IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la Formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Nicolas RIHOUEY

Titre : Application des méthodes d'apprentissage à la modélisation de la prime pure en santé collective

Confidentialité : ☐ NON ☑ OUI (Durée : ☐ 1an ☑ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

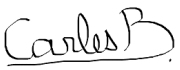
Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : AVIVA ASSURANCES
Nom : Anne-Sophie VERSCHAVE
Signature et Cachet :

Directeur de mémoire en entreprise :
Nom : Brice CARLES
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(s) du candidat(s)

SAS au capital de 96 550 euros - Déclaration d'activité enregistrée sous le n° 11 75 44451 75 auprès du Préfet de Région IDF
4, rue Chauveau-Lagarde - 75008 Paris - Tél : 01 44 51 72 79 - Fax : 01 44 51 72 73
Email : formation@institutdesactuaires.fr
Siret : 512 264 441 000 19 - Code NAF : 8559 A - TVA intracommunautaire : FR 25 512264441 - RCS Paris : 512 264 441

Résumé

Les évolutions législatives fréquentes et la concurrence exacerbée sur le marché de la santé collective obligent les organismes de complémentaires à ajuster régulièrement leurs tarifs afin d'intégrer les changements de comportement des assurés et préserver leur rentabilité. Ce contexte implique également de veiller à ce que les approches utilisées pour l'élaboration des tarifs soient performantes et bien appropriées au risque assuré.

L'approche classique repose sur l'utilisation des modèles linéaires généralises (GLM). Cette méthode présente certains avantages comme la lisibilité du tarif mais aussi certaines contraintes théoriques liées à la distribution des données. Par ailleurs, elles ne permettent pas toujours de prendre pleinement en considération les interactions entre les variables. Le recours aux algorithmes d'apprentissage statistique s'intensifie avec l'augmentation de l'utilisation des données externes car ils permettent de capter ces effets sans à avoir à les spécifier et n'imposent pas de propriétés aussi fortes sur la structure des données.

L'objectif de ce mémoire est de tester l'utilisation de deux méthodes d'apprentissage statistique, CART et Random Forest (forêts aléatoires) et de comparer leurs performances avec une modélisation à l'aide des GLM. Des tests de sensibilité sur la performance consistant à faire varier le type d'approche (fréquence-coût moyen ou coût total) et certains paramètres des algorithmes CART et Random Forest seront réalisés. Une analyse complémentaire sur la réforme du 100% santé est également menée afin de mesurer ses premiers effets sur les postes optique et dentaire.

Mots clés : Tarification, Complémentaire Santé collective, Modèle linéaire généralisé, GLM, CART, forêts aléatoires

Abstract

Frequent legislative changes and fierce competition in the collective healthcare market force complementary organizations to regularly adjust their rates in order to integrate changes in policyholder behaviors and preserve their profitability. This context also implies ensuring that the approaches used for the development of tariffs are efficient and well match to the insured risk.

The classical approach relies on the use of generalized linear models (GLM). This method has some advantages such as policyholders readability as well as some theoretical constraints linked to the distribution of the data. Moreover, they do not always allow to fully take into account interactions between variables. The use of statistical learning algorithms increases with the raise of external data usage as they allow to captured these effects without having to specify them nor impose such strong properties on the data structure.

The objective of this thesis is to test the use of two statistical learning methods, CART and Random Forest and to compare their performances with a modeling using GLMs. Sensitivity tests on performance consisting in varying the type of approach (frequency-average cost or total cost) and some parameters of the CART and Random Forest algorithms will be carried out. An additional analysis on the 100% health reform is also being carried out in order to measure its first effects on the optical and dental care.

---

key words : Pricing, Collective complementary insurance policies, generalised linear model, GLM, CART, Random Forest

Remerciements

Tout d'abord, je tiens à remercier Fabienne GOURINCHAS directrice de la Data Factory, qui m'a permis d'intégrer la formation du CEA.

Je remercie Naoufal RAKAH, responsable du service actuariat Santé, pour ses conseils lors de la réalisation de ces travaux.

Un grand merci également à Brice CARLES pour son accompagnement et ses relectures attentives lors de la finalisation de ce mémoire.

Plus généralement, j'adresse mes remerciements à toutes les personnes qui m'ont aidé et soutenu dans le cadre de ce travail.

Enfin, je remercie ma famille et plus particulièrement ma compagne Elodie pour sa compréhension et son soutien pendant la réalisation de ce mémoire.

Table des matières

Introduction

Le marché de la complémentaire santé collective est un marché extrêmement concurrentiel. En 2019, 87% des cotisations ont été reversées sous forme de prestations aux assurés. C'est également un marché très réglementé, plusieurs grandes réformes ont eu lieu ces dernières années avec notamment la mise en place du contrat responsable, la généralisation de la complémentaire santé pour les salariés et plus récemment la réforme du 100% santé. Ce contexte oblige les acteurs de ce marché à revoir régulièrement leurs tarifs afin de les affiner et de les adapter aux évolutions règlementaires s'ils veulent rester compétitifs et rentables.

La prime pure constitue le socle de l'élaboration du tarif, sa finesse et sa justesse conditionnent la rentabilité future du portefeuille. La modélisation de la prime représente donc un enjeu majeur. Les méthodes de modélisation de la prime pure sont nombreuses, la plus fréquemment utilisée reposant sur la théorie des modèles linéaires généralisés (GLM). Cette méthode apporte une facilité de lecture du tarif et ses bonnes performances lui confère aujourd'hui encore une place de premier plan. Cependant, elle comporte certaines contraintes parfois difficile à satisfaire, notamment l'adéquation des données à des lois théoriques. D'autre part, le caractère multiplicatif du tarif résultant de la modélisation n'est pas toujours adapté. Il est peu probable que chaque facteur tarifaire joue avec la même intensité indépendamment des autres caractéristiques du souscripteur.

Les méthodes d'apprentissage statistique non paramétriques constituent une alternative aux modèles linéaires généralisés. Elles permettent de s'affranchir des hypothèses fortes de distribution de la variable à expliquer et de capter les interactions entre les données sans avoir à les spécifier. La méthode CART (Classification And Regression Trees) est l'une des méthodes d'apprentissage les plus populaires. Développée dans les années 80, cette méthode représente une réelle amélioration des techniques d'arbre de décision. Elle permet de répondre à de multiples problématiques de classification et de régression et facilite l'interprétation des résultats. Elle est donc adaptée à des travaux de tarification mais présente un caractère instable, elle est très sensible à l'échantillon de données. Pour palier à cette faiblesse, des méthodes d'agrégation de modèles ont été développées. Les forêts aléatoires ou Random Forest dans les années 2000 ont apporté une solution à ce problème d'instabilité en associant une stratégie d'agrégation de modèles et une dimension aléatoire dans l'élaboration de chacun des modèles.

L'objectif de ce mémoire est de tester l'utilisation de ces deux méthodes d'apprentissage non paramétriques, CART et Random Forest et de comparer leur performance avec l'approche classique des modèles linéaires généralisés. Cette analyse sera menée sur un portefeuille d'assurés santé collective ayant la particularité d'être relativement récent. Les changements législatifs sont parfois générateurs d'opportunité, ce portefeuille en est l'illustration. Il a été principalement développé à la suite de la loi ANI qui a obligé les entreprises à proposer à leurs

salariés une complémentaire santé à partir du 1^{er} janvier 2016. Cette étude sera l'occasion d'affiner la connaissance du portefeuille et d'identifier d'éventuels ajustements nécessaires sur le tarif en vigueur. Par ailleurs, le contexte règlementaire actuel avec la réforme du 100% santé va modifier le comportement des assurés, une étude d'impact de cette réforme sera également menée.

La premier chapitre de ce mémoire sera consacré à la présentation du contexte de l'étude. Le fonctionnement du système de santé en France et les mécanismes de remboursements des soins y seront détaillés. Après une revue des dernières évolutions réglementaires, une attention plus particulière sera portée sur la réforme en cours du 100% santé. Les grands changements liés à cette réforme y seront exposés.

Le second chapitre sera dédié à l'analyse préliminaire des données. Cette étape est fondamentale lors de la réalisation de travaux de modélisation. Elle consiste à procéder aux retraitements des données et à des analyses univariées et bivariées permettant d'avoir une première vision de la capacité prédictive des variables. Le risque géographique est une composante importante du niveau de consommation médicale. Une analyse plus particulière sera menée sur cet effet avec l'élaboration d'un zonier par poste de soins (hospitalisation, soins courants, pharmacie, dentaire, optique et bien-être).

Le troisième chapitre sera consacré à la modélisation de la prime pure par poste de soins. Les fondements théoriques des différentes méthodes GLM, CART et Random Forest seront d'abord présentés et ensuite une mise en application sur le portefeuille sera réalisée.

Le quatrième chapitre portera sur l'analyse des résultats et une étude des impacts de la réforme du 100% santé. Cette étude sera menée à partir des données observées sur les dix premiers mois de l'année 2020.

Chapitre 1

Contexte de l'étude

1.1 Le système de santé en France

Le système de santé en France s'organise autour de deux piliers : l'assurance maladie obligatoire, première branche de la sécurité sociale et l'assurance maladie complémentaire gérée par des organismes privés. En 2017, selon l'étude « La complémentaire santé - Acteurs, bénéficiaires, garanties » (DREES, 2019), l'assurance maladie obligatoire représente 155,1 milliards d'euros de prestations, soit 77,8% de la consommation de soins et de biens médicaux (CSBM) tandis que l'assurance complémentaire représente 13,2% avec 26,3 milliards d'euros. Les ménages représente 7,5% et les 1,5% restants sont financés par l'état et la couverture maladie universelle complémentaire (CMU C).

