurs ne sont donc plus autorisés à différencier les tarifs en fonction du sexe dans certains domaines de l'assurance, tels que l'assurance vie ou l'assurance automobile.

Malgré cela, le sexe reste un facteur pris en compte dans la tarification de la complémentaire santé collective. La proportion de femmes dans la population assurée est incluse dans la tarification.

La figure 35 ci-dessous illustre clairement que les femmes ont une fréquence de consommation de soins plus élevée que les hommes, ce qui est vrai pour l'ensemble des postes de soins. Cette surconsommation souligne l'importance de prendre en compte le sexe dans la tarification d'une couverture santé.

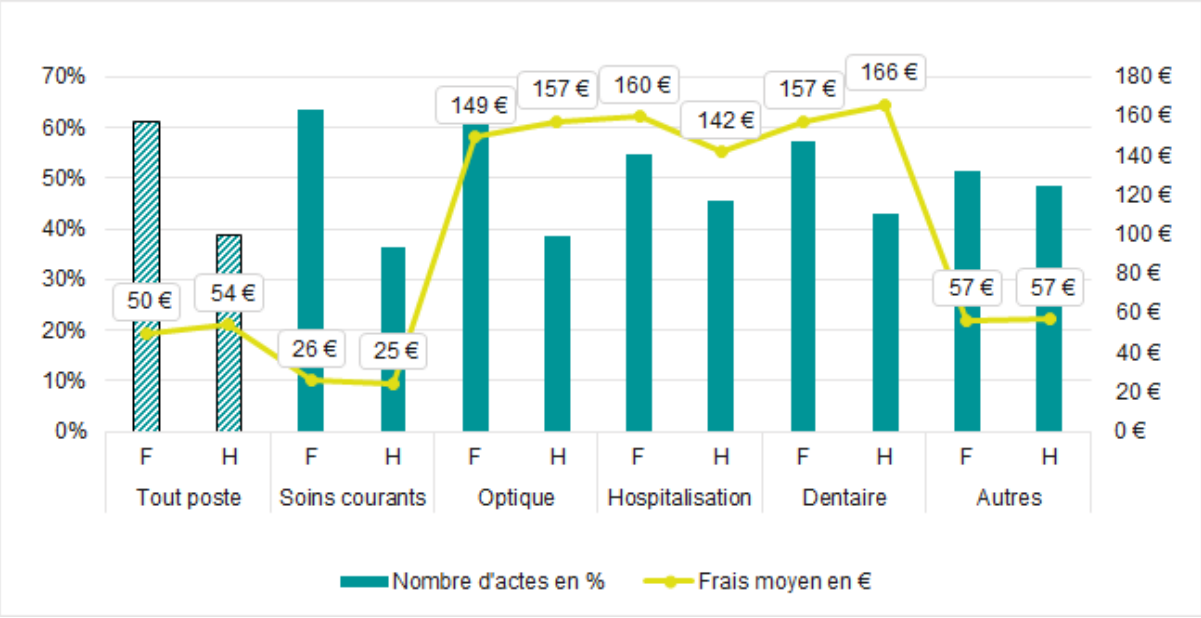


FIGURE 35 – La consommation de soins et de biens médicaux en fonction du sexe

Néanmoins, en ce qui concerne le coût moyen des prestations, la tendance s'inverse, les hommes ayant des frais de soins moyens plus élevés avec un coût moyen de 54€ contre 50€ pour les femmes.

Cette tendance est généralement observée sur la plupart des postes de soins, sauf en ce qui concerne l'hospitalisation, où la tendance s'inverse. Cette inversion peut s'expliquer par le fait que certains actes médicaux spécifiques aux femmes, tels que les accouchements ou les interventions gynécologiques, peuvent s'avérer très coûteux et peuvent considérablement augmenter les frais de soins pour les femmes.

Lors du choix d'une assurance complémentaire santé pour l'ensemble des employés, une personne morale a la possibilité d'opter pour plusieurs régimes, un régime par catégorie socioprofessionnelle.

— Un régime cadre (C) est une couverture santé couvrant uniquement les salariés cadres.

— Un régime non cadre (NC) est une couverture santé couvrant uniquement les salariés non cadres.

— Un régime ensemble du personnel (EP) est une couverture santé couvrant l'ensemble des salariés sous la même couverture santé.

La catégorie socioprofessionnelle est un caractère important dans la consommation de biens médicaux, en effet le profil de consommation de l'une des catégories est clairement identifiable dans la figure 36.

Tout d'abords, précisons que la majorité des contrats détenus en portefeuille sont du régime ensemble du personnel, c'est pourquoi cette catégorie est clairement majoritaire en consommation sur l'ensemble des postes de soins.

En outre, lorsqu'il y a des distinctions des régimes selon la CSP, très souvent celui des cadres est privilégié en termes de garanties. En effet, ils ont la plupart du temps une meilleure couverture, notamment sur les postes ayant le plus de reste à charge, c'est-à-dire le dentaire et l'optique.

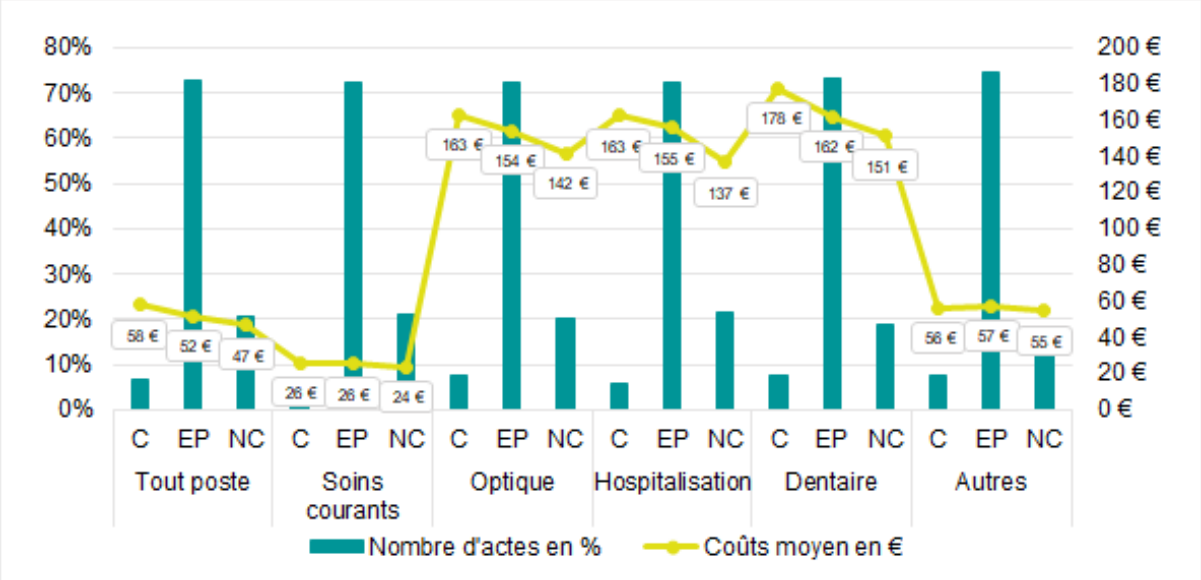


FIGURE 36 – La consommation de soins et de biens médicaux en fonction de la CSP

Ce dernier privilège souvent accordé au régime cadre engendre une consommation plus importe autant dans la fréquence qu'en termes de coût, en particulier sur les postes de soins les plus coûteux tel que l'optique ou encore le dentaire.

À présent, voyons comment la consommation de soins et de biens médicaux se comporte en fonction de l'âge. En assurance santé collective, la grande majorité des assurés sont des personnes actives, en effet, l'adhésion des ayants droit n'est obligatoire que selon la structure de cotisation que nous avons abordée dans la section 2.2.3. De plus, les anciens actifs faisant partie d'une structure d'assurance collective d'une entreprise au titre de retraité sont regroupés et sont souvent sujets à des majorations régulières indépendamment du régime des actifs, ce qui les poussent à passer vers une couverture individuelle plus avantageuse en termes de coût. Ainsi, nous avons une présence moindre sur la tranche d'âge des retraités.

La figure 37 soutient la présence majoritaire d'actifs de 23 ans à 61 ans en portefeuille, ainsi qu'une part moins importante sur la tranche d'âges des enfants. En ce qui concerne le coût des frais de soins, il y a une tendance de coût important chez les adolescents qui s'explique par une consommation importante de soins d'orthodontie, un acte de soins dentaire couteux.

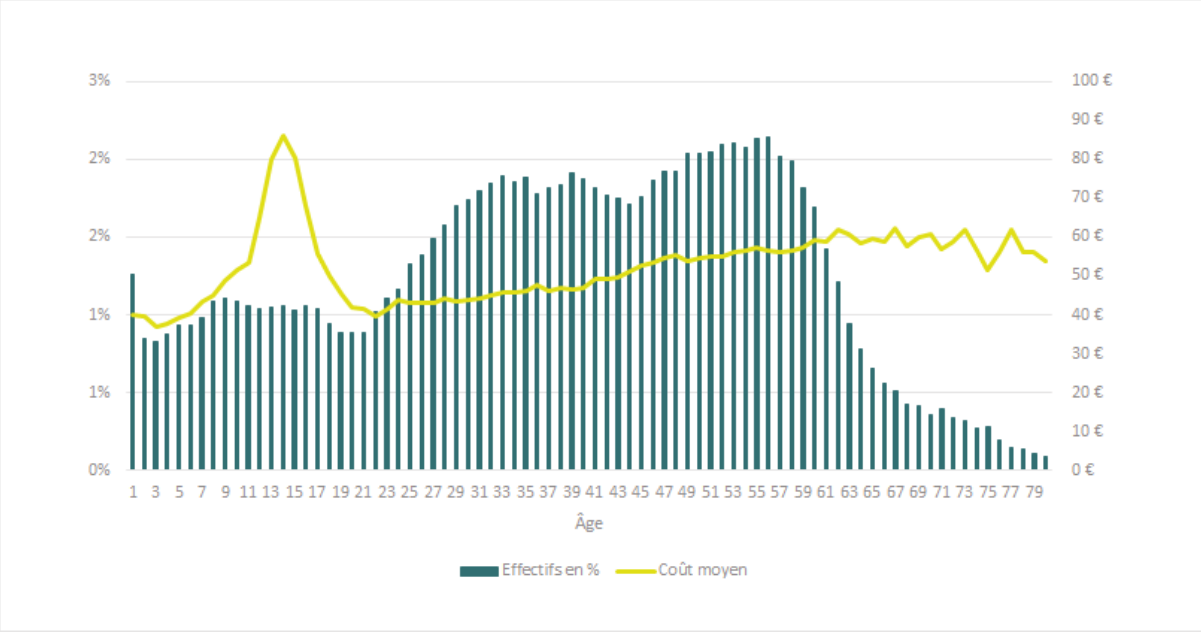


FIGURE 37 – La consommation de soins et de biens médicaux en fonction de l'âge

Le pic apparent sur la tranche d'âge de 9 ans à 21 ans en figure 37 confirme bien cette tendance de surcout chez les adolescents et nous constatons que le coût augmente globalement avec l'âge. Néanmoins, la tendance sur la tranche d'âge 65 ans et plus n'est pas stable par la part faible de cette population en portefeuille.

La figure suivante représente les coûts par les catégories d'activités suivantes :

— Les actifs sont les adhérents encore en activité dans l'entreprise proposant la couverture ainsi que leurs ayants droits (conjoint et enfants à charge).

— Les périphériques sont les adhérents désignés dans l'article 4 de la loi Evin *« L'assureur doit prévoir les modalités et les conditions tarifaires du maintien des garanties santé au profit d'anciens salariés (retraités, bénéficiaires d'une rente d'incapacité ou d'invalidité, licenciés) et des ayants-droit d'assurés décédés pendant une durée minimale de douze mois à compter du décès, qui en font la demande dans les 6 mois qui suivent la rupture du contrat de travail ou la date du décès de l'assuré, ou l'expiration de la couverture de la portabilité »*
Cette catégorie désigne essentiellement les retraités ainsi que leurs ayants droits qui sont la plupart du temps leur conjoint.

— Les personnes en portabilité désignent *« Tout salarié perdant son emploi hors faute lourde, qui bénéficie d'une couverture complémentaire de santé et de prévoyance (décès, incapacité, invalidité) obligatoire ou facultative (option, ...) au sein de son ancienne entreprise, peut continuer à en bénéficier pour une durée égale à celle du ou des derniers contrats de travail consécutifs, exprimée en mois entier, dans la limite de 12 mois. »*

Les frais de soins pour les périphériques sont en moyenne plus importants sur pratiquement l'ensemble des postes de soins

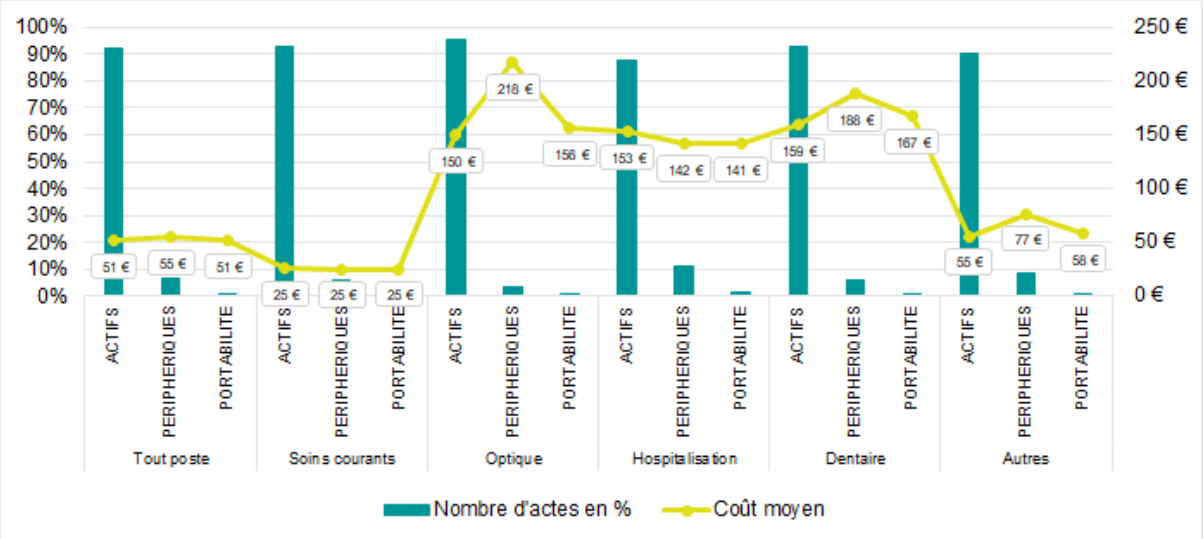


FIGURE 38 – La consommation de soins et de biens médicaux en fonction de l'activité

Dans le domaine de la santé collective, l'adhésion des ayants droit n'est pas toujours exigée, notamment en ce qui concerne le conjoint de l'adhérent. Toutefois, il est observé que les conjoints adhérents ont tendance à consommer davantage de soins, entraînant ainsi un coût moyen plus élevé pour l'ensemble des postes de dépenses.