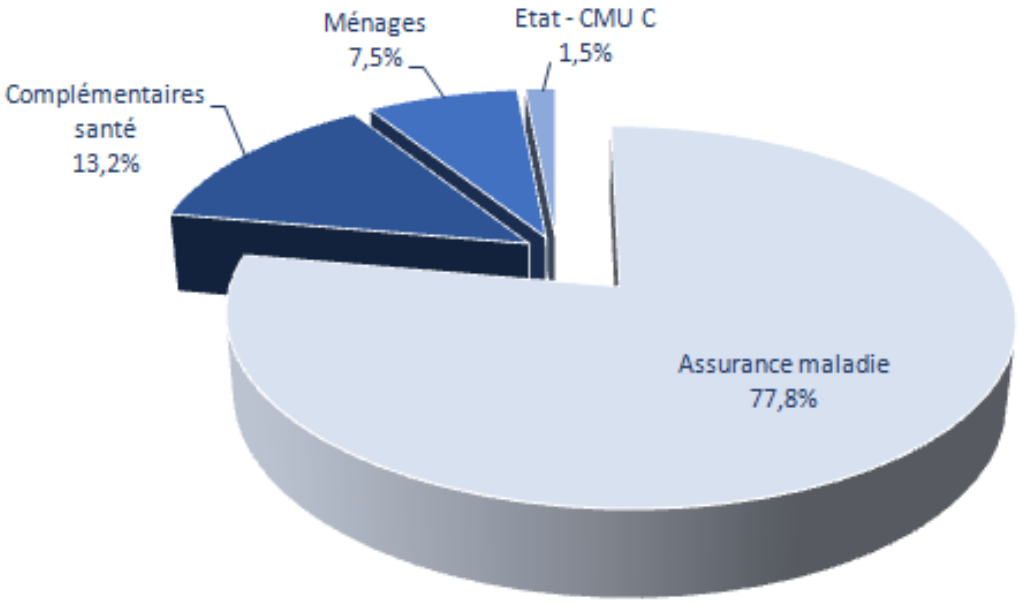


FIGURE 1.1 – Répartition du financement de la consommation médicale en 2017 (DREES, 2019)

1.1.1 Le premier pilier : l'assurance maladie obligatoire

L'assurance maladie obligatoire est l'une des 4 branches de la sécurité sociale. Les trois autres branches sont la branche famille, la branche accidents du travail et maladies professionnelles et la branche retraite. Toutes les personnes qui travaillent ou résident en France ont l'obligation de s'affilier à la sécurité sociale. Cette obligation remonte à 1945, date de la mise en place de la sécurité sociale visant à couvrir l'ensemble de la population. L'assurance maladie est le premier pilier du système de santé. Sa mission consiste à assurer la prise en charge des dépenses de santé des assurés et à garantir l'accès aux soins. Elle a également un rôle de prévention et contribue à la régulation du système de santé. Elle représente 49% du budget de la sécurité sociale.

Elle est constituée de 2 régimes principaux :

- **le régime général** géré par la caisse d'assurance maladie. Il concerne les salariés du secteur privés ainsi que les travailleurs indépendants suite à la suppression régime social des indépendants (RSI) voté en 2018. Ce régime couvre 88% de la population française.

- **le régime agricole** géré par la caisse centrale de la mutualité sociale agricole (MSA). Il couvre les exploitants agricoles, les salariés agricoles et les entreprises de travaux agricoles. Ce régime concerne 5% de la population.

Il existe également des régimes spéciaux. Ils concernent les fonctionnaires, la SNCF, EDF-GDF, les employés et clercs de notaires etc... On en recense 27 et ils couvrent 7% de la population française. Le régime Alsace-Moselle est l'un de ces régimes spéciaux et ne concerne pas une profession particulière. Il s'adresse à l'ensemble des habitants de l'Alsace-Moselle. Il intervient en complément du régime général des salariés, il est donc plus avantageux que le régime de la sécurité sociale. Ce régime est une information à prendre en compte dans la tarification car pour les assurés dépendant de ce régime, la prise en charge par le régime obligatoire est plus importante et la part potentielle prise en charge par la complémentaire santé s'en trouve ainsi réduite.

1.1.2 Le second pilier : l'assurance maladie complémentaire

Les complémentaires santé constituent le second pilier du système de santé français. Elles viennent compléter la couverture de l'assurance maladie obligatoire. La couverture de base ne rembourse pas l'intégralité des dépenses de santé, il reste une partie à la charge de l'assuré que la complémentaire va rembourser partiellement ou totalement. Le niveau de remboursement de la complémentaire sera fonction des garanties définies dans le contrat souscrit par l'adhérent. En 2017, elles ont versé 26,3 milliards de prestations en soins et biens médicaux, cela représente 13,2% des dépenses de santé. Elles financent principalement l'optique (73,1% de la dépense optique totale) et le dentaire (40,9%). Elles contribuent également fortement au financement des soins de ville (médecins, auxiliaires médicaux, laboratoires d'analyse) (16,1%) et les médicaments (12,6%). Concernant les soins hospitaliers, leurs contributions est beaucoup plus réduite avec 5,1% de la dépense, ce type de soins étant largement remboursé par l'assurance maladie.

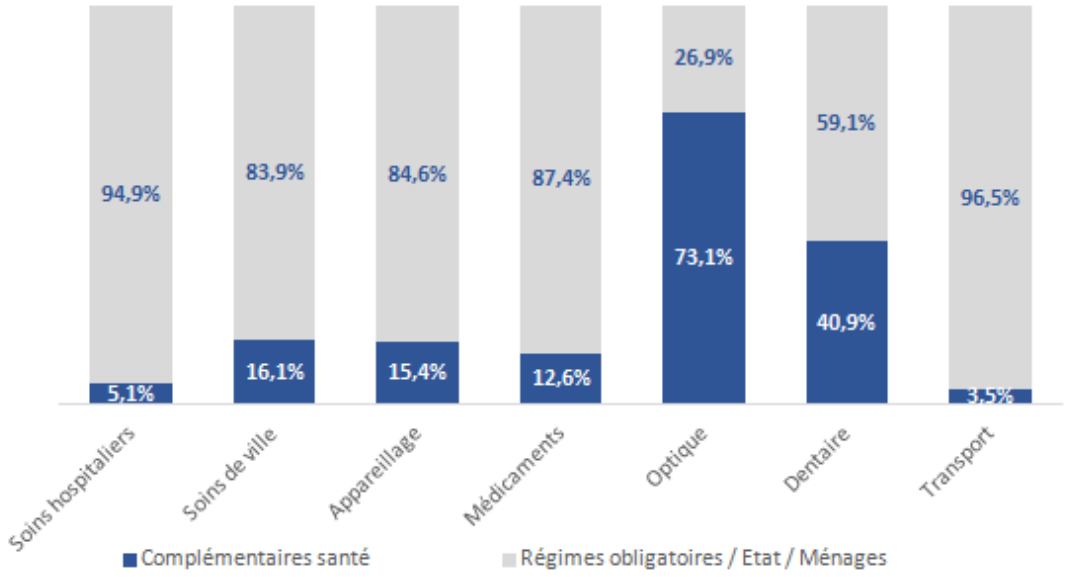


FIGURE 1.2 – Pourcentage de la consommation médicale par poste financée par les complémentaires santé en 2017 (DREES, 2019)

1.1.3 Les mécanismes de remboursement des frais de soins

La complémentaire santé comme son nom l'indique intervient en complément du régime obligatoire. Pour déterminer les montants remboursés par la complémentaire santé, il est nécessaire de bien comprendre les mécanismes de remboursement du régime obligatoire.

1.1.3.1 Mode de calcul du montant remboursé par le régime obligatoire

Le calcul des remboursements des soins médicaux par le régime obligatoire fait intervenir plusieurs paramètres :

-La Base de Remboursement de la Sécurité Sociale (BRSS)
-Le taux de remboursement de la sécurité sociale
-La participation forfaitaire / la franchise médicale

La Base de Remboursement de la Sécurité Sociale (BRSS) nommée parfois « tarif de convention » désigne le tarif de référence d'un acte médical. Son montant est fixé en concertation par l'État, la caisse nationale d'assurance maladie et les syndicats professionnels de santé. Ce montant ne correspond pas toujours au prix réel de l'acte, c'est notamment le cas lorsque les praticiens pratiquent des dépassements d'honoraires.

Le taux de remboursement de la sécurité sociale définit le pourcentage de la base qui sera remboursé par le régime obligatoire. Ce taux de prise en charge est déterminé pour chaque acte médical et peut varier selon les individus, en raison de leur situation médicale par exemple. Les personnes en affection longue durée (ALD) peuvent être exonérées du ticket modérateur, c'est-à-dire qu'elles seront remboursées à 100% de la base de remboursement.

La participation forfaitaire est une participation financière qui s'applique à toutes les consultations ou actes réalisés par un médecin, mais également sur les examens radiologiques et les analyses de biologie médicale. Elle concerne les personnes âgées de plus de 18 ans et son montant total est plafonné à 50 euros par an et par assuré.

La franchise médicale sur le même principe que la participation forfaitaire s'applique aux médicaments. Son montant est de 50 centimes par boite de médicament et elle est déduite du montant remboursé.

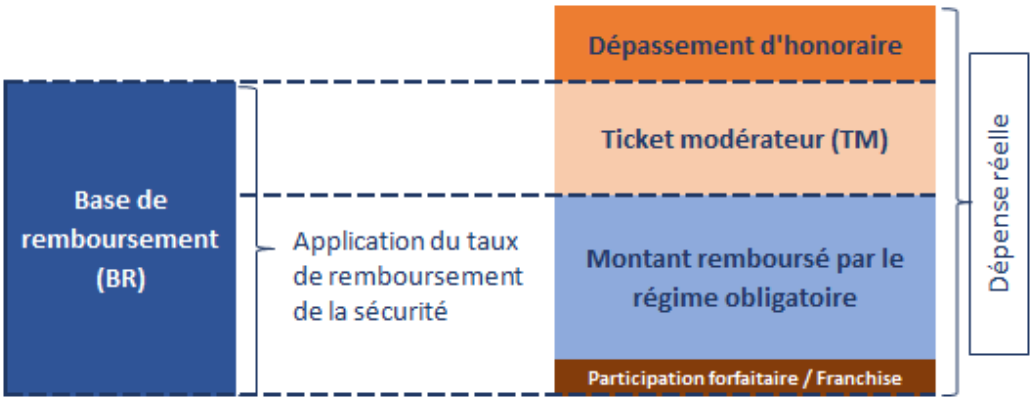


FIGURE 1.3 – Schéma de remboursement par l'assurance maladie

Formule du calcul du montant remboursé par la sécurité sociale :

*Montant remboursé = Base de remboursement * Taux de prise en charge – Participation forfaitaire / Franchise*

1.1.3.2 Mode de calcul du montant remboursé par la complémentaire frais de soins

La complémentaire santé intervient en complément de la sécurité sociale mais elle peut aussi rembourser des actes non pris en charge par celle-ci. Il existe différentes manières de formuler les niveaux de garantie dans les contrats.

Voici les formulations les plus courantes :

Le remboursement en pourcentage des Frais Réels (%FR) : le montant remboursé par la complémentaire sera un pourcentage de la dépense réelle auquel sera déduit le montant remboursé par le régime obligatoire et les éventuelles participations forfaitaires et franchises.

Le remboursement en pourcentage de la base de remboursement (%BR) : le calcul prend en compte le produit de la base de remboursement par le taux de la garantie auquel sera déduit le montant remboursé par le régime obligatoire et les éventuelles participations forfaitaires et franchises.

Le remboursement en pourcentage du plafond mensuel de la sécurité sociale (%PMSS) : il s'agit d'un montant annuel maximum que la mutuelle rembourse. En 2020, le PMSS est de

3 428 euros, une garantie à 10% du PMSS sera équivalente à 342,80 euros.

Le remboursement au forfait : il s'agit d'un montant maximum en euros et cela concerne souvent les frais optiques ou les garanties des postes bien-être non pris en charge par le régime obligatoire.

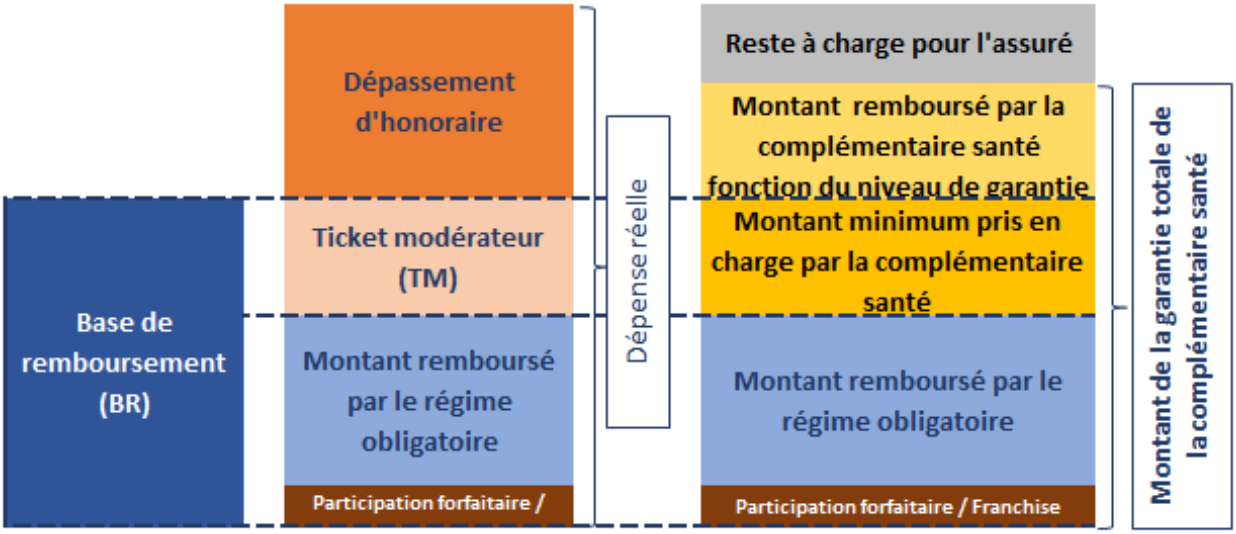


FIGURE 1.4 – Schéma de remboursement par la complémentaire santé

1.1.4 Les complémentaires santé collective

1.1.4.1 Qu'est ce qu'un contrat santé collectif ?

Un contrat santé collectif est un contrat souscrit entre une compagnie d'assurance et un employeur afin d'offrir aux salariés une couverture complémentaire santé. La notion de collectif sous entend que ce contrat concerne l'ensemble des salariés de l'entreprise.

1.1.4.2 Les spécificités de la santé collective

L'employeur est le souscripteur du contrat santé collective. Il peut définir seul ou en concertation avec ses salariés le niveau de garantie du contrat. Il a l'obligation d'opter pour un contrat respectant les minimas de garanties imposées par la convention collective.

L'entreprise a également la possibilité de proposer des garanties différentes aux salariés en fonction de leur appartenance à une catégorie objective de salariés. Les catégories objectives de salariés sont généralement les cadres, les non cadres ou l'ensemble du personnel. Si l'employeur souhaite faire une distinction entre les cadres et les non cadres, les garanties seront définies pour chacun des collèges cadres et non cadres. Dans le cas contraire, les garanties seront définies sur le collège ensemble du personnel.

Une autre particularité de la santé collective est le financement de la cotisation. Une partie de la cotisation, au minimum 50%, est payée par l'entreprise.

Les contrats santé collective proposent également différentes structures tarifaires. Le choix de la structure est défini au niveau du collège de salariés et s'applique à tous les salariés qui y sont rattachés. Les structures les plus souvent proposées sont la structure adulte/enfant, la structure isolé/duo/famille et la structure unique.

Voici les particularités des différentes structures tarifaires :

- structure adulte/enfant : les salariés ont la possibilité d'affilier des ayant-droits avec lesquels ils ont un lien de parenté de type conjoint ou enfant. La cotisation des ayant-droits ne comporte pas de participation de l'employeur.

- structure isolé/duo/famille : les salariés ont la possibilité de choisir l'une des 3 formules en fonction de leur situation familiale et de leur souhait. S'ils ne souhaitent pas affilier de membre de leur famille, ils peuvent opter pour la formule isolé. S'ils souhaitent affilier un enfant ou son conjoint, ils peuvent opter pour la formule duo. Enfin, si ils souhaitent affilier deux membres ou plus de leur famille, ils peuvent choisir la formule famille.

- structure unique : les salariés peuvent affilier les membres de leur famille gratuitement.

Les assurés ont également la possibilité de souscrire des renforts individuels si les garanties du contrat proposé par l'employeur ne leur paraissent pas suffisantes. La cotisation de ces renforts est entièrement à la charge du salarié.

1.2 La santé, un marché très concurrentiel avec un cadre législatif en constante évolution

1.2.1 Un nombre important d'acteurs opèrent sur ce marché

La concurrence sur le marché de la santé est très forte notamment du fait du nombre important d'acteurs présents sur ce marché. En 2017, il y avait 474 organismes de complémentaires santé dont 346 mutuelles, 103 sociétés d'assurances et 25 institutions de prévoyance.

Cependant, depuis plusieurs années, un mouvement de concentration du marché est en place, avec une baisse du nombre d'organismes. En effet, la forte concurrence et les contraintes règlementaires en terme de solvabilité obligent les acteurs à se regrouper. Le graphique de la figure 1.5 ci-après issu de l'étude « La complémentaire santé - Acteurs, bénéficiaires, garanties » (DREES, 2019) donne une évolution du nombre d'organismes de 2001 à 2017. Leur nombre était de 1702 en 2001, soit 72% d'acteurs en moins sur la période 2001-2017. Cette tendance touche principalement les mutuelles dont le nombre a été divisé par quatre pendant la période et les institutions de prévoyance dont le nombre a été divisé par 2.

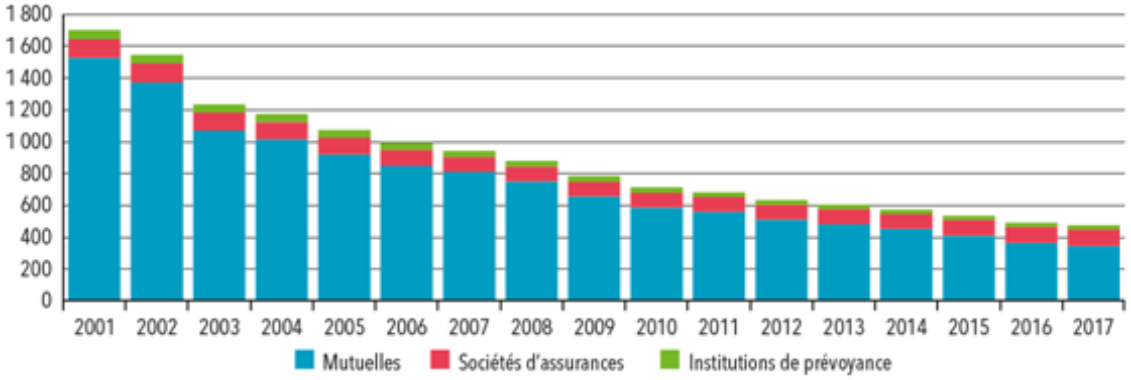


FIGURE 1.5 – Évolution du nombre d'organismes de complémentaire santé de 2001 à 2017 (DREES, 2019)

1.2.2 Les évolutions législatives récentes

Le marché de la complémentaire santé est essentiellement géré par des organismes privés néanmoins c'est un marché fortement régulé. Plusieurs réformes majeures ont eu lieu ces dernières années, la mise en place des contrats responsables, l'accord ANI avec la généralisation de la complémentaire santé pour les salariés et plus récemment la réforme du 100% santé.

1.2.2.1 Le contrat responsable

Le contrat responsable est entré en vigueur le 1^{er} janvier 2006, il a ensuite évolué dans le cadre d'une nouvelle réforme au 1^{er} janvier 2015. Ce contrat a été mis en place par les pouvoirs publics avec l'objectif de réduire le déficit de la sécurité sociale en encourageant les assurés à maîtriser leurs dépenses de santé et à respecter le parcours de soins coordonnés. Ce contrat bénéficie d'aides fiscales et sociales à condition de respecter un cahier des charges précis en terme de garantie. Pour être responsable, le contrat doit respecter des minimas et des plafonds de garantie mais également ne pas prendre en charge certains frais.

Le contrat responsable doit couvrir l'intégralité du ticket modérateur à la charge de l'assuré pour les soins de ville. Ces soins concernent notamment les consultations et actes réalisés par les professionnels de santé (médecin généraliste, dentiste, ophtalmologue, auxiliaires médicaux,...). Il doit également couvrir l'intégralité du forfait journalier hospitalier sans limitation de durée Il doit couvrir les frais d'optique soit à hauteur de 100% du ticket modérateur, soit par un forfait en euros.

1.2.2.2 La loi ANI

La loi ANI adoptée définitivement par le parlement le 13 mai 2013 est la retranscription dans la législation de l'accord national interprofessionnel (ANI) qui a été conclu entre les partenaires sociaux le 11 janvier 2013. Cet accord se traduit par un certain nombre d'avancées concernant les droits sociaux des salariés avec notamment la généralisation de la couverture complémentaire santé à l'ensemble des salariés. En effet, depuis le 1^{er} janvier 2016, toutes les entreprises à partir d'un salarié et un dirigeant doivent proposer une complémentaire santé à leurs salariés dans le cadre d'un contrat collectif et obligatoire.

Concernant le financement de cette couverture, la loi impose l'obligation pour l'employeur de prendre à sa charge au moins 50 % de la cotisation, la partie restante étant financée par le salarié. Elle impose également que la complémentaire santé proposée par l'entreprise respecte le panier de soins, c'est à dire un niveau de couverture minimal pour différentes garanties.