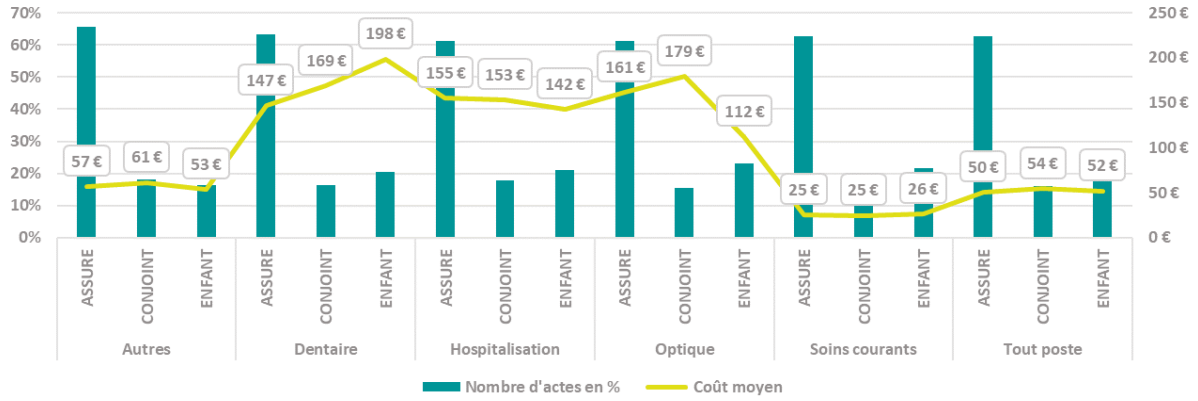


FIGURE 39 – La consommation de soins et de biens médicaux en fonction du type de bénéficiaire

Suite à ces constatations, nos variables internes explicatives se résument à :

— l'âge moyen ;
— la proportion de femme représentera le sexe ;
— les effectifs des différentes catégories socioprofessionnelles ;
— la proportion de retraités ;
— les proportions de chaque bénéficiaire.

3.3.2 Étude du coût moyen par rapport aux variables externes

Rappelons que les variables externes décrivent par département la démographie de la population, des indicateurs socio-économiques, les professions médicales ainsi que les établissements médicaux. Ils sont en nombre de 35, certaines sont potentiellement corrélées entre elles, c'est les raisons pour lesquelles nous les avons résumées à travers trois nouvelles variables données par l'ACP qui expliquent près de 75% de la variance des 35 données externes.

Montrons qu'il existe une relation entre ces dernières et notre variable cible avec la matrice de corrélation suivante.

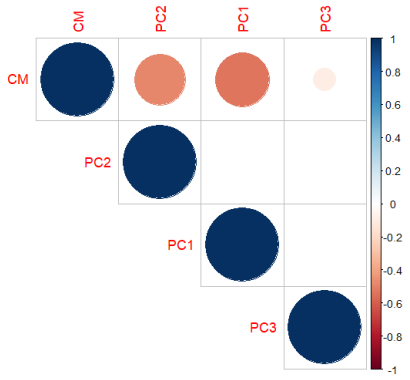


FIGURE 40 – La matrice de corrélation du coût moyen et des variables externes

3.3.3 Étude par département

Dans certains départements, les adhérents à un régime santé collective de Crédit Agricole Assurances ne sont pas assez nombreux pour différentes raisons que nous n'aborderons pas ici. Il se trouve que dans ces zones géographiques que la population et la consommation y sont particulières. En effet, l'âge moyen des effectifs consommant est entre 30 et 35 ans, il y a 1 département ou l'âge moyen est plus bas et 10 départements ou l'âge moyen est plus élevé.

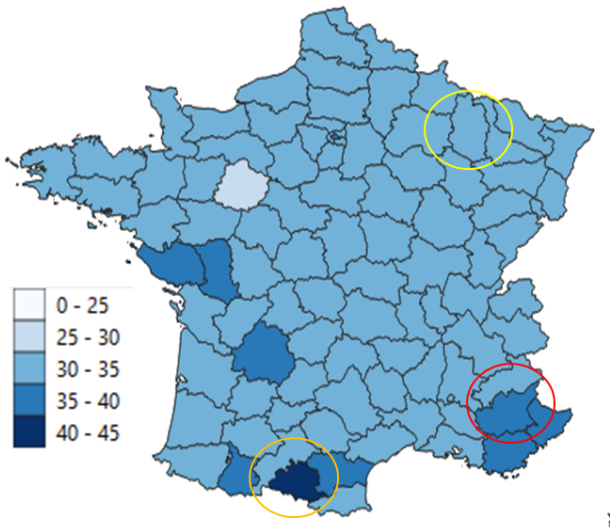


FIGURE 41 – Âge moyen par département

Dans l'Ariège (encercle en orange) et les Alpes de haute Provence (encercle en rouge), les écarts d'âge moyen avec la majorité de départements sont essentiellement dus aux faibles effectifs.

L'effectif moindre dans certaines zones impacte également le coût moyen, nous pouvons l'observer avec le département de la Meuse (encercle en jaune) où le coût fait partie des 20% les plus élevés alors qu'il fait partie des départements où il y a le moins d'effectifs.

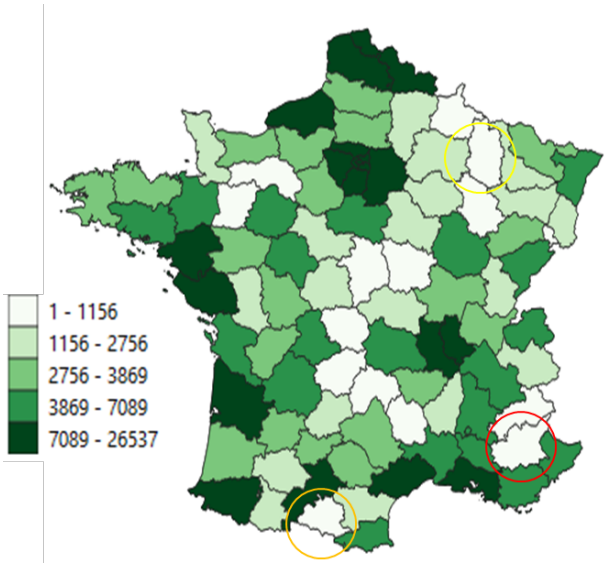


FIGURE 42 – Effectifs consommant par département

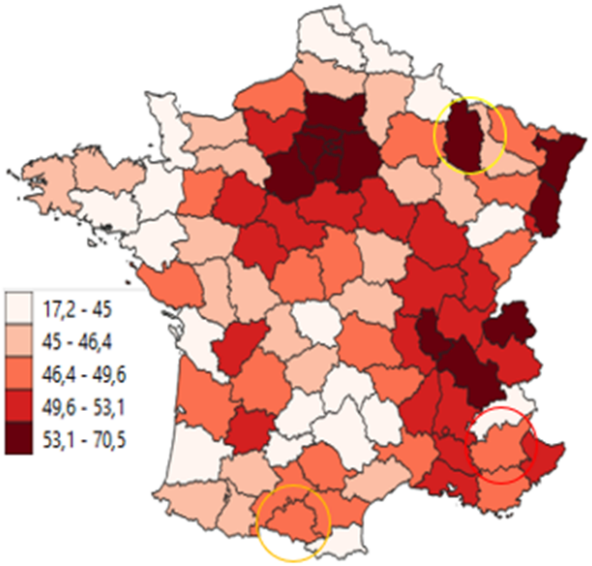


FIGURE 43 – Coût moyen par département

Das suite de ce mémoire, nous appellerons effet d'effectif, l'ensemble des inconvénients causés des faibles effectifs.

4 Modélisation du risque géographique

L'impact géographique sur la consommation médicale est un sujet complexe, tant sur le plan politique qu'actuariel. En assurance santé, la situation géographique constitue un des principaux éléments dans la tarification, mais également dans la mesure de performance d'un régime de complémentaire santé, il est donc nécessaire d'introduire la notion de zone géographique.En effet, la confrontation du niveau de garantie et de la localisation est essentiel pour évaluer et comparer des couvertures de santé.

L'objectif de ce chapitre est de présenter les aspects théoriques des différentes méthodes abordées dans ce mémoire pour modéliser le risque géographique pour l'étude de performance. Nous avons décidé de confronter trois algorithmes de régression dans l'élaboration d'un zonier.

4.1 Les modèles Linéaires Généralisés

4.1.1 Régression linéaire classique

La régression linéaire également appelée modèle linéaire est un des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle.

Nous cherchons donc à prédire une variable Y quantitative à partir d'une combinaison linéaire d'une ou de plusieurs variables explicatives $X_1, ..., X_p$ quantitatives. Pour ce faire, nous disposons d'observations de ces variables sur n individus, c'est-à-dire d'un tableau de données de la forme :

$$\begin{array}{c|cccc} y_1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ y_2 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{array}$$

Dans le modèle de régression linéaire, la quantité à expliquer Y s'écrit alors la façon suivante :

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

avec $Y = (Y_1, ..., Y_n) \in \mathbb{R}^n$ les valeurs à prédire, $X = (\mathbb{1}, X_1, ..., X_p) \in \mathbf{R^{n*p+1}}$ les données de prédiction, les paramètres $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbf{R^{p+1}}$ sont des coefficients réels inconnus à estimer et enfin $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ une variable quantitative de valeur moyenne nulle qui représente une somme d'erreurs aléatoires résumant l'information manquante dans l'explication linéaire des valeurs y_i par les $x_{i,1}, ..., x_{i,p}$ pour $i \in 1, ..., n$.

Dans ce contexte, l'objectif principal est d'estimer convenablement les coefficients $\beta = (\beta_1, ..., \beta_p)$ par $\hat{\beta}$ à l'aide des données. Entre autres, cela permet de mesurer l'importance des variables $X_1, ..., Xp$ dans l'explication de Y. Le meilleur estimateur $\hat{\beta}$ minimise l'erreur

de prédiction ε s'exprimant comme la distance entre Y et la projection orthogonale de Y dans l'espace donnée par $X_1, ..., X_p$.

$$\varepsilon = ||Y - \hat{\beta}X|| = \arg \min_{\beta \in \mathbf{R^{p+1}}} ||Y - \beta X||$$

Nous l'appelons alors l'estimateur des moindres carrés, il suppose que X est de rang plein, autrement dit que les variables explicatives sont indépendantes et sont données par la formule suivante :

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Ce modèle restreint énormément la forme de la variable à expliquer.

4.1.2 Régression linéaire généralisée

Les modèles linéaires généralisés sont des généralisations de la régression linéaire, en effet, ils introduisent une fonction lien qui définit la relation entre la combinaison linéaire des variables $X_1, ..., X_p$ et la variable Y que nous retrouvons dans la section précédente. Dans le cas présent, nous souhaiterions prédire une quantité continue, nous parlons donc de régression linéaire généralisée, celle-ci se décompose en trois composantes :

- Une composante aléatoire : la loi de distribution de la variable explicative Y ;
- Une composante déterministe : l'ensemble des variables explicatives ;
- La fonction lien : une fonction monotone et dérivable qui relie la composante aléatoire et la composante déterministe.

La composante aléatoire

La première composante constitue la variable Y à expliquer qui est considérée comme aléatoire et à laquelle nous associons une loi de probabilité appartenant à la famille exponentielle. La variable Y appartient à la famille des lois exponentielles si et seulement si sa fonction densité est de la forme suivante :
Pour $y \in \mathbb{R}$

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

avec

- θ le paramètre naturel de la fonction exponentielle ;
- ϕ le paramètre de dispersion de la fonction exponentielle ;
- $b(.)$ une fonction de classe C^3 et de dérivée première inversible ;
- $a(.)$ et $c(.)$ deux fonctions dérivables ;

L'écriture de l'espérance et de la variance de Y appartenant à la famille des lois exponentielles implique deux paramètres : le paramètre naturel et le paramètre de dispersion.

$$\mathbb{E}(Y) = b'(\theta) = \mu$$

$$\mathbb{V}(Y) = b''(\theta)a(\phi) = \sigma^2$$

Ainsi, le paramètre θ influence l'espérance de la distribution de Y mais également sur sa variance, d'où l'appellation paramètre de dispersion.

La plupart des lois usuelles s'écrivent comme une loi exponentielle, notamment la loi gaussienne, gamme, poisson ou encore la loi binomiale.

Distribution de $Y = y$	$b(\theta)$	ϕ	$a(\phi)$
Gaussienne $\mathcal{N}(\mu; \sigma^2)$	$\frac{\theta^2}{2}$	σ^2	ϕ
Gamma $\Gamma(\alpha, \beta)$	$\beta log(\beta y) - log(y) - log(\Gamma(\beta))$	$\frac{-1}{\alpha}$	$-log(-\theta)$
Poisson $\mathcal{P}(\lambda)$	$exp(\theta)$	1	ϕ
Binomiale $\mathcal{B}(n, p)$	$log(1 + exp(\theta))$	$\frac{1}{p}$	ϕ

La composante déterministe

Les variables explicatives à intégrer dans le modèle sont à choisir avec soin. En effet, le nombre de variables utilisé ne doit pas être trop important pour que le modèle soit utilisable en pratique, mais il doit être suffisant pour que le modèle, soit cohérent et performant. C'est pourquoi, il est nécessaire de sélectionner, parmi toutes les variables à disposition, les variables dont le pouvoir explicatif est le plus important afin de réussir à obtenir un juste équilibre.

La fonction de lien

La troisième et dernière composante d'un modèle linéaire généralisé est la fonction de lien $g(.)$ monotone et dérivable qui décrit la relation entre la composante aléatoire et la composante déterministe. Cette fonction permet plus particulièrement d'établir un lien entre l'espérance de la variable cible et une combinaison linéaire des variables explicatives.

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

Distribution de Y	Fonction Lien $g(.)$
Normal $\mathcal{N}(\mu; \sigma^2)$	$g(\mu) = \mu$
Gamma $\Gamma(\alpha, \beta)$	$g(\mu) = \frac{1}{\mu}$
Poisson $\mathcal{P}(\lambda)$	$g(\mu) = log(\mu)$
Binomiale	$g(\mu) = log(\frac{\mu}{1-\mu})$

Dans notre étude, nous modélisons le coût moyen d'une prestation, une quantité à valeur dans $\mathbb{R}^+$ que nous avons modélisée par des lois de distributions continues positives, notamment la loi log-normale et la loi gamma qui sont les plus appropriées par rapport à notre contexte.

Cas d'une distribution log-normale

Une variable continue positive Y suit une loi log-normale $Log-\mathcal{N}(\mu,\sigma)$ si $X = log(Y)$ suit la loi gausienne de paramètres $\mu > 0$ et $\sigma^2 > 0$ notée $\mathcal{N}(\mu,\sigma^2)$. La densité de Y f_Y s'écrit alors :

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp(\frac{ln(y)-\mu}{2\sigma^2}) \text{ pour x} > 0$$

Cas d'une distribution gamma

Une variable continue positive Y suivant la loi de distribution gamma de paramètres $\alpha > 0$ et $\beta > 0$ notée $\Gamma(\alpha,\beta)$ est définie par la fonction densité $f_{\alpha,\beta}$ et s'écrit :

$$f_{\alpha,\beta}(y) = \frac{\beta}{\Gamma(\alpha)} y^{\alpha-1} exp(-\beta y)$$

Où $\Gamma(x) = \int_0^{+\infty} exp(-u)u^{x-1}du$
Définie ainsi, cette loi fait partie de la famille des lois exponentielles avec :

- Le paramètre naturel $\theta = -\beta$;
- Le paramètre de dispersion $\phi = 1$;
- $u(\phi) = 1$ et $v(\theta) = -\alpha ln(\theta)$;
- $c(y,\phi) = (\alpha-1)(ln(y) - ln(\Gamma))$

Rappelons que l'espérance et la variance d'une variable aléatoire dont la densité est de la forme exponentielle sont définies de la façon suivante,

$$\mathbb{E}[Y] = b'(\theta) \text{ et } \mathbb{V}(Y) = b''(\theta)a(\phi)$$

Pour chaque loi de la famille exponentielle, il exsite une fonction de lien qui permet de faire le lien entre l'espérance μ et le paramètre naturel θ de la loi de Y appelée la fonction de lien canonique que nous notons g_c :

$$\theta = g_c(\mathbb{E}[Y]) = g_c(g^{-1}(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p))$$

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

$$\theta = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

L'ensemble des paramètres inconnus du modèle linéaire généralisé est déterminé par maximum de vraisemblance.