1.2.2.3 La nouvelle réforme du 100% santé

La réforme du 100% santé est une réforme visant à améliorer l'accès au soins et à réduire le renoncement aux soins pour raison financière. Elle entre en vigueur de manière progressive à partir du 1^{er} janvier 2019 et concerne les postes optique, dentaire et audio-prothèse. Les détails de cette réforme seront présentés dans la section suivante.

1.3 La réforme du 100% santé en détail

La réforme du 100% santé a été initiée en partant du constat qu'un trop grand nombre de français renonçaient pour des raisons financières à changer de lunettes, à se faire poser une prothèse dentaire ou à s'équiper d'une aide auditive. Cette réforme est un engagement du président de la république lors de la campagne de 2017. L'objectif est de permettre à tous les français de bénéficier d'un ensemble de prestations de soins et d'équipement en optique, dentaire et audio prothèse intégralement pris en charge par le régime obligatoire et la complémentaire santé. Tous les français affiliés à une complémentaire santé responsable ont la possibilité de bénéficier de l'offre 100% santé. En optant sur des prestations du panier 100% santé, les assurés n'auront plus de reste à charge mais ils ont toujours la possibilité de choisir des soins en dehors de ce panier.

La réforme du 100% santé a commencé à entrer en vigueur le 1^{er} janvier 2019. Son déploiement est progressif sur trois années 2019, 2020 et 2021. La réforme modifie également les critères du contrat responsable et concerne à la fois les contrats individuels et collectifs. A partir du 1^{er} janvier 2020, pour continuer à être responsable, un contrat a l'obligation de proposer l'offre de soins et d'équipement du panier 100% santé.

1.3.1 La réforme 100% santé en optique

La réforme du 100% santé en optique concerne les montures et les verres. Elle est entrée en vigueur le 1^{er} janvier 2020 et oblige les opticiens à proposer une gamme d'équipement 100% santé composée d'une sélection de lunettes de vue (monture et verres) intégralement remboursables et dont les tarifs sont plafonnés.

Avec la réforme, deux classes d'équipement se distinguent :

Classe A (Panier 100% Santé) :

Montures : il s'agit d'une gamme de montures répondant aux normes de qualité européennes. Chaque opticien doit disposer d'au moins 17 modèles adultes et 10 modèles enfants. Ils doivent être disponibles en 2 coloris différents. Leur prix limite de vente (PLV) est inférieur ou égal à 30 €.

Verres : leur tarif est plafonné (PLV) et ils sont mieux remboursés par l'assurance maladie. Ils couvrent tous les besoins de correction visuelle et répondent à des critères de qualité à la fois esthétiques (amincissement obligatoire) et techniques (traitement anti-rayures et anti-reflet). Selon la correction, le plafond de remboursement total pourra aller jusqu'à 800 € (monture + verres).

Classe B (Tarifs libres) :

Montures : elles sont prises en charge selon les conditions définies par le contrat de complémentaire santé, dans la limite de 100 € (contre 150 € avant la réforme).

Verres : les bases de remboursement de l'assurance maladie sont abaissées à 0,05 euro. La complémentaire santé intervient dans la limite des garanties du contrat.

1.3.2 La réforme 100% santé en dentaire

La réforme de 100% santé en dentaire est entrée en vigueur le 1^{er} janvier 2020 et sa mise en œuvre s'étale sur trois ans 2020, 2021 et 2022. Elle a pour objectif de proposer une large gamme de prothèses dentaires (couronnes, inlay-core, bridge) sans reste à charge et d'améliorer les soins préventifs.

Trois types de panier dentaire ont été définis.

Panier 100% Santé :

Ce panier se compose de prothèses dentaires intégralement remboursables par l'assurance maladie obligatoire et les complémentaires santé. Il couvre un large choix de prothèses fixes ou mobiles, avec des matériaux (céramo-métallique, céramique monolithique. . .) dont la qualité esthétique est adaptée à la localisation de la dent (distinction entre les dents « visibles » et les dents « non visibles ») :

- Couronnes céramiques monolithique (autre que zircone) et céramo-métalliques sur les dents visibles (incisives, canines et 1ère prémolaire).
- Couronnes céramique monolithique zircones (incisives, canines et prémolaires).
- Couronnes métalliques toute localisation.
- Inlays core et couronnes transitoires (liées aux couronnes définitives).
- Bridges céramo-métalliques (incisives).
- Bridges métalliques toute localisation.
- Prothèses amovibles à base résine.

Au 1^{er} janvier 2020, ce panier contient 8 types de prothèses et au 1^{er} janvier 2021, 50 prothèses entreront dans ce panier.

Panier aux tarifs maîtrisés

Le Panier aux tarifs maîtrisés comprend des prothèses dentaires dont les prix sont plafonnés afin de limiter le reste à charge pour l'assuré. Au 1^{er} janvier 2020, ce panier contient 6 types de prothèses. Au 1^{er} janvier 2021, 4 prothèses entreront dans ce panier et enfin au 1^{er} janvier 2022, 57 prothèses intégreront ce panier.

Panier aux tarifs libres

Le panier libre concerne tous les autres actes, avec un reste à charge plus conséquent pour l'assuré.

1.3.3 La réforme 100% santé sur les audio-prothèses

La mise en place de la réforme 100% santé sur les audio-prothèses est également progressive. Elle est entrée en vigueur le 1^{er} janvier 2019 et sera pleinement effective au 1^{er} janvier 2021. La progressivité de la réforme concerne les remboursements qui augmenteront chaque année jusqu'à atteindre un remboursement intégral en 2021 sur le panier 100% santé. Avant la réforme, les appareils étaient classés en quatre catégories, ils sont désormais classés en 2 catégories.

Aides auditives de classes I (Panier100% Santé) :

Ces appareils sont pris en charge intégralement à partir du 1^{er} janvier 2021.

Aides auditives de classes II (Panier libre) :

Les appareils de ce panier ont des prix libres et possèdent des fonctionnalités qui ne permettent pas de garantir une absence de reste à charge.

Chapitre 2

L'analyse préliminaire des données

2.1 Présentation du portefeuille

Le portefeuille d'étude concerne les affiliés santé collective de l'assureur Aviva sur les produits de la gamme standard. Les données considérées portent sur 3 exercices de survenance 2017, 2018 et 2019.

2.1.1 Présentation des données

Les contrats santé collective font l'objet d'une délégation de gestion, les données sont issues de différentes bases de données transmises par celui-ci.

Données tarifaires de l'entreprise

- Le nombre de salariés de l'entreprise
- L'âge moyen des salariés de l'entreprise
- Le département de l'entreprise
- Le zonier (4 zones géographiques)
- L'activité de l'entreprise (Code NAF)
- Le collège (Non cadre, cadre, ensemble du personnel)
- La structure tarifaire (adulte-Enfant, isolé-duo-famille, unique)

Les informations sur les garanties souscrites

- Le produit souscrit (ANI / Gamme Standard)
- Poste (hospitalisation, soins courants, dentaire, optique et bien-être)
- Le niveau de formule du poste (Niveau 0 : produit ANI / niveau 1 à 6 : gamme standard)
- Le régime (RG : Régime général / AM : Régime Alsace-Moselle)
- La souscription d'une option (oui / non)

Les informations sur les bénéficiaires

- Le type de bénéficiaire (Affilié, conjoint, enfant)
- La date de naissance du bénéficiaire
- Sexe du bénéficiaire

- Le code postal de l'adresse de l'affilié

Les données de consommation médicale

- Date de soin
- Acte médical
- Nombre d'actes
- Libellé de l'acte
- Famille de l'acte
- Base de remboursement
- Taux de remboursement du régime obligatoire
- Montant de la dépense
- Montant remboursé par la mutuelle
- Montant remboursé par la sécurité sociale

2.1.2 Traitement des données

2.1.2.1 Création d'indicateurs

Certains indicateurs ne figurant pas directement en lecture dans les bases de données peuvent être calculés. Voici quelques indicateurs qui ont été construits à partir des données disponibles :

- Âge du bénéficiaire : il est déterminé au 31 décembre de l'année de survenance
- Le nombre de salariés assurés par la complémentaire
- L'exposition : il s'agit de la fraction de l'année où l'assuré a été couvert. (Vaut 0 si l'assuré n'a pas été couvert dans l'année, vaut 1 si l'assuré a été couvert l'année entière)
- Option : l'assuré a souscrit ou non une option (0 : pas de souscription d'option en cours, 1 : l'assuré a souscrit une option en cours de validité)
- Secteur d'activité : le code Naf est extrêmement fin, un regroupement en 10 modalités détaillé ci-après a été réalisé à partir de la nomenclature des activités de niveau 1.

LIBELLE NAF 1	Secteur regroupé	secteur
Agriculture, sylviculture et pêche	Agriculture, sylviculture et pêche	Secteur 1
Industrie manufacturière	Industrie manufacturière, industries extractives et autres	Secteur 2
Industries extractives		
Production et distribution d eau , assainissement, gestion des déchets et dépollution		
Production et distribution d électricité, de gaz, de vapeur et d air conditionné		
Construction	Construction	Secteur 3
Commerce , réparation d automobiles et de motocycles	Commerce de gros et de détail, transports, hébergement et restauration	Secteur 4
Hébergement et restauration		
Transports et entreposage		
Information et communication	Information et communication	Secteur 5
Activités financières et d assurance	Activités financières et d assurance	Secteur 6
Activités immobilières	Activités immobilières	Secteur 7
Activités spécialisées, scientifiques et techniques	Activités spécialisées, scientifiques et techniques et activités de services administratifs et de soutien	Secteur 8
Activités de services administratifs et de soutien		
Administration publique	Administration publique, enseignement, santé humaine et action sociale	Secteur 9
Enseignement		
Santé humaine et action sociale		
Autres activités de services	Autres activités de services	Secteur 10
Activités extra-territoriales		
Arts, spectacles et activités récréatives		

TABLE 2.1 – Regroupement des secteurs d'activité

2.1.2.2 Écrêtement des valeurs extrêmes

En tarification de produits d'assurance non vie, la modélisation du coût des sinistres repose sur l'hypothèse fondamentale selon laquelle le portefeuille est composé de risques relativement homogènes. Cette hypothèse implique un faible poids des sinistres « graves » dans la sinistralité globale. Une part trop importante des sinistres graves pourrait conduire à biaiser les estimations. La solution à ce problème consiste à écrêter les valeurs extrêmes, la charge résiduelle pouvant faire l'objet d'une modélisation spécifique ou être répartie sur le portefeuille selon une règle d'affectation.

En santé, le poids des sinistres graves reste relativement modéré. Le poste le plus concerné par ce phénomène est le poste hospitalisation bien que le risque reste mesuré. Dans le cas de pathologie grave nécessitant une hospitalisation de longue durée, le régime obligatoire prend souvent l'intégralité du coût en charge. Le coût pour l'assureur portera alors principalement sur les postes sans prise en charge par la sécurité sociale comme la chambre particulière. Pour les postes optique et bien-être, les garanties sont souvent plafonnées.

Le tableau des quantiles ci-dessous montre une faible part des sinistres graves.

Niveau	Hospitalisation	Soins courants	Pharmacie	Dentaire	Optique	Bien-être
100Max 100%	18 041,2 €	7 853,8 €	809,8 €	7 482,0 €	2 500,0 €	3 269,0 €
99,5%	1 864,5 €	73,8 €	76,5 €	972,5 €	849,9 €	350,0 €
99%	347,5 €	28,2 €	19,2 €	263,4 €	550,0 €	90,0 €
95%	198,6 €	22,2 €	14,0 €	145,7 €	425,9 €	62,7 €
90%	116,2 €	18,8 €	11,4 €	81,4 €	355,0 €	57,5 €
75% Q3	69,0 €	16,4 €	9,5 €	33,3 €	300,0 €	53,5 €
50% Médiane	35,9 €	14,5 €	8,1 €	17,2 €	277,6 €	50,0 €
25% Q1	19,9 €	12,9 €	6,9 €	13,0 €	247,4 €	50,0 €
10%	13,5 €	11,4 €	5,8 €	11,6 €	200,0 €	47,5 €
5%	7,9 €	9,9 €	4,6 €	9,5 €	168,6 €	40,0 €
1%	5,1 €	8,2 €	3,4 €	8,2 €	100,0 €	32,0 €
0% Min	0,5 €	0,5 €	0,5 €	0,8 €	3,4 €	0,7 €

TABLE 2.2 – Tableau des quantiles du coût moyen par poste

Compte tenu des quantiles, un seuil d'écrêtement à 99.5% a été fixé, ce qui représente une charge de sinistre écrêtées pesant pour 2,2% de la charge de sinistres totale.

Risque	Poids des sinistres écrêtés
Hospitalisation	3,78%
Soins courant	2,51%
Pharmacie	3,10%
Dentaire	1,80%
Optique	0,58%
Bien-être	2,78%
Total	**2,2%**

TABLE 2.3 – Poids des sinistres écrêtés par poste

2.1.3 Exclusion de périmètre

Certains assurés doivent être exclus du périmètre car ils font l'objet d'une tarification spécifique. Deux types d'assurés sont concernés, les assurés en situation de portabilité et les assurés retraités.

2.1.3.1 Exclusion des assurés en portabilité

En santé collective, sous certaines conditions, les assurés quittant leur entreprise peuvent continuer à bénéficier de leur contrat santé, c'est le droit à la portabilité. Ce droit leur permet de continuer à bénéficier de leur contrat santé pendant une durée maximale de 12 mois avec la gratuité des primes. Cet avantage social représente un coût pour l'assureur, généralement autour de 3%. Il peut être mutualisé sur l'ensemble des assurés ou faire l'objet d'une modélisation particulière. En effet, certains segments peuvent être plus exposés notamment les secteurs d'activité faisant appel à beaucoup de salariés en contrat à durée déterminée (CDD).

	Poids du nombre de bénéficiaires	*Exposition moyenne*	*Poids des prestations*	*Poids de la consommation médicale*
Affilié Portabilité	*3,4%*	*123 jours*	*2,1%*	*2,0%*
Ayant-droit Portabilité	*2,0%*	*137 jours*	*0,9%*	*0,9%*
Total	***5,4%***	***128 jours***	***3,0%***	***3,0%***

TABLE 2.4 – Tableau du poids de la portabilité dans le portefeuille sur les exercices 2017-2019

2.1.3.2 Exclusion des assurés en loi évin

Les assurés en loi Evin sont des anciens salariés qui lors de leur départ en retraite ont souhaité conserver le contrat de complémentaire santé de leur entreprise. Ils ne font plus partis des effectifs de l'entreprise et sont assurés à titre individuel avec une tarification spécifique. A ce titre, ils sont exclus de l'étude.

	Poids du nombre de bénéficiaires	*Exposition moyenne*	*Poids des prestations*	*Poids de la consommation médicale*
Affilié loi Evin	*0,2%*	*203 jours*	*0,3%*	*0,3%*

TABLE 2.5 – Tableau du poids des assurés loi Evin dans le portefeuille sur les exercices 2017-2019

2.1.4 Segmentation de la base de données

La base de données est constituée de 3 exercices 2017, 2018 et 2019. La santé étant une branche courte, les données à fin septembre comportent l'ensemble des prestations de l'exercice précédent. Sur les 3 exercices, la consommation médicale est stable sur ce portefeuille. L'analyse de la fréquence et du coût moyen par poste ne montre pas la nécessité de corriger les données de la dérive.

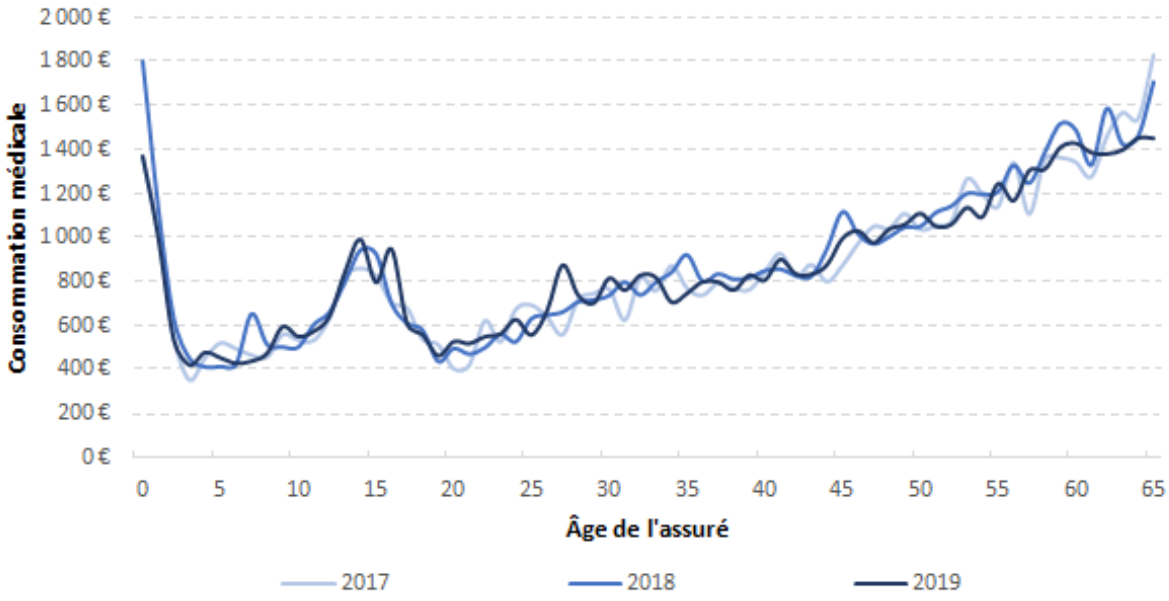


FIGURE 2.1 – Consommation médicale en fonction de l'âge de l'assuré

Dans le cadre de la modélisation, la base de données a été segmentée en deux bases, une base d'apprentissage comportant 70% des effectifs et une base de test avec 30% des effectifs.

2.1.5 Analyse descriptive du portefeuille

Le portefeuille se compose de 48 845 bénéficiaires au 31/12/2019 avec une exposition moyenne annuelle de 0.82 année.

2.1.5.1 Réseau de distribution

Ce portefeuille est issus de deux canaux de distribution, des agents et des courtiers avec une large partie des bénéficiaires provenant du réseau des agents.

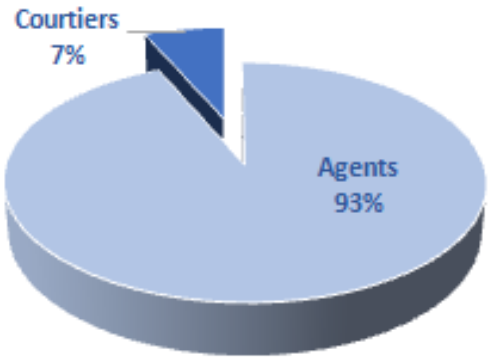


FIGURE 2.2 – Répartition des bénéficiaires par réseau de distribution

2.1.5.2 Démographie du portefeuille

La pyramide des âges de la figure 2.3 ci-dessous montre une répartition du portefeuille par âge relativement équilibrée avec une légère sur-représentation des 25-35 ans. Le nombre d'hommes est légèrement supérieur au nombre de femmes, ils représentent 53% des affiliés.

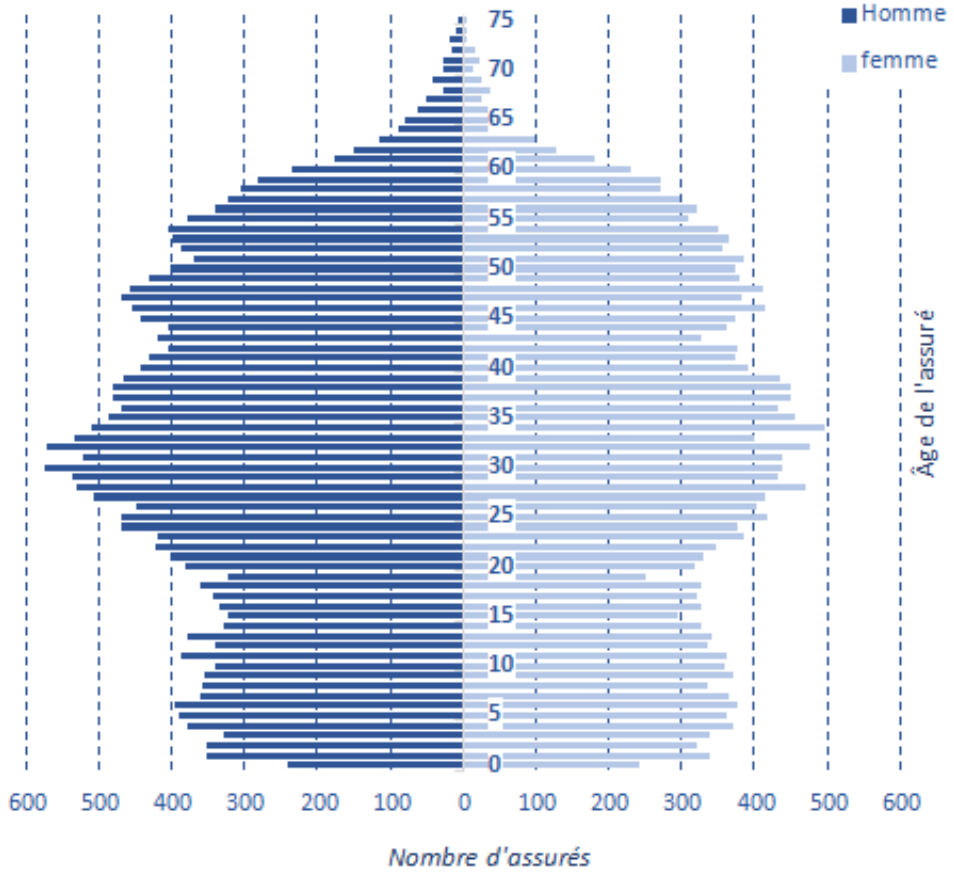


FIGURE 2.3 – Pyramide des âges du portefeuille au 31/12/2019

2.1.5.3 Répartition géographique du portefeuille

La répartition du portefeuille par département est très hétérogène. Les départements avec une densité urbaine plus importante sont naturellement davantage représentés. La carte de la répartition du portefeuille par département indique sans surprise que les départements composés de grandes agglomérations telles que Paris, Nantes, Lille, Bordeaux, Rouen, Lyon concentre une partie plus importante du portefeuille.