4.1.3 Estimation des coefficients par maximum de vraisemblance

En considérant $y = (y_1, ..., y_n)$ comme étant une réalisation de l'échantillon de n variables aléatoires indépendantes et identiquement distribuées, $Y_1, ..., Y_n$ dont la fonction de densité est issue de la famille exponentielle et pour chaque i, y_i la réponse en $x_i = (x_1, ..., x_n)$, la vraisemblance de y s'écrit :

$$\mathcal{L}(y; \theta, \phi) = \prod_{i=1}^{n} f_{\theta;\phi}(y_i)$$

Où le paramètre naturel θ et le paramètre de dispersion sont inconnus et θ est en fonction des coefficients $\beta_1, ...\beta_p$. Maximiser la vraisembmance revient à maximiser la log-vraisemblance qui s'écrit comme suivant :

$$L(y; \theta, \phi) = \sum_{i=1}^{n} (\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi))$$

La valeur maximale de $L(y; \theta, \phi)$ est obtenue par la résolution des équations suivantes :

$$\begin{cases} \dfrac{\partial L(y;\theta,\phi)}{\partial \theta} \\ \dfrac{\partial L(y;\theta,\phi)}{\partial \phi} \end{cases} \Leftrightarrow \begin{cases} \dfrac{\partial L(y;\theta,\phi)}{\partial \beta_i} \text{ pour } i = 1, ..., p \\ \dfrac{\partial L(y;\theta,\phi)}{\partial \phi} \end{cases}$$

Dans les modèles linéaires généralisés, les paramètres estimés par maximum de vraisemblance sont non linéaires et les solutions de ces équations ne sont pas explicites. En pratique, il faut recourir à des algorithmes d'optimisation itératifs pour maximiser la fonction de vraisemblance. Il existe de nombreux algorithme d'optimisation, parmis elles , l'algorithme de Newton-Raphson et l'algorithme du Fisher-scoring sont les plus courants.

4.2 Arbre de regression

Afin de challenger, la méthode classique pour l'élaboration d'un zonier, nous avons opté pour une deux autres méthodes différentes qui sont des algorithmes d'appretissage incontournables en machine learning, basés sur l'assemblage d'arbres de décision.

4.2.1 Les arbres de régression de type CART

Les arbres de décision sont des méthodes d'apprentissage statistiques visant à construire un modèle de prédiction, nous parlons de classification lorsqu'il est question de quantités discrètes, dans le cas contraire, on parle régression pour la prédiction de quantités continues.

Il existe différents algorithmes de construction d'arbre de décision, cependant, dans la suite, nous aborderons uniquement l'algorithme CART (Classification And Regression Trees) introduit par Breiman, Friedman, Olshen et Stone en 1984. Le principe de cette approche est de partitionner récursivement et de manière binaire et homogène l'espace des variables explicatives afin d'obtenir toutes les valeurs possibles de la variable à prédire.

Arbre binaire

Un arbre binaire est une construction hiérarchique constituée de plusieurs éléments : la racine qui représente le premier nœud et comprend la totalité des observations, la racine est liée à deux autres nœuds fils, qui eux-mêmes peuvent être liés à deux nœuds fils, et ainsi de suite. Les nœuds n'ayant pas de descendant sont appelés des feuilles. Ainsi, un arbre se parcourt de la racine aux feuilles.

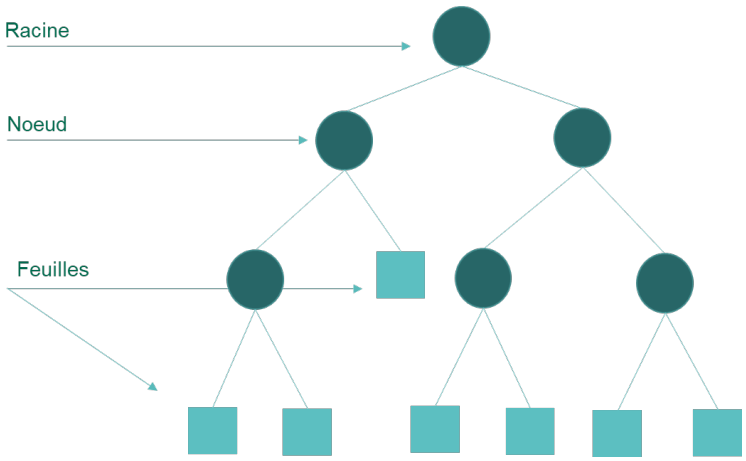


FIGURE 44 – Exemple d'arbre de décision

Arbre de régression de type CART

L'algorithme de CART est une approche non paramétrique reposant sur le partitionnement récursif de l'espace des observations, noté $\chi = (Y, X_1, ..., X_p) \in \mathbf{R^{n*p+1}}$, ce qui se représente par un arbre binaire de décision. Les valeurs prédites sont basées sur un partitionnement récursif reposant des conditions judicieusement établies à partir des données explicatives afin que les observations à chaque nœud soient les plus homogènes possible. Le partitionnement de l'arborescence est donc caractérisée par :

— des règles d'arrêt : à chaque création d'un nœud fils, les critères d'arrêt sont vérifiés, et si aucune des conditions n'est remplie, le nœud est séparé à son tour en deux noeuds fils. Ces règles ont un impact direct sur la taille de l'arbre en particulier le nombre de noeud. Les arrêts reposent sur des principes simples tel que la quantité minimum d'observation à avoir dans chaque feuille ou encore la profondeur de l'arbre.

— des conditions de séparation : contrairement aux arrêts, les conditions de division reposent sur des critères mathématiques afin de sélectionner la meilleure variable de séparation que nous aborderons un peu plus loin.

En la pratique, afin de déterminer le "meilleur" arbre, l'algorithme CART procède en trois étapes :

La première étape consiste à construire un arbre de taille maximal, noté T_{max}, autrement dit construire une suite de partitions de plus en plus fines de l'espace des observations

X jusqu'à ce que chaque feuille ne contienne qu'une seule observation ou des observations ayant la la même valeur à prédire. À chaque nœud, la condition de division est sélectionnée de manière optimale afin que les noeuds fils soient le plus homogène possible.

La notion d'homogénéité dans chaque noeud se traduit par une fonction d'impureté notée $i()$. Il existe plusieurs fonctions qui permettent de définir l'impureté $i(N)$ d'un nœud N, mais dans le cas présent de la prédiction d'une variable aléatoire Y continue, nous prendrons la moyenne des erreurs de prédiction au carré. L'impureté d'un nœud parent restant constante indépendamment du découpage réalisé pour créer les nœuds fils. Maximiser l'homogénéité dans les nœuds fils est alors équivalent à maximiser la variation d'impureté engendrée par une division.

Principe de construction d'une division optimale

Soit N un nœud de T_{max}, soit N_d son descendant droit et N_g son descendant gauche, descendant engendré par une division δ. Notons $p_g = \frac{p(N_g)}{p(N)}$ et $p_d = \frac{p(N_d)}{p(N)}$ les proportions d'observation envoyées respectivement dans N_d et N_d.
La variation d'hétérogénéité générée par δ est définie par :

$$\Delta i(\delta, N) = i(N) - p_g(N_g) - p_d(N_d)$$

La division optimale du nœud N est donnée par :

$$\delta^*(N) := \delta^* = argmax \Delta i(\delta, N)$$

Il est à noter que la division optimale est récursivement définie localement afin que celle-ci devienne un optimal global, il faut donc vérifier que la différence d'impureté Δi soit concave, autrement dit $\Delta i \geq 0$.

L'arbre maximal T_{max} n'est pas exploitable, en effet, le nombre de feuilles est bien trop important et en assimilant le nombre de feuilles à la complexité du modèle, nous obtenons un modèle de très complexe qui est beaucoup trop fidèle aux données et va donc sur-apprendre.

La deuxième étape consiste à élaguer l'arbre T_{max}, l'idée de cette seconde phase consiste à créer, à partir de l'arbre maximal, une suite de sous-arbres moins complexes et bien adaptés au problème. cette suite de sous arbres est construite autour d'un critère d'élagage.

Principe d'élagage

Soit T un arbre et N un nœud non terminal de T. Élaguer T à partir de N consiste à créer un nouvel arbre T' qui n'est autre que T privé de tous les descendants de N.

Une fois l'arbre de régression construit, si le nombre de feuilles est jugé trop grand, on peut le simplifier en élaguant ses branches de bas en haut.

Un élagage judicieux s'arrête quand on atteint un bon compromis entre la complexité de l'arbre et la précision de la prédiction.

Les arbres de régression présentent plusieurs avantages, tout d'abord ce sont des modèles simples à comprendre et à interpréter. En effet, il peut être représenté de manière explicite ce qui permet une compréhension facile. De plus, il n'y a pas de contrainte sur la nature des données explicatives, elles peuvent être quantitatives ou qualitatives.

Cependant, il arrive souvent que ce type de modèle soit trop dépendant de l'échantillon qu'il a parcouru, en particulier des points exceptionnels voir aberrants. Il retient alors ces traits exceptionnels et les considère comme des comportements normaux, ce qui génère un biais. Ce problème est t'autant plus présent lorsqu'il y a un très grand nombre de feuilles/noeuds. Ainsi, l'élagage de l'arbre permet de ne pas prendre en compte des individus atypiques non pertinents à l'étude.

Par ailleurs, les modèles agrégés permettent de sélectionner le critère qui est redondant sur l'ensemble des arbres construits et mesurent sa contribution dans la construction de la variable d'intérêt dans chaque arbre.
Un arbre de regression étudie les variables explicatives successivement. Les nœuds sont construits de façon enchaînée et un critère choisi pour figurer à un emplacement de l'arbre n'est plus réétudié par la suite. Ce qui suggère que modifier en amont la construction d'un critère fort remet en question la construction de l'intégralité de l'arbre. L'agrégation des modèles permet ici aussi de remonter les variables les plus déterminantes dans l'explication de la variable cible.

4.2.2 Principe du Bagging

Le Bagging, également connu sous le nom de Bootstrap Aggregation, est une méthode utilisée pour améliorer les performances des arbres de décision, en particulier les arbres de régression présentés précédemment.

Les arbres de décision ont tendance à produire beaucoup de variance, en d'autres termes, si nous entraînons deux arbres de décision différents sur deux échantillons aléatoires différents de la même base de données, les résultats peuvent différer considérablement.

Afin de limiter cet inconvenient, le Bagging vise à réduire la variance de l'estimateur en corrigeant l'instabilité des arbres de décision en se basant sur le principe du bootstrap pour créer de nouveaux échantillons aléatoires en tirant des échantillons avec remplacement de l'échantillon original.

Considérons (X, Y) un vecteur aléatoire où X prend ses valeurs dans $\mathbb{R}^p$ décrivant Y dans R. Notons $D_n = (X_1, Y_1), ..., (X_n, Y_n)$ un n-échantillon i.i.d. et de même loi que (X, Y) et $m(x) = \mathbb{E}[Y|X = x]$ la fonction de régression. Pour $x \in \mathbb{R}^p$, notons également l'erreur quadratique moyenne d'un estimateur et sa décomposition biais-variance :

$$\mathbb{E}[(\hat{m}(x) - m(x))^2] = (\mathbb{E}[\hat{m}(x)] - m(x))^2 + \mathbb{V}(\hat{m}(x)).$$

Le bagging est une méthode d'agrégation, elle consiste à agréger un nombre B d'estima-

teurs $\hat{m}1, ..., \hat{m}_B$:

$$\hat{m}(x) = \frac{1}{B}\sum_{k=1}^{B} \hat{m}_k(x)$$

.

Le biais de l'estimateur agrégé est donc le même que celui des $\hat{m}_k$ mais la variance diminue.

L'approche bagging consiste à tenter d'atténuer la dépendance entre les estimateurs agrègés en les construisant sur des échantillons bootstrap. Les étapes de l'algorithme sont les suivantes :

Algorithm 1 Bagging

Entrées :

- d_n l'échantillon ;
- un regresseur de type arbre CART θ ;
- B le nombre d'estimateurs à agréger.

Pour k=1 à B :

1. Tirer un échantillon boostrap d_n^k dans d_n
2. Ajuster le régresseur sur cet échantillon bootstrap : $\hat{m}_k$

Sorties : L'estimateur $\hat{m}(y) = \frac{1}{B}\sum_{k=1}^{B} \hat{m}_k(y)$

.

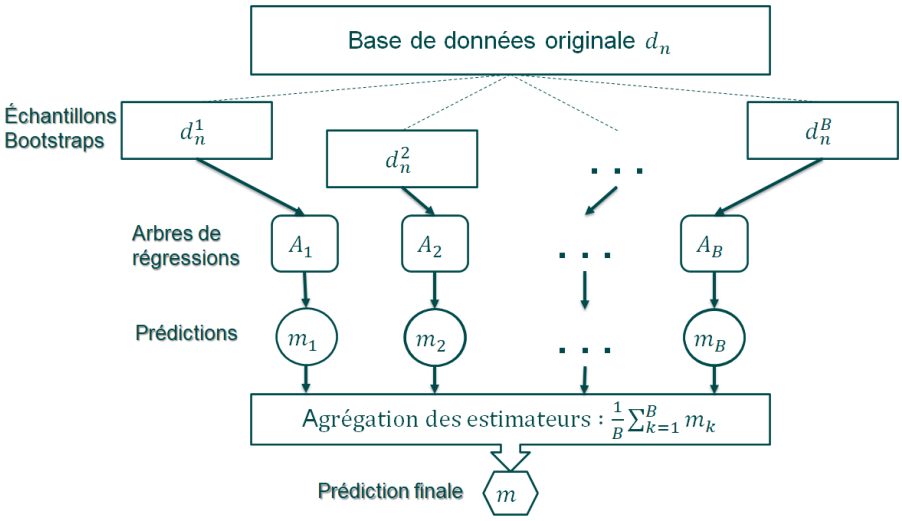


FIGURE 45 – Illustration du bagging

4.2.3 Les fôrets aléatoires

Les forêts aléatoires offrent une amélioration supplémentaire par rapport aux d'arbres du bagging grâce à un ajustement qui permet de décorréler les arbres. En effet, comme dans le bagging, ils sont contruits à partir d'un certain nombre d'arbres de décision utilisant des échantillons bootstrap. Cependant, lors de la construction des arbres, à chaque nœud, un échantillon aléatoire de m variables est choisi parmi l'ensemble de p variables de base pour définir le nœud. Ainsi, chaque nœud est construit avec une seule variable de l'échantillon.