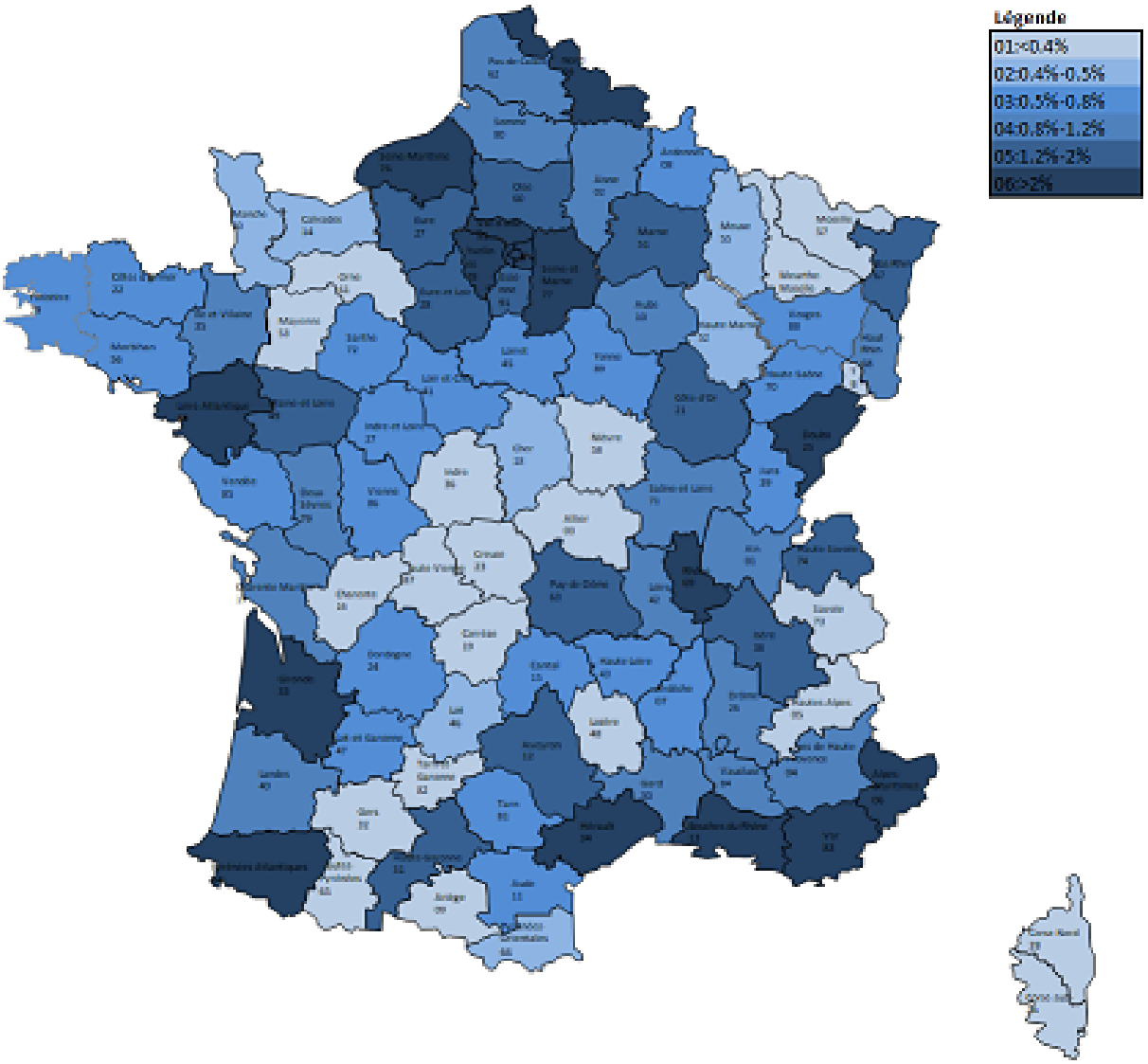


FIGURE 2.4 – Répartition géographique du portefeuille au 31/12/2019

2.2 Analyse des variables explicatives de la consommation médicale

2.2.1 Analyse bivariée

L'analyse bivariée est une étape préliminaire à la modélisation dont l'objectif est de détecter les variables présentant une influence sur le niveau de la consommation médicale de l'assuré. Cette analyse permet également de fournir une première vision de l'intensité de la relation. Seules les variables les plus importantes seront présentées dans cette section, l'analyse sur les autres variables est disponible en annexe A.

2.2.1.1 L'âge moyen des salariés

En santé collective, l'information tarifaire n'est pas l'âge des bénéficiaires mais l'âge moyen des salariés. Cependant, avant de regarder la consommation médicale selon l'âge moyen des salariés une analyse graphique de la consommation médicale selon l'âge des bénéficiaires apparaît importante.

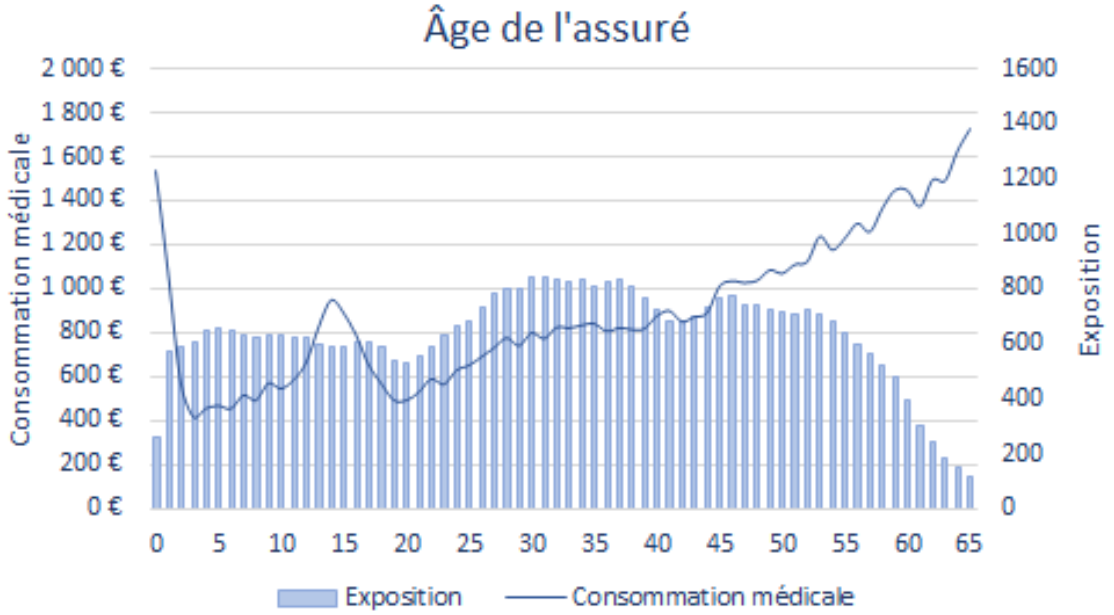


FIGURE 2.5 – Consommation médicale annuelle moyenne en 2018 fonction de l'âge de l'assuré

La consommation médicale est très corrélée avec l'âge de l'assuré. Le graphique de la figure 2.5 met en évidence la présence d'une corrélation positive entre l'âge de l'assuré et sa consommation médicale annuelle moyenne. Le graphique fait également apparaître deux périodes de forte consommation médicale chez les mineurs. La première concerne les jeunes enfants avec un pic correspondant à la première année. Cette forte consommation s'explique par le fait que durant les deux premières années, les enfants font l'objet d'un suivi médical très régulier. La seconde période de forte consommation des mineurs concerne les 12-18 ans avec un pic de consommation à 15 ans. Cette hausse de la consommation est liée au poste dentaire et plus particulièrement aux dépenses d'orthodontie.

Chez les adultes, la consommation médicale croît tout au long de la vie mais cette croissance n'est pas uniforme, elle est relativement linéaire de 20 ans à 45 ans et ensuite elle s'accélère avec une forme plutôt exponentielle. Une forte hausse de la consommation est observée à partir de 43 ans liée à la presbytie.

L'évolution croissante de la consommation médicale avec l'âge du bénéficiaire se confirme également en considérant l'âge moyen des salariés. Le graphique 2.6 met en évidence une évolution exponentielle de la consommation médicale avec l'âge moyen des salariés.

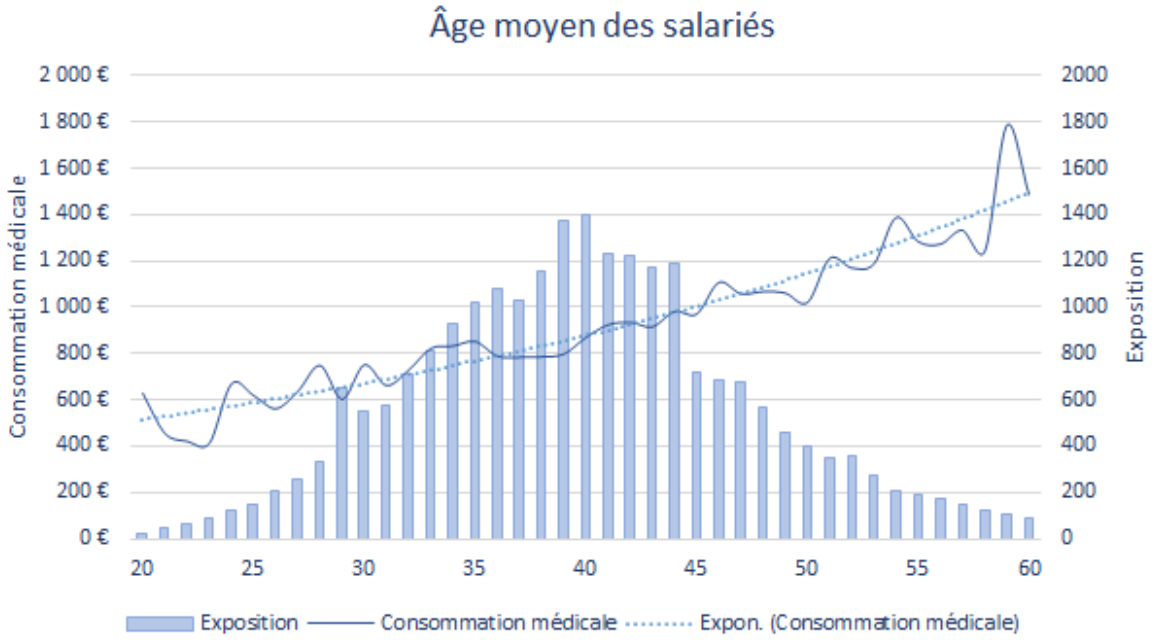


FIGURE 2.6 – Consommation médicale annuelle moyenne en fonction de l'âge moyen des salariés

L'analyse graphique ne permet pas d'identifier clairement un découpage en classe d'âge mais des effets d'escaliers s'observent à 40 ans, 45 ans et 50 ans. Le graphique met également en évidence une concentration du portefeuille sur la tranche d'âge 35-45 ans. Compte tenu des ces éléments, un découpage en classe d'âge de 5 ans entre 30 et 50 ans apparaît pertinent.

2.2.1.2 Le niveau de garantie

Le niveau de garantie de la complémentaire santé est un facteur influençant la consommation médicale. Le graphique 2.7 ci-après montre une croissance monotone entre le niveau de garantie et la consommation médicale. Le niveau moyen de la consommation médicale par assuré est deux fois plus élevé sur le niveau 6 que sur le niveau 0.

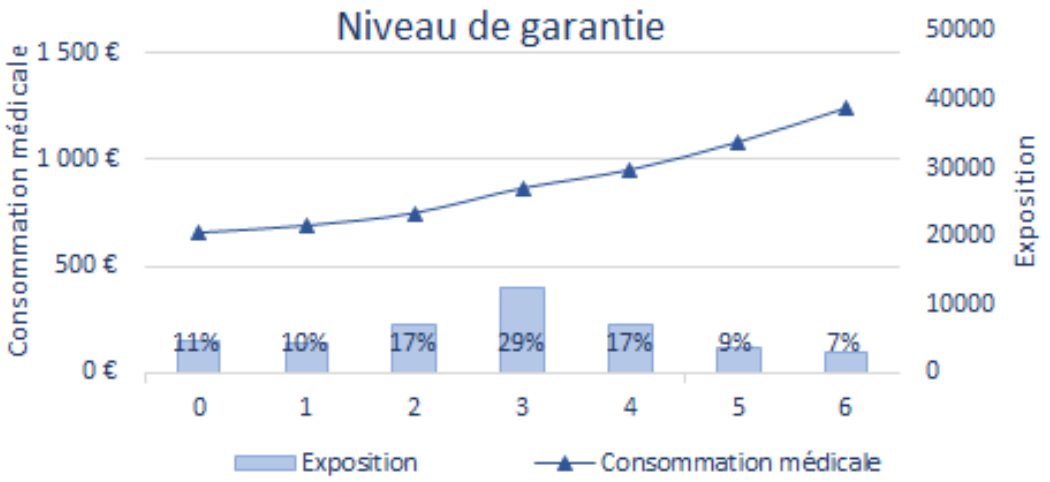


FIGURE 2.7 – Consommation médicale en fonction du niveau de garantie

La figure 2.8 montre que cette tendance est particulièrement prononcée sur les postes optique et dentaire. La dépense annuelle par assuré en optique est 2,7 fois plus élevée sur le niveau 6 que sur le niveau 0 et en dentaire c'est encore plus marqué avec un rapport égal à 3,5 fois.

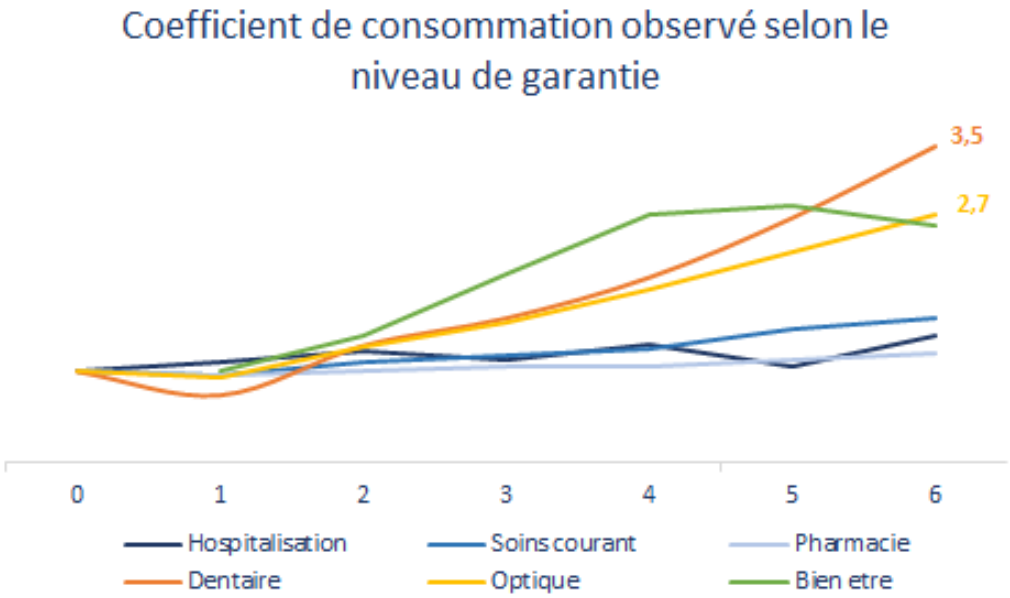


FIGURE 2.8 – Coefficient de consommation médicale en fonction du niveau de garantie

Ce constat est la manifestation d'une modification éventuelle du comportement des assurés en fonction de leur niveau de couverture ainsi que d'un possible effet d'anti-sélection. Cependant s'agissant d'un contrat collectif, l'effet d'anti-sélection doit être moindre tout du moins sur les entreprises de taille importante, les salariés n'étant pas souscripteur du contrat.

2.2.1.3 Le nombre de salariés

Le nombre de salariés semble influencer dans une moindre mesure que les variables précédentes la consommation médicale. Le graphique 2.9 ci-dessous de la consommation médicale moyenne en fonction des classes d'effectifs salariés de l'entreprise montre de légères variations.

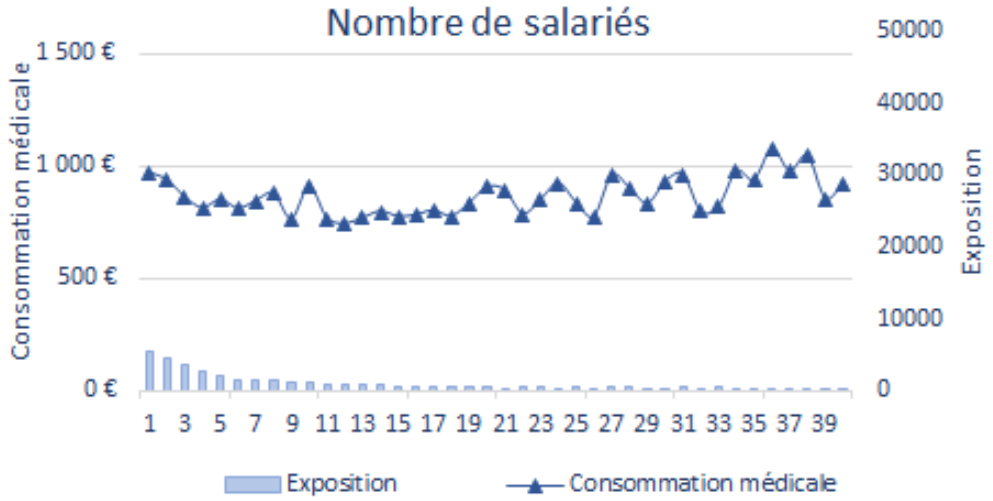


FIGURE 2.9 – Consommation médicale en fonction du nombre de salariés dans l'entreprise

Le graphique met en évidence une plus forte consommation pour les entreprises de moins de 4 salariés et plus particulièrement concernant les entreprises avec un et deux salariés. Il s'agit ici vraisemblablement d'un effet d'anti-sélection. En effet, sur les petites entreprises, les salariés affiliés peuvent correspondre au chef d'entreprise et à leur conjoint. A partir de ces observations, un regroupement en classe de nombre de salariés (figure 2.10) a été effectué.

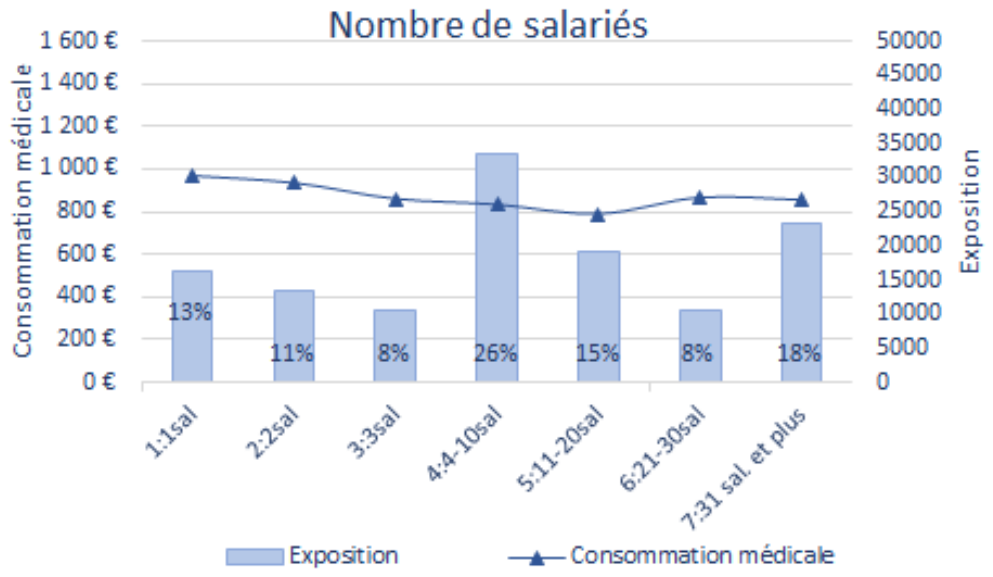


FIGURE 2.10 – Consommation médicale en fonction du nombre de salariés dans l'entreprise

2.2.1.4 Les secteurs d'activité

Le graphique 2.11 ci-dessous met en évidence une variabilité de la consommation médicale moyenne selon le secteur d'activité. Deux secteurs présentant une consommation plus basse se dégagent, il s'agit du secteur 1 et 3 correspondant respectivement à l'agriculture et la construction. A l'inverse, les secteurs 6 (Activités financières et Assurances), 7 (Activités immobilières) et 9 (Administration publique, enseignement, santé humaine et action sociale) présentent une consommation plus importante.