Algorithm 2 Fôrets aléatoires

Entrées :

- x l'observation à prédire ;
- d_n l'échantillon ;
- B le nombre d'abres ;
- $m < p \in \mathbb{N}$ le nombre de variables candidates pour découper un nœud.

Pour k=1 à B :

1. Tirer un échantillon boostrap dans dn
2. Construire un arbre CART sur cet échantillon bootstrap, chaque coupure est sélectionnée en minimisant la fonction de coût de CART sur un ensemble de m variables choisies au hasard parmi les p. On note $h(., \theta_k)$ l'arbre construit.

Sorties : L'estimateur $h(x) = \frac{1}{B}\sum_{k=1}^{B} h(x, \theta_k)$

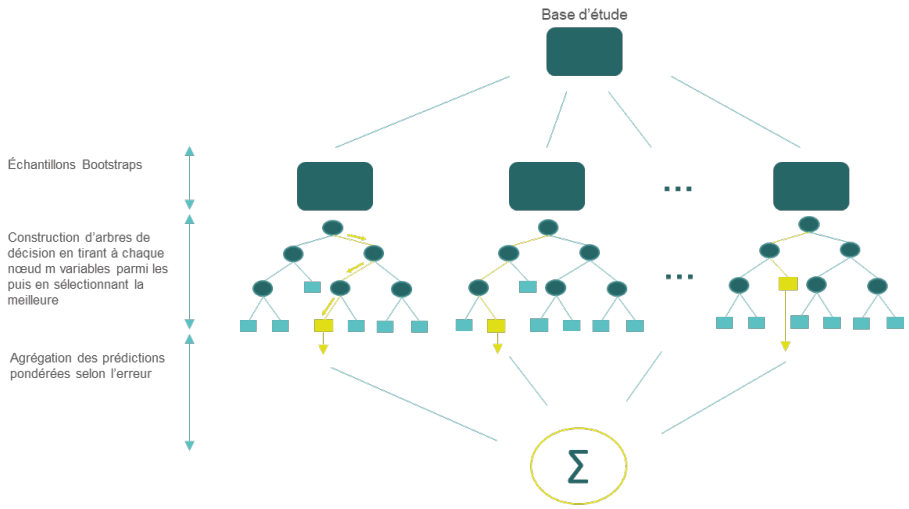


FIGURE 46 – Illustration du principe de fôret aléatoire

En d'autres termes, contrairement au bagging, l'algorithme de la forêt aléatoire ne considère pas seulement la moitié des variables existantes lors de la construction de chaque nœud de l'arbre. Cela peut sembler contre-intuitif, mais c'est en fait une solution efficace

pour éviter les prédictions très corrélées qui se produisent souvent avec le bagging lorsque certaines variables ont un pouvoir prédictif très élevé.

En effet, si une variable avec un pouvoir prédictif très élevé était utilisée pour les premiers nœuds de la plupart ou de tous les arbres, les prédictions de bagging seraient très corrélées. Malheureusement, la moyenne de quantités très corrélées ne permet pas de réduire la variance autant que la moyenne de quantités décorrélées.

Pour éviter cela, la méthode des forêts aléatoires oblige chaque nœud à considérer seulement une partie des variables. Ainsi, en moyenne, il y a des chances qu'un nœud ne prenne même pas en compte la variable à très fort pouvoir prédictif, ce qui donne plus de chances aux autres variables d'être sélectionnées.

La principale différence entre le bagging et les forêts aléatoires réside dans le choix de m, le sous-ensemble de variables retenu. Si m = p, la forêt aléatoire correspond au bagging. En choisissant de petites valeurs de m lors de la construction d'une forêt aléatoire, on peut tirer parti de la capacité de la méthode à gérer les variables corrélées, ce qui peut être particulièrement utile lorsque le nombre de variables est élevé.

4.2.4 Le Gradient Boosting

Le Gradient Boosting est une méthode de machine learning qui combine deux procédés, le boosting qui combine plusieurs arbres de décision faibles pour créer un modèle plus robuste et précis.

Le boosting suit une approche entraînant plusieurs arbres de régression faibles successifs sur les résidus de l'arbre précédent. Chaque arbre est construit pour corriger les erreurs de prédiction du modèle précédent. Le boosting permet ainsi de réduire l'erreur de biais et d'améliorer la précision prédictive.

Contrairement au bagging, qui combine les prédictions de plusieurs arbres indépendants, le boosting combine les prédictions de plusieurs arbres successifs faibles en ajustant leurs poids pour minimiser la fonction de coût globale. Ainsi, le boosting est plus efficace que le bagging pour les ensembles de données bruyants ou déséquilibrés et peut souvent obtenir des performances prédictives supérieures.

Le second procédé utilisé est la technique de la descente de gradient pour ajuster les poids de chaque arbre de décision faible afin de minimiser la fonction de perte globale. La descente de gradient est une méthode d'optimisation qui ajuste les paramètres d'un modèle en minimisant progressivement la fonction de coût globale.

Enfin, les prédictions de tous les arbres de décision faibles sont agrégés en une seule prédiction finale en additionnant les prédictions individuelles pondérées par les poids calculés par la descente de gradient.

Plus formellement, pour prédire la variable $Y = f(X)$ où $X = (X^1, ..., X^p) \in \mathbb{R}^p$ est l'ensemble des prédicteurs de y. Considérons une fonction de pertes L (loss function) à minimiser. Il existe différentes fonctions de pertes L, les principales sont les suivantes :

— $L(y, f(x)) = (y - f(X))^2$

— $L(y, f(x)) = ||y - f(X)||$

— $L(y, f(x)) = \begin{cases} 1 \text{ si } y = f(x) \\ 0 \text{ sinon.} \end{cases}$

L'espérance mathématique de cette mesure définit la fonction de risque (ou erreur)

$$\mathcal{R}(f(X)) = \mathbb{E}[L(Y, f(X)].$$

Le but est alors de trouver une approximation $\hat{f}(x)$ de la fonction $f^*(X)$ qui minimise l'espérance de $L(Y, f(X))$:

$$f^*(x) = argmin_f \mathbb{E}(L(y, f(x)))$$

L'algorithme se définit par les étapes suivantes :

Algorithm 3 Grandient Boosting

Entrées :

— $d_n = \{(x_i, y_i)\}_{i=1}^n$ l'échantillon de données ;

— Une fonction perte $L(y, f(x))$ différentiable ;

— M le nombre d'itérations.

— λ le taux d'apprentissage ou shrinkage.

1. Initialisation du modèle avec une constante : $\hat{f}_0(x) = argmin \sum_{i=1}^n L(y_i, \gamma)$
2. Pour m=1 à M :
   (a) Calculer les pseudo-résidus pour $i = 1, ..., n$:

$$r_{m_i} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{m-1}(x)}$$

   (b) Ajuster un arbre de regression c_m aux couples $((x_i, r_{m_i})_i$;
   (c) calculer $\gamma_m = argmin_\gamma \sum_{i=1}^n L(y_i, f_{m-1}(x_i)) + \gamma_m c_m(x_i)$
   (d) Mettre à jour $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m c_m(x)$

Sorties : $\hat{f}_M(x) = \sum_{m=1}^M \gamma_m c_m(x) + f_0(x)$

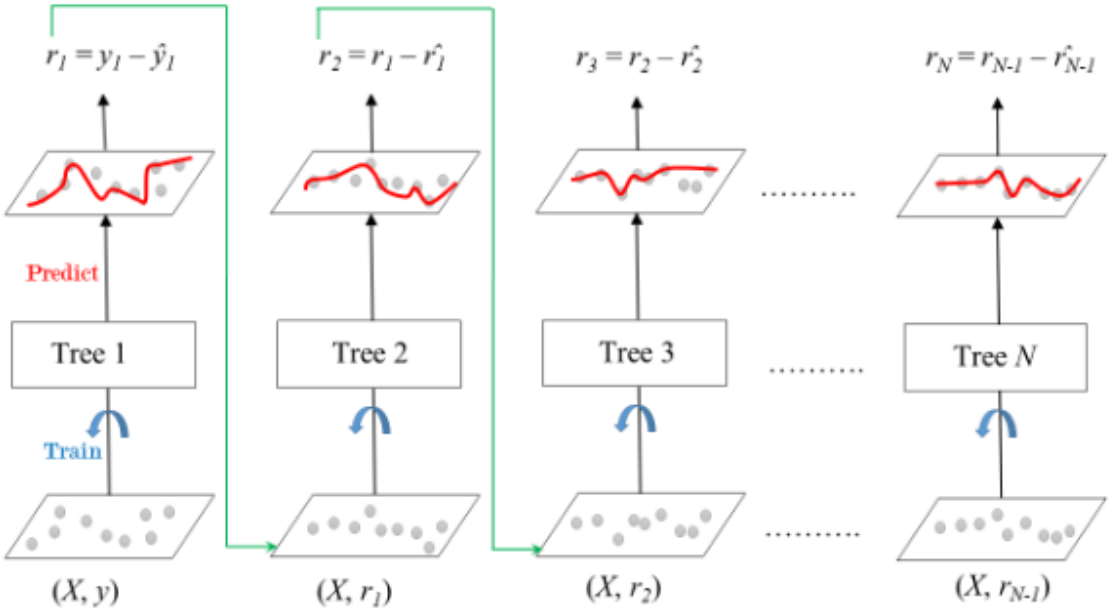


FIGURE 47 – Illustration du boosting

L'ensemble des méthodes de machine learning que nous avons évoquées sont sujet au problème de surapprentissage, En effet un ajustement trop proche des données d'apprentissage peut entraîner une détérioration de la capacité de généralisation du modèle. Plusieurs techniques dites de régularisation réduisent cet effet de surajustement en encadrant la procédure d'ajustement par l'hyperparametrage.

L'hyperparamétrage est le processus de sélection des paramètres pour contrôler l'apprentissage à travers :

- Le nombre d'intération M, s'il augmente cela permet de réduire l'erreur de prédiction sur les données d'apprentissage, mais peut entraîner un surajustement. Une valeur optimale de M est souvent sélectionnée par validation croisée.
- Le taux d'apprentissage (shrinkage) qui permet de controler la vitesse de convergence. Si sa valeur est petite cela conduit à accroître le nombre d'ajustements de base nécessaire et entraîne généralement une amélioration de la qualité de prédiction.
- La profondeur maximale de l'arbre est un autre paramètre à optimiser pour contrôler la complexité de l'algorithme. Plus la profondeur maximale est grande, plus l'algorithme devient complexe, mais est capable de capter les interactions complexes entres les variables.

4.3 Métriques de comparaison et de validation des modèles

Dans toute modélisation, en sus de la fiabilisation de la base de données d'étude, il est important de valider et de comparer différents modèles pour s'assurer de choisir le modèle le plus adapté aux données et au contexte du projet.

Dans notre cadre, la validation de modèle passe par la validation des hypothèses retenues avec des tests d'hypothèses et d'adéquation mais c'est aussi le cas de la mesure de la qualité de la prédiction.

La mesure de la qualité de la prédiction permet de quantifier la capacité d'un modèle à se généraliser sur de nouvelles données. Cette mesure s'obtient à l'aide d'une métrique d'évaluation appelée aussi métrique de validation, il en existe plusieurs. Les plus connues pour les problèmes de régression sont le R^2, RMSE, MAE. Cependant, ces métriques donnent une vision globale de la qualité de la prédiction du modèle sur l'ensemble des données. De plus, certaines métriques sont moins adaptées pour des données assurantielles.

Dans le cadre de ce mémoire, nous avons élaboré un processus d'évaluation mieux adapté à nos objectifs du métier afin d'évaluer efficacement nos différents modèles.

4.3.1 Mean Squared Error (MSE)

Le Mean Squared Error (MSE), correspond à l'erreur quadratique moyenne d'un estimateur. Dans notre cas, si nous notons $(Y_i)_i$ les coûts moyens des prestations de soins observés dans la base et $(\hat{Y}_i)_i$ les coûts moyens des prestations prédits par un modèle, alors l'erreur quadratique du modèle est calculée comme suivant :

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

Le MSE est facilement compréhensible dans sa définition : si le MSE est plus élevé, cela signifie que la moyenne des coûts prédits est plus éloignée de la moyenne des coûts observés, ce qui indique que le modèle est moins ajusté. En revanche, un MSE de zéro indique une estimation parfaite, car toutes les valeurs prédites sont égales aux valeurs observées. Pour évaluer deux modèles, il suffit de comparer leur MSE et de choisir celui qui a la valeur la plus petite.

4.3.2 La Déviance

La déviance D d'un modèle statistique est une mesure utilisée dans les modèles de régression pour évaluer l'ajustement du modèle aux données. Elle est souvent utilisée pour comparer deux modèles de régression différents pour déterminer lequel a un meilleur ajustement aux données.

La déviance est définie comme la différence entre la log-vraisemblance du modèle ajusté et la log-vraisemblance d'un modèle nul, qui est le modèle de référence qui ne contient que l'ordonnée à l'origine. Plus précisément, la déviance est donnée par la formule suivante :

$$D = -2(L(m) - L(m_0))$$

Avec :

— $L(m_0)$ la log-vraisemblance du modèle ajusté

— $L(m)$ la log-vraisemblance du modèle null

La déviance est une mesure de la différence en termes de l'ajustement aux données entre le modèle ajusté et le modèle nul. Elle mesure à quel point le modèle ajusté est meilleur que le modèle nul au vu de sa capacité à prédire les valeurs observées.

En d'autres termes, une valeur de déviance plus faible indique un meilleur ajustement du modèle aux données. Cependant, la déviance ne fournit pas d'informations sur la qualité de l'ajustement global du modèle ou sur la précision de la prédiction. Il est donc important de considérer d'autres mesures d'évaluation pour évaluer les performances globales du modèle et sa précision de prédiction.

4.3.3 L'AIC et le BIC

L'AIC (Akaike Information Criterion) et le BIC (Bayesian Information Criterion) sont des mesures utilisées pour évaluer la qualité de l'ajustement d'un modèle. Les deux critères sont basés sur la log-vraisemblance du modèle ajusté et prennent également en compte la complexité du modèle.

$$AIC = 2L(\hat{\beta}) + 2k$$
$$BIC = 2L(\hat{\beta}) + klog(n) \quad [$$

où k est le nombre de paramètres dans le modèle, le coefficient 2 est utilisé pour pénaliser les modèles avec un grand nombre de paramètres et n est la taille de l'échantillon.

En général, l'objectif est de minimiser l'AIC ou le BIC pour obtenir le modèle le plus approprié. Cependant, il est important de considérer d'autres mesures d'évaluation pour évaluer les performances globales du modèle et sa précision de prédiction.

L'AIC et le BIC sont souvent utilisés pour comparer différents modèles de régression en utilisant la même base de données. Cependant, ils ne doivent pas être utilisés pour comparer des modèles différents qui utilisent des bases de données différentes. De plus, ces critères supposent que le modèle est correctement spécifié et que les résidus suivent une distribution normale.

4.3.4 Test d'adéquation

Le test de Kolmogorov-Smirnov est un test statistique non paramétrique utilisé pour tester si un échantillon de données suit une distribution donnée. Ce test compare la distribution observée d'un échantillon statistique à une distribution théorique. Il est basé sur la comparaison des fonctions de répartition.

4.4 Classification hiérarchique ascendante CAH

Il existe de nombreuses méthodes classifications, elles visent toutes à construire des regroupements les plus homogènes possibles et se distinguent en deux catégories :

— La classification supervisée consiste à définir les caractéristiques d'observations déjà classées afin de prédire de nouvelles observations de classe inconnue à partir de leurs caractéristiques.

— La classification non supervisée correspond quant à elle à la situation dans laquelle nous disposons d'observations non classées et nous souhaitons les regrouper en classes homogènes.

Dans notre contexte, nous devons utiliser des méthode de classifications non supervisée puisque nous souhaitons classer les départements. Nous procédons par la méthode hierarchique ascendente.

La classification ascendante est un algorithme automatique qui construit une suite de partitions emboîtées des données en n classes, $n-1$ classes, ... , 1 classe. Le plus souvent pour cette méthode classification, nous utilisons la distance de Ward qui défini la distance entre deux classes par la distance euclidienne de leurs barycentres au carré, pondérée par les effectifs des deux classes.

Si d est la distance euclidienne, la méthode de Ward :

— La distance de Ward pour les classes c_i et c_j et leurs barycentres respectifs g_i et g_j :

$$d_w(c_i, c_j) = d(g_i, g_j)^2$$

— Minimise l'inertie interclasse : notons n_i l'effectif de la classe c_i, et g le barycentre de l'ensemble des observations.

$$I_{inter} = \frac{1}{n} \sum_{i=1}^{k} n_i d(g_i, g)^2$$

— Maximise l'inertie intraclasse : k un nombre de classes fixé

$$I_{intra} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i d(g_j, g_i)^2$$

L'algorithme est le suivant :

1. À l'état initial, chaque observation constitue une classe, la variance interclasse est à son maximum et la variance intraclasse est nulle.
2. États intermédiaires : les distances deux à deux entre classe sont calculées, et les deux classes les plus proches (au sens de la distance de ward) sont réunies, cette étape se répète jusqu'à obtenir une seule classe. La variance interclasse diminue et la variance intraclasse augmente.
3. À l'état final, il ne reste plus qu'une classe, la variance interclasse est nulle et la variance intraclasse est à son maximum.

4.5 Lissage spatial par la théorie de crédibilité

Plutôt dans ce présent mémoire, nous avons relevé le manque de robustesse dans des indicateurs, notamment le coût moyen des frais de soins dès lors que nous passons sur une maille plus fine. En somme, la maille départementale nous apporte de la précision, mais elle fragilise la robustesse d'indicateurs calculés, ce qui est dû à un effet d'effectifs.

Pour atténuer l'inconvénient engendré par l'utilisation de la maille départementale, nous avons opté pour un lissage spatial. Il existe plusieurs méthodes de lissage spatiale, mais dans ce mémoire, nous aborderons uniquement le lissage par la théorie de crédibilité de Bühlmann-Straub.

4.5.1 En théorie

Le modèle de Bûhlmann-Straub comme tous les autres modèles de crédibilité introduit un coefficient de crédibilité dans l'estimation d'une cible à modéliser telle que la prime pure, le coût moyen ou encore le nombre de sinistres dans la tarification de différents produits en assurance.

Posons le cadre théorique. Ce modèle part d'un portefeuille de contrats d'assurance définis chacun par un paramètre de risque θ qui contient toute l'information nécessaire pour décrire le risque. Soient X une variable aléatoire discrète représentant une grandeur statistique (coût moyen, nombre de sinistres, ...) et θ un paramètre de risque.

À présent notons par I le représentant du portefeuille de risque, par $X_{i,j}$ la grandeur statistique 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. Le modèle de Bülmann-Straub suposse les deux hypothèses suivantes,

1. Les variables aléatoires $X_{i,j}$ sont conditionnellement à Θ_i, indépendantes, et de moments conditionnels
$$\mu(\Theta_i) = \mathbb{E}[X_{i,j}|\Theta_i]$$
$$\sigma^2(\Theta_i) = \omega_{i,j}\mathbb{V}ar[X_{i,j}|\Theta_i]$$
2. Pour un j fixe, les couples $(\Theta_1, X_{1,j})$,..., $(\Theta_I, X_{I,j})$ sont indépendants et les Θ_i sont indépendants et identiquement distribués pour $i = 1, ..., I$.

Ainsi la variance des $X_{i,j}$ s'écrit :

$$\begin{aligned}\mathbb{V}ar[X_{i,j}] &= \mathbb{V}ar[\mathbb{E}[X_{i,j}|\Theta_i]] + \mathbb{E}[\mathbb{V}ar[X_{i,j}|\Theta_i]] \\ &= \mathbb{V}ar[\mu(\Theta_i)] + \frac{1}{\omega_{i,j}}\mathbb{E}[\sigma^2(\Theta_i)] \qquad [\end{aligned}$$

Nous noterons $a := \mathbb{V}ar[\mu(\Theta_i)]$ et $s^2 := \dfrac{1}{\omega_{i,j}}\mathbb{E}[\sigma^2(\Theta_i)]$

Fixons maintenant i, le meilleur estimateur sans biais de $\mu(\Theta_i)$ est,

$$\begin{aligned}\hat{\mu}(\Theta_i) &= \bar{X}_i \\ &= \sum_j \frac{\omega_{i,j}}{\omega_i} X_{i,j} \text{ avec } \omega_i = \sum_j \omega_{i,j} \qquad [\end{aligned}$$

Et $\mathbb{V}ar[\bar{X}_i] = a + \dfrac{\sigma^2(\Theta_i)}{\omega_i}$

Dans le modèle de Bühlmann-Straub l'estimateur appelé estimateur de crédibilité est de la forme suivante :

$$\mu(\hat{\Theta}_i) = Z_i X_i + (1 - Z_i)\mu_0.$$

Avec Z_i le facteur de crédibilité associé au risque i et satisfait la relation suivante :

$$Cov(\mu(\hat{\Theta}_i), X_i) = Z_i Cov(X_i, X_i) = Cov(\mu(\Theta_i), X_i). \quad [$$

Par conséquent, le facteur de crédibilité est redéfini par l'expression suivante :

$$Z_i = \frac{\omega_i}{\omega_i + \dfrac{\sigma^2}{s^2}} \in [0, 1]$$

4.5.2 En pratique

L'adaptation du modèle de Bûhlmann-Straub consiste à attribuer à chaque département un coefficient de crédibilité reflétant la robustesse de celle-ci jugée par les effectifs consommant. Ainsi, le coût moyen d'un département dont la crédibilité est faible et sera estimé avec les coûts moyens de départements alentours. La proximité des départements est introduite par la notion de distance entre les départements.

En effet, le lissage «distance» utilise la méthode de crédibilité exposée ci-dessus, et lisse le risque d'un département à partir de la moyenne de risque des autres départements pondérée par l'effectif et une fonction distance. La fonction en question assigne une influence plus grande aux département plus proches. L'expérience lissée est basée sur une moyenne pondérée par l'effectif consommant du département et des département aux alentours, avec une influence plus grande assignée aux département 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 un département $i = 1, ..., 94$ et un rayon $\delta > 0$, le lissage distance du coût moyen c_i est donné par la formule suivante :

$$C_i = Z_i c_i + (1 - Z_i)\bar{c_i}$$

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

$$Z_i = \frac{\omega_i}{\omega_i + \dfrac{\sigma^2}{s^2}}$$

. Ici ω_i est le poids de la crédibilité, correspondant donc à l'effectif par rapport à l'effectif total, et $\omega_0 = \frac{\sigma^2}{s^2}$ qui compense le poids de la crédibilité, correspondant donc au rapport des variances inter et intra departements, qui seront estimées de manière empirique.

Le risque individuel c_i correspond au résidu du département i.
Le risque collectif $\bar{c_i}$:

$$\bar{c_i} = \frac{\sum_{j=1}^{94} c_i d_{i,j} \omega_j^{\delta_i}}{\sum_{j=1}^{94} d_{i,j} \omega_j^{\delta_i}} \text{ pour } i \neq j \text{ et } d_{i,j} < \delta_i$$

Ou,

— $\omega_j^{\delta_i}$ est le poids du département j dans le rayon d'influence du département i, il correspond donc à l'éffectif du département j par rapport à l'effectif total de la zone d'influence du département i.

— $d_{i,j}$ la distance entre le département i et le département j. En pratique, nous utilisons les distances entre les départements données par Google Map, que nous avons extrait à partir de l'API google.

— $\bar{c_i}$ correspond donc à la moyenne des coûts moyens de la zone d'influence non lissés, pondérés par leur distance avec les départements j et leur poids $\omega_j^{\delta_i}$ tel que $d_{i,j} < \delta_i$.

Ainsi, les données peu robustes sont estimées avec la moyenne pondérée des données proches démilitées par un rayon.

5 Applications des modèles

Dans cette partie, nous allons mettre en application les modèles présentés dans la partie précédente afin d'expliquer le coût réel moyen et d'en extraire l'effet géographique par le biais des résidus et ainsi construire un zonage à partir des résidus accompagnés de variable conteant une partie de l'information gégraphique.

Nous précisons que les études de ces modèles sont sur deux mailles différentes :

— la maille région : partir des variables définies pour chaque région française de la métropole et construire des regroupements de régions ayant un risque similaire. Ce niveau agrégation présente l'inconvénient de filtrer trop d'information. C'est pourquoi nous ferons l'étude d'une seconde maille.

— la maille départementale : partir des variables définies pour chaque département français de la métropole et construire des regroupements de département ayant un risque similaire. Ce choix d'agrégation permet de récupérer plus d'information, néanmoins, il rend les variables impliqués dans les modèles moins robustes.

La question de modéliser le coût moyen par poste de soins s'est posée, des études également ont été faites par poste de soins. Cependant, nous avons arbitrairement fait le choix de garder les résultats du risque pour tout poste de soins confondus.

Le zonier intégré dans l'outil est celui correspondant à la maille régionale qui reste la plus fiable. Cependant, dans la suite de ce mémoire, nous présenterons les résultats obtenus des différents modèles établis sur la maille départementale, car nous avons jugé plus intéressant de présenter la démanche mise en place pour tenter de corriger la non-robustesse des paramètres.

À travers le schéma suivant, nous pouvons voir toutes les étapes misent en place dans l'élaboration du zonier.

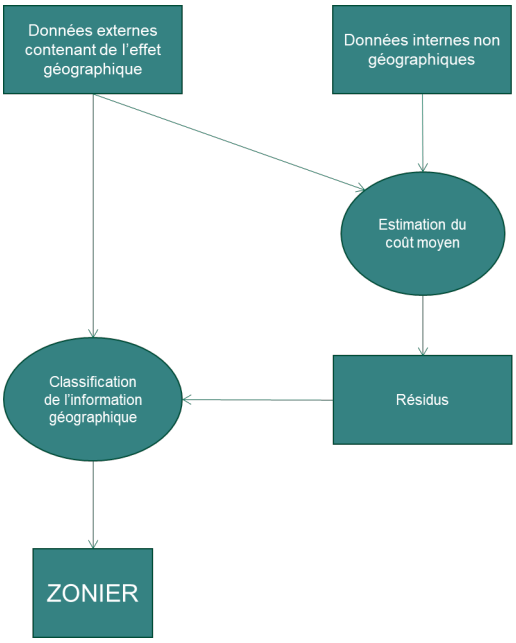


FIGURE 48 – Les étapes de la construction du zonier

5.1 Modélisation du risque géographique

Dans notre contexte, le risque correspondant à l'effet géographique est contenu dans le coût moyen des prestations médicales. Ce risque est complexe, car il n'est pas explicite. Pour estimer le risque géographique, nous suivons la procédure suivante :

1. Estimation des coûts moyens à partir des variables internes ainsi que les variables externes. Nous supposons que les variables internes sont non géographiques, les variables externes contienent une partie de l'information géographique et que le cout moyen est décomposé comme suivant :

$$Cout\_moyen = variables\_internes + variables\_externes + residus$$

Avec

$variables\_externes = effets\_non\_gographiques + effet\_gographique + bruit_1$ et

$$variables\_internes = effets\_non\_gographiques + bruit_2$$

Et les résidus s'écrivent comme la somme d'un bruit et d'une partie de l'effet géographique

$$residus = bruit + effet\_geographique$$

Pour réduire au mieux le bruit, nous garderons les résidus donnés par le meilleur estimateur.

2. Et enfin, nous récupérons toute l'information géographique en combinant les résidus et les variables non géographiques pour une classification.

Le coût moyen est estimé par trois méthodes, le GLM, la forêt aléatoire et le gradient boosting. Le meilleur estimateur est déterminé par les différentes métriques de comparaisons présentées.

5.1.1 Estimation des données non robustes par lissage spatial

Lors de l'agrégation des données par département, nous pouvons relever un manque d'information dans certains départements qui engendre dans le cas du calcul du montant de frais moyen des valeurs peu fiables causé par une présence faible d'assurés dans ces zones.

l'estimation des coûts moyens des départements peu robuste est établie un lissage spatial par la théorie de crédibilité, la robustesse est estimée à travers les effectifs des consommateurs du département en question. Cette étape permet d'atténuer l'effet d'effectifs.

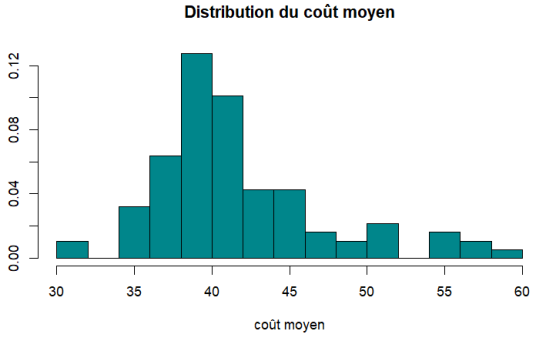


FIGURE 49 – Distribution du coût moyen avant lissage

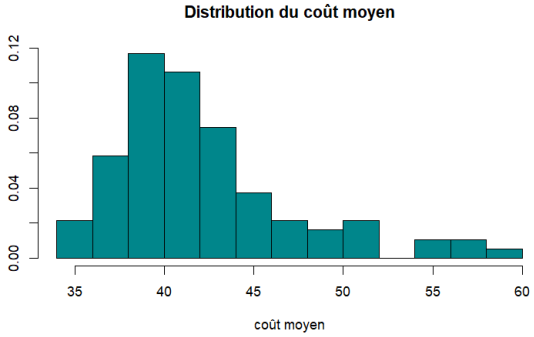


FIGURE 50 – Distribution du coût moyen après lissage

les histogrammes ci-dessus représentent la distribution du coût moyen de prestations avant et après le lissage spatial par théorie de crédibilité présenté dans la partie théorique. Nous constatons que le procédé ne dénature pas la distribution.

5.1.2 Estimation du cout moyen par GLM

Choix de la loi pour la modélisation du coût moyen

Dans un modèle linéaire généralisé, la variable à estimer Y est supposée suivre une loi de la famille exponentielle. Ainsi, il faut donc déterminer la loi exponentielle qui se rapproche le mieux de la distribution de notre risque. Ce dernier correspond au coût moyen par département prenant des valeurs continues et positives. Dans les lois usuelles présentées, deux d'entre elles sont souvent utilisées pour modéliser des coûts.

La figure 51 ci-dessous présente la comparaison de la distribution empirique (histogramme), du coût moyen, avec la distribution théorique d'une loi Gamma représentée par la courbe jaune et d'une loi Log-normale représentée par la courbe bleue.

Les paramètres des deux lois sont déterminés par maximum de vraisemblance à l'aide des fonctions *egamma* et *elnorm* de *R*.

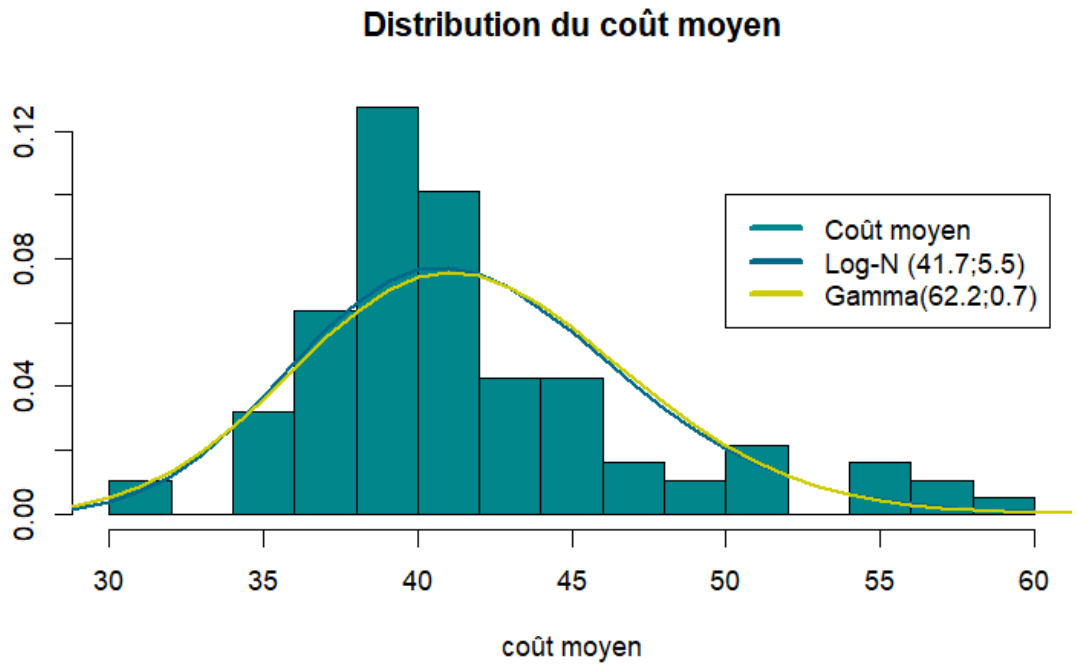


FIGURE 51 – Distribution du cout moyen

Graphiquement, les deux lois théoriques semblent donner des résultats assez similaires à la distribution empirique. Afin de choisir de manière optimale la loi qui servira à la modélisation du coût moyen, nous regardons les résultats des tests d'adéquation de Kolmogorov-Smirnov, de Cramer-Von Mises et ainsi que les résultats des métriques de comparaison données par les GLM de chaque loi. Nous conservons la loi ayant les meilleurs résultats.

Test d'adéquation entre la distribution du coût moyen et la loi Log-N(μ=3.7;σ²=0.1)			
Test	**Statistique**		**P-value**
Kolmogorov-Smirnov	D	0.12517	0.09639
Cramer-von Mises	ω²	0.32921	0.112

Test d'adéquation entre la distribution du coût moyen et la loi Gamma γ(α=56.8,β=0.7)			
Test	**Statistique**		**P-value**
Kolmogorov-Smirnov	D	0.97004	< 2.2e-16
Cramer-von Mises	ω²	31.138	< 2.2e-16

Nous constatons donc que la loi Log-normale présente une meilleure adéquation que la loi Gamma. Le résultat de la comparaison des deux lois avec la déviance et le critère AIC confirme la tendance reflétée par les tests d'adéquation.

Critère	Log-N(μ=3.7,σ²=0.1)	Gamma γ(α=56.8,β=0.7)
Déviance	1.18155	1.5158
AIC	-202.71	509.69

Dans le cadre du modèle GLM, nous retenons donc l'hypothèse que le coût moyen suit une loi Log-normale et la fonction de lien couramment associée à cette loi est la fonction identité.

L'ajustement du modèle

Après avoir sélectionné la loi la plus appropriée pour notre variable cible, nous effectuons une sélection des variables explicatives les plus pertinentes afin d'obtenir une meilleure estimation. Il existe plusieurs méthodes pour sélectionner les variables, mais l'objectif est de ne garder que les plus pertinentes. Nous utilisons la méthode basée sur l'AIC du modèle, qui est un critère de pénalisation de la log-vraisemblance du modèle en fonction du nombre de paramètres retenus.

Nous utilisons la fonction *step* de R avec la « sélection descendante » sur le critère AIC qui consiste à commencer avec toutes les variables explicatives et à les supprimer une par une, en commençant par celle qui a la plus petite contribution au modèle, jusqu'à ce que le modèle final ait le plus petit AIC possible. Cela signifie que les variables les moins importantes sont éliminées en premier, jusqu'à ce que le modèle ne puisse plus être amélioré de manière significative en supprimant davantage de variables.

Les variables retenues pour optimiser le GLM sont les suivants :

— la proportion d'hommes ;
— l'âge moyen ;
— la proportion de retraités ;
— les trois variables représentant les données externes.

La pertinence du modèle

Métrique	GLM	GLM optimisé
MSE	58.22797	9.855609
Déviance	1.18155	1.18155
Residual déviance	0.48304	0.49948
AIC	-202.71	-209.56

Comme nous l'avons évoqué dans la partie théorique des modèles, les modèles GLM sont très répandus dans la construction d'un zonier. Ces modèles ont la particularité de détecter les effets non linéaires et prennent en compte le caractère non gaussien dans la distribution des résidus. Cependant, bien que performants par rapport aux modèles de régression classique, les contraintes imposées par ces modèles telles que les interactions entre les variables explicatives ou encore les contraintes liées à la structure du risque font que ces modèles peuvent conduire à des résultats non pertinents. Procédons à présent à des modèles de Machine Learning impliquant des arbres de decision.

5.1.3 Estimation du cout moyen par Forêt aléatoire

Rappelons que la méthode de Random forest reprend l'algorithme du Bagging par la construction de plusieurs arbres de régression sur la base d'échantillons bootsraps, pour ensuite en faire une moyenne. La particularité des forêts aléatoires réside dans le fait de s'intéresser à un nombre réduit de variables pour chaque arbre, choisi aléatoirement au lieu de les prendre toutes en compte.

En pratique, nous utilisons la fonction *randomForest* du package h2o

Fonction RandomForest du package *h2o*	
randomForest (y=y.dep,x=x.dep, training_frame=,nfolds=, ntrees=,max_depth=,mtries=)	
Paramètres	description
y	La variable à expliquer
x	Les variables explicatives
training_frame	La base d'apprentissage
nfolds	Le nombre de partions dans la validation croisée
ntrees	Le nombre d'arbres retenus
max_depth	La profondeur maximale de chaque arbre
mtries	Le nombre de varibles tirées aléatoirement à chaque nœud

Dans la modélisation du coût moyen, nous adopterons le procédé suivant :

1. Nous lançons un premier modèle avec les paramètres par défaut pour calculer les métriques de comparaison sur ce modèle dans le but de le construire un modèle optimisé :
   - $mtries = p/3$, avec p le nombre de variables explicatives ;
   - $nfolds = 5$, nombre de partitions dans la validation croisée ;
   - $ntrees = 50$ arbres ;
   - $max\_depth = 17$, la profondeur maximale des arbres.
2. Nous procédons à l'hyperparamétrage en testant des valeurs différentes pour les paramètres qui impactent la complexité de l'algorithme :
   - $ntrees \in \{10, 20, 30, ..., 500\}$;
   - $mtries \in \{2, ..., 11\}$;
   - $max\_depth = 10$

3. Nous procédons à l'élaboration de la forêt aléatoire une fois que les hyperparamètres ont été déterminés. Nous traçons un graphique qui répertorie l'évolution de la MSE en fonction du nombre d'arbres. Cette étape nous permettra de prendre une décision sur le paramètre *ntrees*.
4. Nous observons l'importance de chaque variable, une variable sera considérée importante dans la construction de l'arbre si la différence entre le taux d'erreur avant et après perturbation est importante. L'importance de la variable dans la forêt correspond à la moyenne sur l'ensemble des arbres de la différence des taux d'erreurs avant et après perturbation, elle est ensuite divisée par l'écart type de cette différence s'il n'est pas nul.
5. Nous mesurons la pertinence des prédictions à l'aide de l'erreur quadratique et de déviance résiduelle.

Nous constatons que l'erreur de prédiction diminue avec le nombre d'arbre, ensuite augmente légèrement et reste plus ou moins constant en augmentant le nombre d'arbres. Le minimum est atteint pour 120 arbres.
Le nombre de variables tirées aléatoirement à chaque nœud est optimal pour $mtries = 9$

Il est toujours intéressant de regarder comment ce modèle a pu modéliser le coût moyen aussi précisément, et donc quelles sont les variables qui vont avoir le plus d'influence sur la réponse. Pour cela, il est possible d'afficher l'importance des variables.

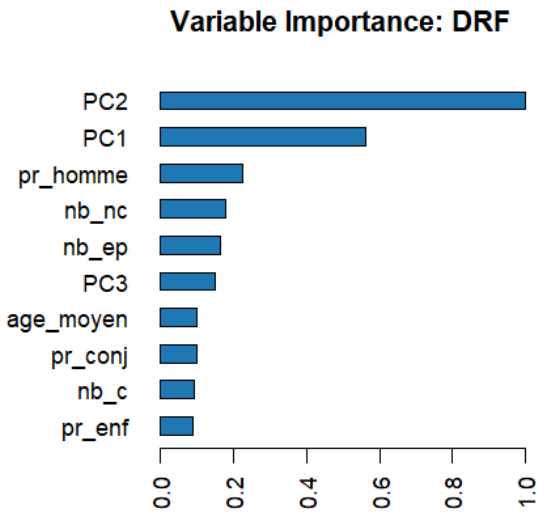


FIGURE 52 – L'importance des variables pour le randomForest

Nous observons que $PC1$, une variable expliquant 52% de la variance des variables externe, prédomine clairement. Elle est suivie de la deuxième variable qui résume les données externes, ce qui prouve le caractère non exhaustif des données interne pour décrire le coût moyen des prestations d'un département.

5.1.4 *Grandient Boosting*

En pratique, nous utilisons la fonction *gbm* du package *h2o* pour modéliser le coût moyen.

Fonction GBM du package *h2o*	
gbm (y=y.dep,x=x.dep, training_frame=,nfolds=, ntrees=,max_depth=,mtries=,learn_rate=)	
Paramètres	description
y	La variable à expliquer
x	Les variables explicatives
training_frame	La base d'apprentissage
nfolds	Le nombre de partions dans la validation croisée
ntrees	Le nombre d'arbres retenus
max_depth	La profondeur maximale de chaque arbre
learn_rate	Le taux d'aprentissage *"shrinkage"*

Pour le GBM, nous mettrons en application les étapes suivantes :

1. Nous lançons un premier modèle avec les paramètres par défaut calculer les métriques de comparaison sur ce modèle dans le but de le comparer au modèle optimisé :
   - $nfolds = 5$, partitions dans la validation croisée ;
   - $ntrees = 100$, arbres ;
   - $max\_depth = 5$, la profondeur maximale des arbres ;
   - $learn\_rate = 0.1$, le taux d'apprentissage.
2. Nous procédons à l'hyperparamétrage en testant des valeurs différentes pour les paramètres qui impactent la complexité de l'algorithme :
   - $ntrees \in \{10, 20, 30, ..., 300\}$;
   - la profondeur est limitée à $max\_depth \in$;
   - $learn\_rate \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$
3. Tracer comme pour les forêts aléatoires et une fois que les hyperparamètres ont été déterminés, l'évolution de la MSE en fonction du nombre d'arbres dans le but de faire un choix sur le paramètre $ntrees$ optimal à retenir.
4. Déterminer l'importance des variables ayant contribué à l'élaboration du modèle.
5. Mesurer la pertinence du modèle retenu.

Le modèle optimal que nous avons retenu à l'hyperparamétrage sont les suivants :

- 500 arbres de décision, en effet, l'erreur quadratique diminue considérablement avec le nombre d'arbres
- Une profondeure maximale de 5 pour les arbres.

En ce qui concerne, l'ordre d'importance des variables explicatives, nous retrouvons

Les caractéristiques principales sont les mêmes, à savoir la prédominance des variables externes PC2 et PC1. Néanmoins, nous observons que le sexe donné par la proportion d'homme et la catégorie socioprofessionnelle ont joué un rôle important que nous avions anticipé au début de l'étude.

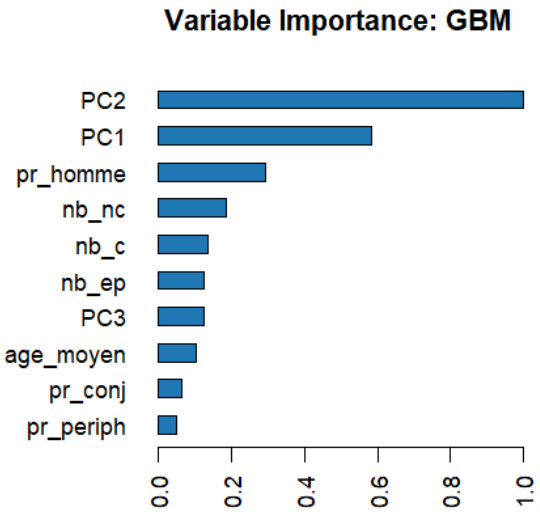


FIGURE 53 – L'importance des variables pour le gradient boosting

Cependant, il est essentiel de nuancer ces résultats. En effet, même si la hiérarchie des variables reste généralement inchangée, nous avons remarqué que les valeurs d'influence varient considérablement en fonction du modèle utilisé. Cela rappelle l'instabilité et l'irrégularité des arbres de décision. Compte tenu de la part d'aléatoire présente dans le modèle, il est fréquent de constater des changements de comportement significatifs lors de la comparaison de différents modèles ayant des performances équivalentes.

En outre, contrairement à la modélisation linéaire généralisée, la mesure de l'influence des variables explicatives n'est pas précise. En effet, l'interprétation de l'influence se limite à la contribution des variables au modèle et ne permet pas de préciser leur impact direct sur le coût moyen.

5.2 Comparaison des estimations

L'optimisation de chaque algorithme quant à elle permet de tirer le meilleur parti de la méthode utilisée.
L'amélioration apportée par celle-ci semble dérisoire lorsqu'on voit à quel point un changement de méthode peut modifier la qualité de prédiction, mais ce n'est pas le cas. En effet, l'avantage de l'optimisation d'un modèle est que celle-ci permet d'améliorer ses performances sans altérer les atouts qu'il possède. Un changement de méthode implique un changement de perspective, et donc des caractéristiques différentes.

	MSE
GLM	11.97178
GLM optimisé	9.855609
RandomForest	17.4192
RandomForest optimisé	13.97998
GBM	19.23063
GBM optimisé	9.25108

FIGURE 54 – Les erreurs de prédiction

En effet, malgré des performances prédictives différentes, chaque modèle possède son lot d'atouts qui le différencie des autres. Le GLM impose une structure à la relation entre la variable réponse et les variables explicatives, qui demande un certain travail de retraitement et de choix de paramétrisation. Cependant, une fois que cela est fait, l'interprétation et l'explication des résultats deviennent intuitifs et compréhensibles.

Le GBM propose une qualité de prédiction au-dessus des deux autres modèles. Néanmoins, la paramétrisation du modèle est assez lourde, une grande quantité de calculs est effectuée ce qui gêne fortement la traçabilité et la vérification du résultat.

5.3 Construction du zonier

Après avoir estimé le coût moyen pour isoler les effets géographiques dans les résidus, nous procéderons à une classification hiérarchique ascendante des résidus en utilisant la méthode présentée précédemment. Ce procédé présente l'avantage de déterminer le nombre de classes optimal à l'aide du critère du coude qui consiste à observer un décrochement dans la décroissance de la variance. Pour cela, nous nous intéressons aux variances associées au nombre de classes, ce qui donne le graphique suivant :

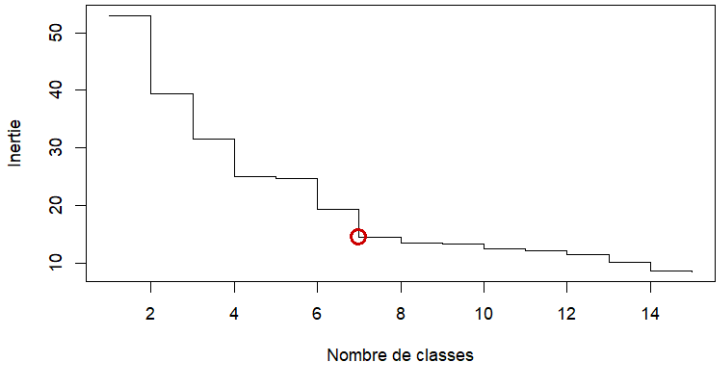


FIGURE 55 – Détermination du nombre de classes

Cette figure nous permet ainsi de déterminer le nombre de groupes optimal. Ici, un coude est observé pour 4 et 7 classes. Généralement, le point optimal est celui du nombre de classes à partir duquel la variance ne se réduit plus significativement. Nous avons choisi donc 7 classes.

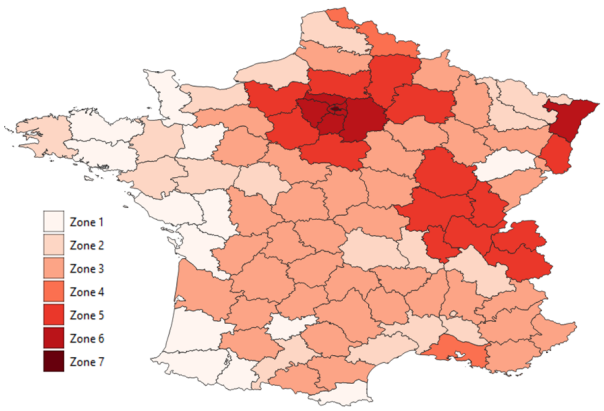


FIGURE 56 – Zonier de 7 classes

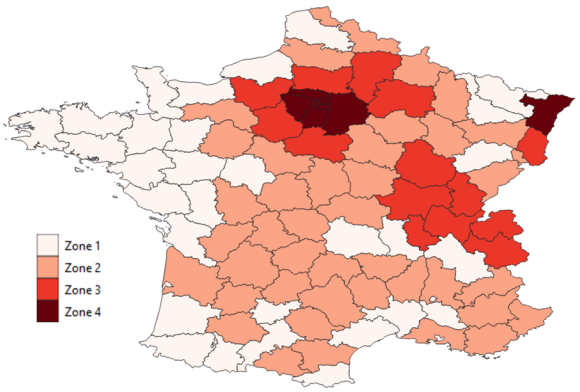


FIGURE 57 – Zonier de 4 classes

Nous avons réduit le nombre de groupes en fusionnant les classes ayant les couts moyens les plus proches afin de réduire le nombre de classes et de regrouper le département de Paris isolé. Ce résultat final présente des similitudes avec le zonier actuel dans lequel l'Île-de-France en zone forte et les départements de l'est dans la zone la plus basse.

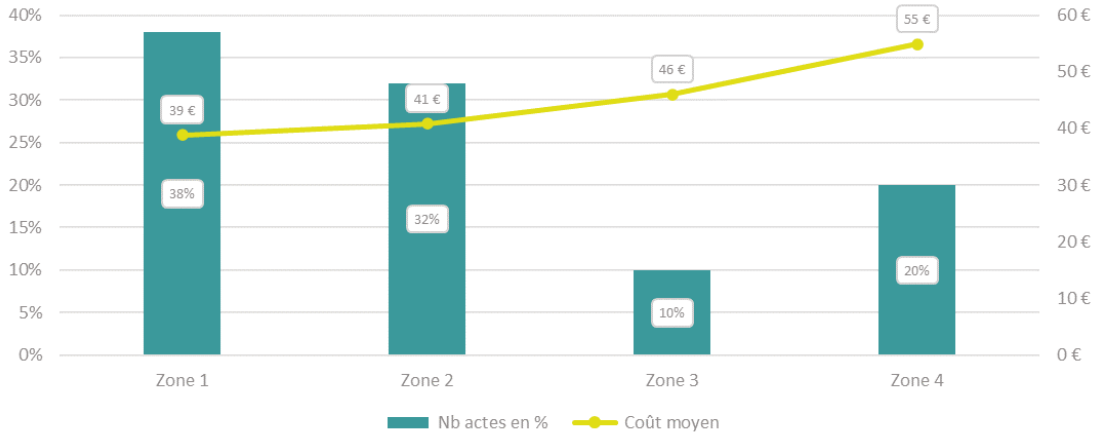


FIGURE 58 – Zonier de 4 classes

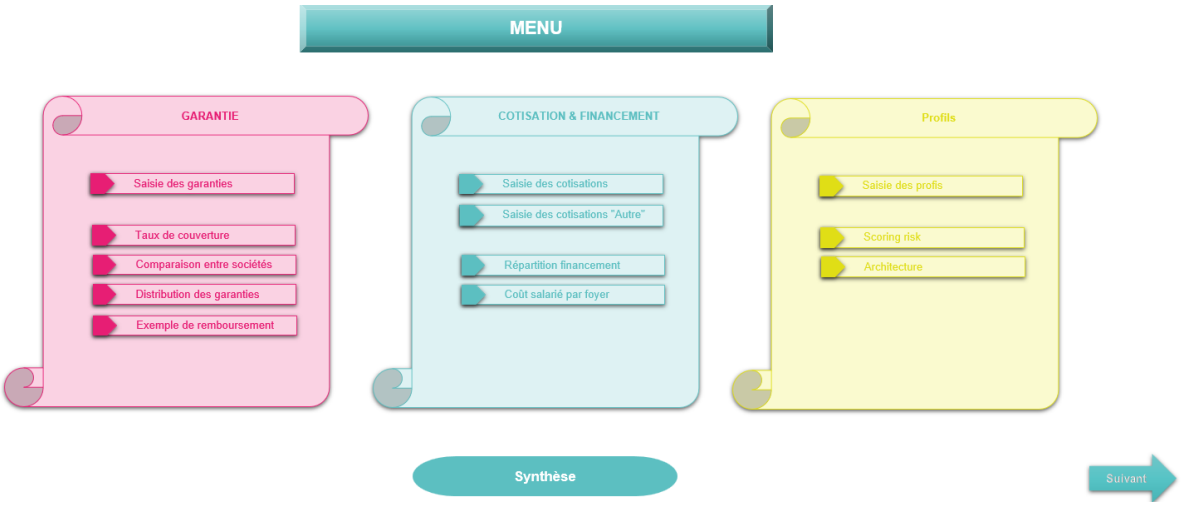
6 Quelles sont les applications de l'outil ?

Le zonier conçut est intégré dans l'outil par des réparations de coûts de prestation, en effet, dans chaque zone identifiée les coûts des prestations établis permettent de construire une table de répartition des montants de prestation pour chaque acte de soins. Ainsi le calcul, de taux de couverture moyen repose sur ces tables de répartition.

Cet outil présente plusieurs usages, sa principale fonction est de mesurer la performance d'une couverture santé en définissant un taux de couverture moyen qui peut également se lire comme le restant à charge moyen pour l'assuré. En effet, la quantité restant à la charge de l'assuré définit explicitement l'efficacité de son assurance complémentaire santé.

De plus, à travers l'analyse des garanties, ce dernier peut également servir de base de comparaison entre différents régimes, en particulier la comparaison d'un régime à ceux d'un groupe concurrent afin de le positionner sur le marché.

Dans le cas pratique que nous allons introduire, l'outil est un support pour l'harmonisation de plusieurs régimes d'un même groupe à travers les volets définis dans le menu.



6.1 Cas pratique : audit d'un régime santé

Cadre de l'étude :

Il s'agit d'un groupe composé de sept entités avec six régimes santé différents. Les entités du groupe sont réparties dans trois régions différentes et du fait de ce nombre important de couvertures santé, ce groupe souhaiterait une étude complète des différents régimes dont elle dispose ainsi qu'une proposition d'harmonisation des régimes.

Le groupe sur le plan démographique :

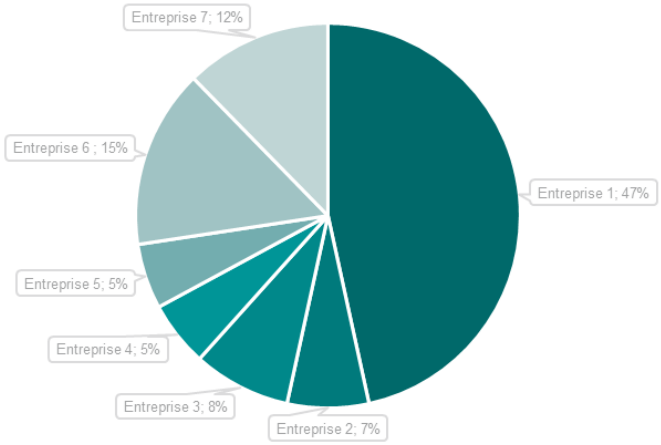


FIGURE 59 – La réparation des effectifs

Nous constatons une part importante des effectifs est comptabilisée dans l'entreprise 1. cette information nous conduit à préconiser une harmonisation vers une structure base avec une option pour se rapprocher de l'architecture de la majorité des effectifs. En effet, plus de 60% des salariés du groupe sont sur un régime base et une option facultative, c'est donc le schéma le plus représenté.

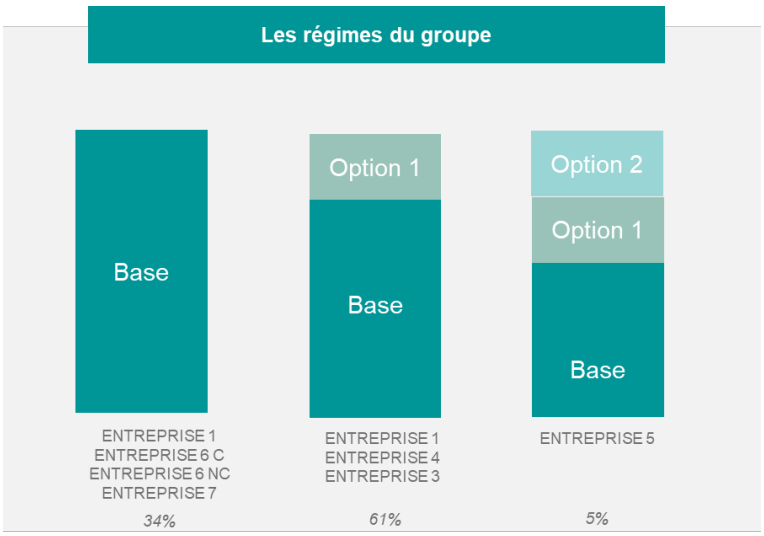


FIGURE 60 – Les architectures des différents régimes

6.1.1 *Analyse des garanties*

Chaque volet d'étude comporte une saisie d'information, en particulier pour les garanties, l'ensemble des niveaux de garantie des actes d'intérêt énumérés en figure 23 sont à saisir pour chaque entité en précisant l'architecture des régimes et ses garanties ainsi que la catégorie socioprofessionnelle et la situation géographique.

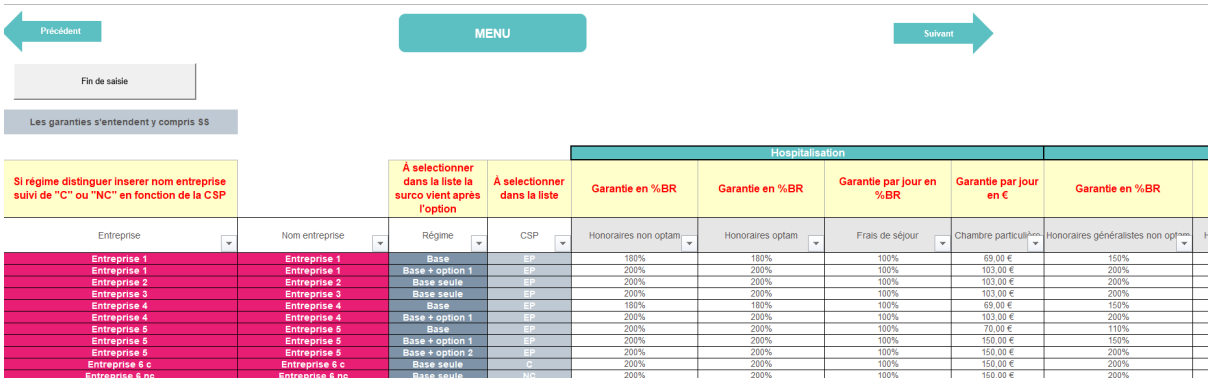
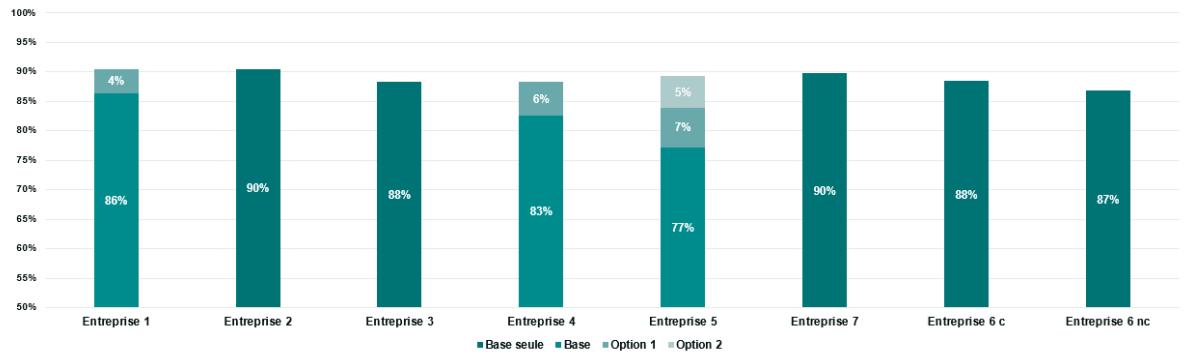


FIGURE 61 – La réparation des effectifs

À la fin de cette saisie, une macro est enclenchée par le bouton "Fin de saisie", cette dernière lance tous les calculs des différents taux de couverture accompagnés des graphes présentant les taux globaux, par poste et par actes de chaque entreprise.

L'analyse des garanties est initiée par une vision globale des régimes avec le figure ci-dessous qui présente les taux de couverture globaux. Nous constatons avec la figure suivante que les régimes base seule ont un socle plus important ceux avec options, mais nous relevons que sur l'ensemble des régimes le taux de couverture moyen est calculé à 87%



Note	Base	Base + Option 1	Base + Option 2	Base + Option 3	Base + Surco 1	Base + Surco 2	Base seule	Base (tout type de régime)	Base + toute option
Benchmark (EP)	84%	89%	89%	0%	0%	0%	89%	87%	89%

FIGURE 62

Passons à présent au niveau inférieur avec les taux de couverture moyen par poste. À ce niveau, les différences sont marquées pour les régimes bases. Nous constatons que plusieurs régimes présentent des lacunes sur les postes optique et dentaire, or, nous avons

montré que ce sont les postes avec le reste à charge le plus important. Bien que les options relèvent les taux de couverture sur ces postes, elles restent bien inférieures au taux de couverture d'autres postes.

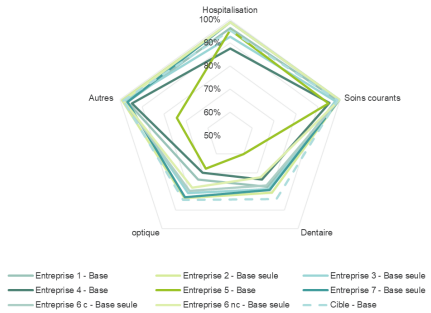


FIGURE 63 – Taux de couverture par poste des régimes "base"

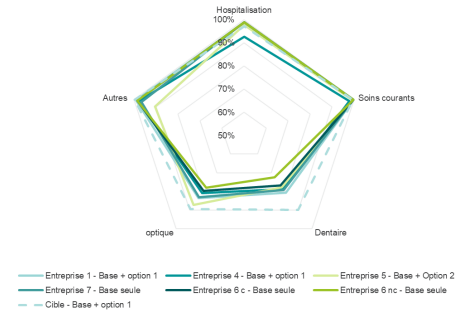


FIGURE 64 – Taux de couverture par poste des régimes base + option(s)

Concentrons-nous sur le domaine dentaire. Les deux figures ci-dessous présentent le niveau de couverture garanti pour les actes de prothèse maîtrisée et les implants dentaires. Nous constatons que ces deux actes sont très coûteux et que la plupart des régimes de base ne couvrent pas la totalité des frais, même pour le 1 quartile. Néanmoins, nous ne recommandons pas d'augmenter le niveau de couverture pour les raisons suivantes :

— Bien que les prothèses maîtrisées présentent encore un coût élevé pour les patients, le programme 100% santé prévoit de couvrir intégralement les frais pour certains types de prothèses répondant à des critères de matériaux spécifiques. Ce qui ne nécessite pas d'intervention sur ce type de prestation, il en est de même pour les prothèses libres et les inlays.

— En revanche, la situation est différente pour les implants dentaires, car cette prestation n'est pas du tout remboursée par la Sécurité sociale. Par conséquent, les assureurs jouent un rôle prépondérant dans cet acte. Toutefois, une augmentation de la couverture pourrait entraîner une hausse significative des tarifs des régimes.

Par ces différents graphiques établit par l'outil, nous pouvons procéder à une étude acte par acte en ciblant les poste de soins ayant des taux de couverture insuffisants.

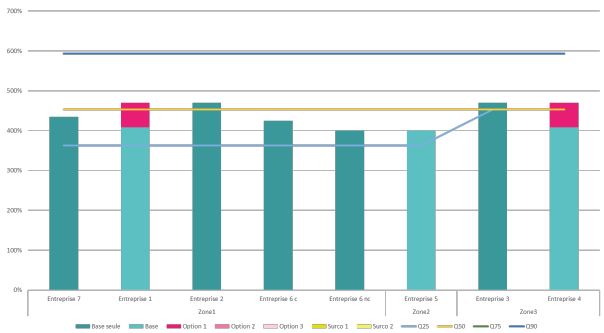


FIGURE 65 – Garantie des prothèses dentaires « maitrisés » par entité

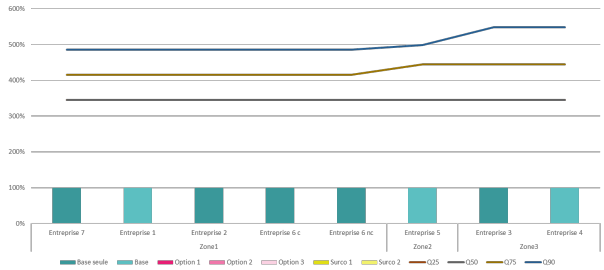


FIGURE 66 – Garantie des inlays par entité

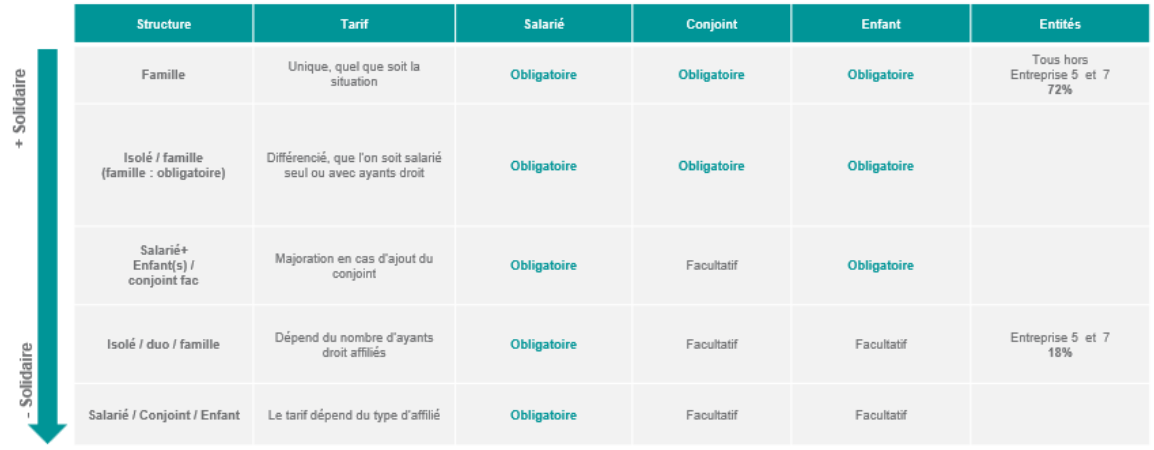
Dans notre cadre actuel, étant donné les contraintes financières, il est impossible de proposer des ajustements pour toutes les garanties présentant des lacunes. En ce qui concerne l'harmonisation, nous recommandons que toutes les entités optent pour le régime majoritaire, qui offre un compromis adéquat en termes de niveau de garantie et permet également de limiter les variations budgétaires.

6.1.2 Analyse des cotisations

De même que pour le volet des garanties, celui des cotisations et des financements est initié par la saisie des taux de cotisation en fonction de la structure de cotisation et financement employeur.

Rappelons que le financement employeur n'est présent que pour le régime obligatoire, c'est-à-dire la base. Les régimes base seule sont plus avantagés que la base des régimes avec option dont le niveau est révélé par les options qui sont à adhésion facultative et totalement à la charge du salarié.

Nous allons maintenant passer à l'analyse financière, qui commence par la liste des structures de cotisations existantes dans le groupe. Il convient de rappeler que l'employeur n'est obligé de prendre en charge qu'au minimum 50% de la cotisation des salariés, mais il a également la possibilité de financer la cotisation des ayants droit en fonction de la structure de cotisation choisie. Ainsi, pour toutes les structures solidaires qui rendent l'adhésion des ayants droit obligatoire, l'employeur participe à leur cotisation.



Structure	Tarif	Salarié	Conjoint	Enfant	Entités
Famille	Unique, quel que soit la situation	Obligatoire	Obligatoire	Obligatoire	Tous hors Entreprise 5 et 7 72%
Isolé / famille (famille : obligatoire)	Différencié, que l'on soit salarié seul ou avec ayants droit	Obligatoire	Obligatoire	Obligatoire	
Salarié+ Enfant(s) / conjoint fac	Majoration en cas d'ajout du conjoint	Obligatoire	Facultatif	Obligatoire	
Isolé / duo / famille	Dépend du nombre d'ayants droit affiliés	Obligatoire	Facultatif	Facultatif	Entreprise 5 et 7 18%
Salarié / Conjoint / Enfant	Le tarif dépend du type d'affilié	Obligatoire	Facultatif	Facultatif	

FIGURE 67 – Les structures de cotisation du groupe

Nous constatons que dans 72% des effectifs, l'employeur est solidaire dans le financement des ayants droit en proposant une cotisation unique pour chaque salarié quelle que soit sa situation familiale. Pour les entités 5 et 7 l'employeur cotise uniquement pour l'adhérent. Toutefois, pour l'entité 7, nous observons dans la figure 68 que l'employeur est solidaire à travers le financement étant pris en charge à 100%.

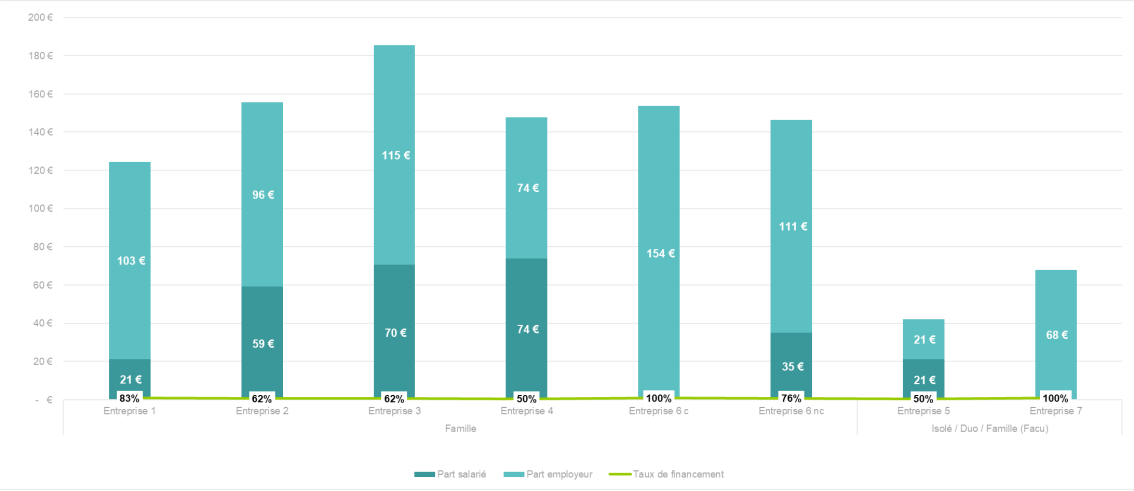


FIGURE 68 – Les cotisations mensuelles pour un salarié seul

6.1.3 Conclusion des analyses

Globalement, l'ensemble des régimes du groupe dispose de garantie avec un taux de couverture satisfaisant mis à part l'entité 5 dont la base est moins élevée avec une couverture moyenne de 77% néanmoins elle dispose de deux couches facultatives qui relève de niveau.

Toutefois, dès lors que nous passons au niveau des postes de soins, nous constatons que les postes optique et dentaire sont relativement bas, les options augmentent les taux de couverture pour l'optique, mais pas suffisamment sur le dentaire.

Pour les deux ou nous avons constaté des taux de couverture moins élevés dispose de panier 100% qui propose des prestations sans restant à charge. Ainsi, relever le niveau de certains actes de ces postes n'est pas nécessaire avec le 100% santé.

Pour minimiser, les variations budgétaires dans l'harmonisation, il est préconisé que toutes les entités passent pour le régime majoritaire, qui propose un bon compromis en termes de niveau de garantie et, mais également de limiter les variations budgétaires.

7 Conclusion

Dans ce mémoire, nous avons dans un premier temps présenté les éléments mis en place pour développer l'outil d'audit. Les dispositions apportées permettent plus de lisibilité dans l'étude et une meilleure interprétation de performance d'une couverture santé ainsi que l'automatisation de plusieurs calculs.

Dans un second temps, l'apport concret et technique de ce mémoire est principalement établi dans la modélisation du risque géographique synthétisé dans un zonier. Les étapes de zonage présenté sont basées sur des variables agrégées par département, les départements peu fiables présentant peu d'assurés sont fiabilisés par l'estimation des coûts moyens de prestations avec un lissage spatial basé sur la théorie de crédibilité. Ce procédé a permis de corriger les indicateurs aberrants tout en conservant la distribution générale du coût moyen.

Pour les estimations des coûts moyens de prestations sur la base des départements, nous avons comparé la méthode classique donnée par le modèle linéaire généralisé avec deux méthodes de machine learning utilisant des arbres de décisions : la forêt aléatoire et le gradient boosting. Nous avons observé des différences significatives entre ces méthodes.

En matière de qualité de prédiction, le gradient boosting est le modèle ayant les meilleures performances, en particulier avec les paramètres optimaux, suivi du GLM et enfin la forêt aléatoire. Cependant, l'écart entre ces trois modèles est significatif, et plus le modèle est performant, plus ses résultats sont difficiles à interpréter. En particulier, le gradient boosting nécessite un grand nombre de calculs pour ajuster le modèle en créant de nombreux arbres, ce qui complique la traçabilité des résultats. En revanche, le GLM calcule une fonction explicite qui permet de déterminer les prédictions, contrairement au gradient boosting qui ne produit qu'un résultat.

En ce qui concerne l'évaluation de l'influence d'une variable sur le coût de prestation, le GLM est la solution la plus pertinente parmi les trois modèles testés. En effet, il estime un coefficient pour chaque variable, qui évalue directement l'impact de celle-ci sur le coût moyen. Les deux autres modèles disposent également d'outils permettant d'évaluer l'influence d'une variable, mais leurs interprétations sont limitées et ne permettent pas de préciser l'impact qu'une variable a sur le coût.

Finalement, nous avons constaté que l'utilisation de techniques spécifiques au Machine Learning dans la modélisation rend les relations plus complexes et conduit à des résultats de meilleure qualité, au détriment de leur interprétabilité, c'est particulièrement le cas du gradient boosting.

Bibliographie

https ://drees.solidarites-sante.gouv.fr/sites/default/files/2022-02/rapport_oc_2021.pdf

—

— Mémoire Actuariat : Modélisation du risque géographique en Santé, pour la création d'un nouveau zonier. Comparaison de deux méthodes de lissage spatial de Catalina SEPULVEDA

— Cours de Théorie de la crédibilité d'Olivier Wintenberger

— Méthodes d'agrégation : boosting, bagging et forêts aléatoires

— Bagging vs Boosting

— Mémoire Actuariat : Une approche individuelle du provisionnement des sinistres corporels automobiles de Fatima-Zohra Zouggagh

— Théorie de la crédibilité de Pierre Thérond

— Projection des dépenses de santé à l'horizon 2060,le modèle PROMEDE, Ministère de Santé