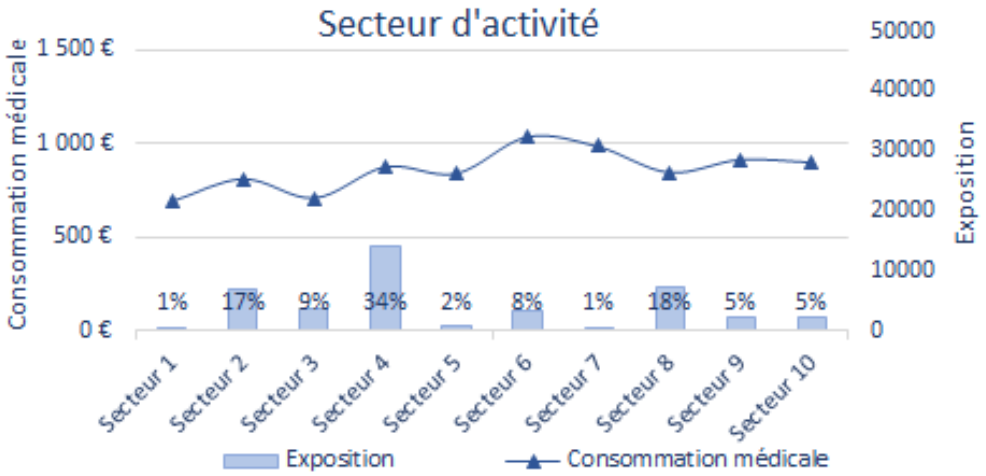


FIGURE 2.11 – Consommation médicale en fonction du secteur d'activité

Ce premier constat devra être vérifié lors de l'étape de modélisation afin de s'assurer que cela n'est pas un effet lié aux autres variables.

2.2.1.5 Le lien de parenté

Le lien de parenté influence également la consommation médicale. Concernant les enfants, cela confirme ce qui a été observé précédemment sur le graphique de la consommation médicale avec l'âge. Généralement, le tarif enfant est de l'ordre de 60% du tarif adulte, ce que semble refléter l'écart de consommation constaté.

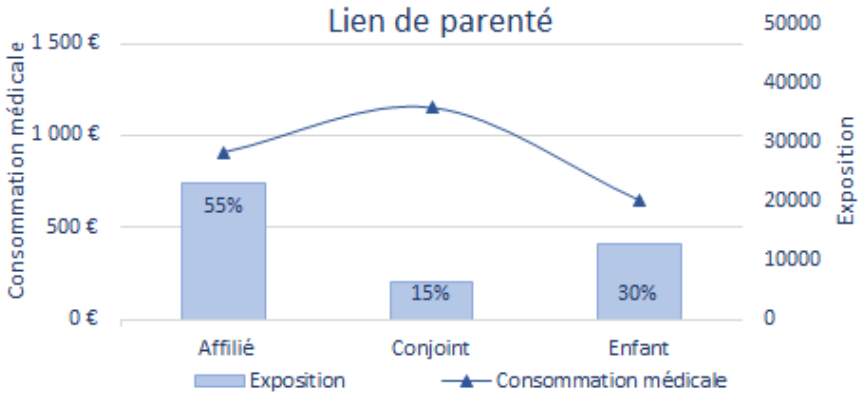


FIGURE 2.12 – Consommation médicale en fonction du lien de parenté

Concernant les conjoints, une surconsommation de 25% est constatée. Elle est liée également à un effet d'anti-sélection, l'affiliation étant optionnelle pour les ayant droits.

2.2.1.6 L'âge et le niveau de garantie

Le graphique croisé des deux variables présentant la plus forte influence sur la consommation médicale à savoir l'âge et le niveau de garantie montre des écarts importants de consommation. Si l'on se focalise sur les adultes, il y a d'un côté les petits consommants avec une moyenne proche de 500 euros par an et de l'autre côté les gros consommants avec une moyenne supérieure à 2000 euros par an. Les petits consommants sont jeunes avec des garanties d'entrées de gamme tandis que les gros consommants sont des salariés proches de la retraite avec des garanties hautes gammes.

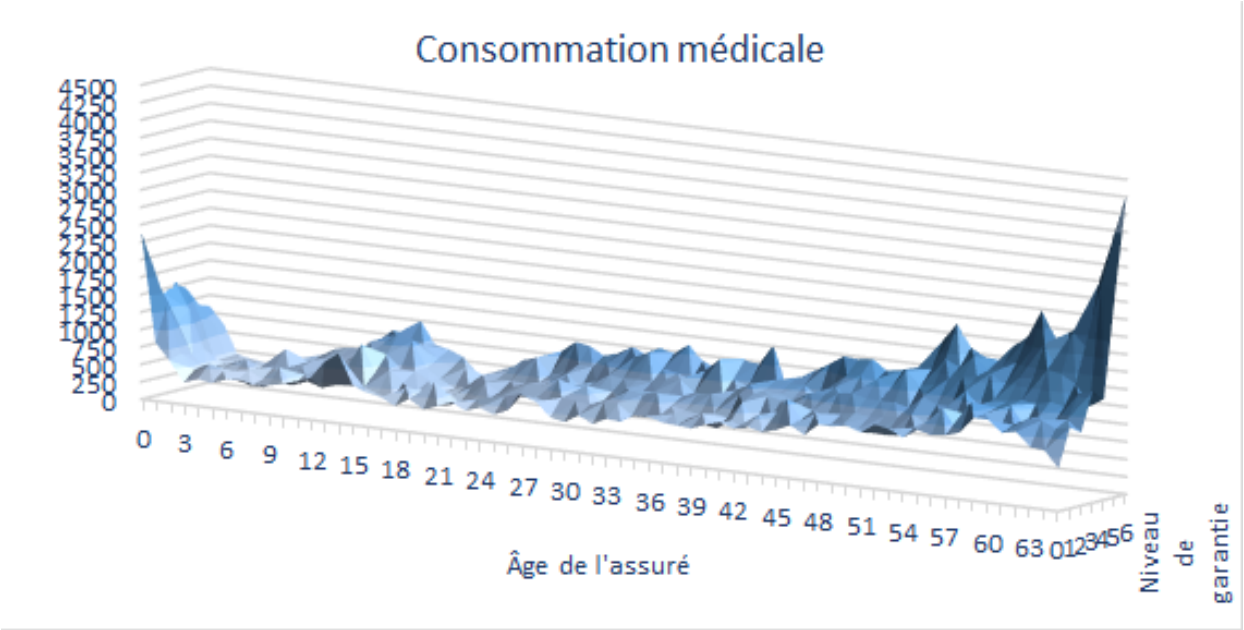


FIGURE 2.13 – Consommation médicale en fonction de l'âge et du niveau de garantie

2.2.2 Etude du zonier

La zone géographique, déterminée à partir du lieu d'habitation de l'assuré est un élément déterminant du niveau de sa consommation médicale. Cette influence s'explique par différents facteurs tels que la densité de médecins, le pourcentage de médecins en secteur 2. Le tarif du médecin varie selon qu'il est conventionné ou non et selon son secteur. En secteur 1, les tarifs du médecin sont fixés par la convention et le médecin ne pratique pas de dépassement d'honoraire. En secteur 2, les médecins fixent eux-mêmes leurs tarifs, ils peuvent pratiquer des dépassements d'honoraires mais avec des tarifs maîtrisés. Les médecins non conventionnés pratiquent des tarifs totalement librement. D'après l'observatoire des pratiques tarifaires, la moyenne nationale du niveau des dépassements chez les médecins de secteur 2 se situe à 52%, mais il y a une forte hétérogénéité selon les départements avec des taux de dépassements qui varient de 11% pour le Cantal à 114% pour Paris.

2.2.2.1 Choix de l'approche

Différentes approches de la détermination du risque géographique existent. Un première approche fréquemment utilisée en assurance non vie consiste à déterminer le risque géographique à partir des résidus d'une modélisation sans effet géographique. Le mémoire « Modélisation du risque géographique en santé, pour la création d'un nouveau zonier. Comparaison de deux méthodes de lissage spatial »(Catalina Sepulveda, 2016) est une illustration de cette approche. Cette approche est également bien adaptée à la réalisation d'un exercice de modélisation de l'effet géographique avec des variables exogènes mais elle nécessite un volume important de données. Une seconde approche fréquemment utilisé en santé consiste à déterminer d'abord les zones géographiques avec des méthodes de classification par exemple et à les intégrer ensuite dans une modélisation. Dans le cadre de cette étude, c'est cette dernière approche qui a été retenue.

Le risque géographique peut être déterminé avec plus ou moins de finesse, selon la problématique traitée et le volume de données disponibles. En santé collective, les zoniers sont souvent déterminés au niveau département ou région. Le risque géographique est défini à partir de l'adresse de l'entreprise et celles-ci sont particulièrement concentrées dans les départements composés de grandes agglomérations. La figure 2.4 de la répartition du portefeuille l'illustre. Un zonier au niveau région ne semble pas suffisamment fin car il existe une forte hétérogénéité des départements composant une région. Il a donc été opté pour la réalisation d'un zonier par département.

Pour l'identification des différentes zones, la méthode de la classification hiérarchique avec la distance de WARD a été retenue, elle est relativement facile à mettre en œuvre et présente l'avantage de ne pas avoir à spécifier le nombre de classes contrairement à la méthodes des K-means par exemple.

2.2.2.2 Méthode de classification hiérarchique ascendante

La classification hiérarchique ascendante (CAH) est une méthode de classification visant à partitionner des individus en un certain nombre de classes. Le processus d'élaboration des classes repose sur deux objectifs fondamentaux : un objectif d'homogénéité au sein de chaque classe (homogénéité intra-classe) et un objectif d'hétérogénéité entre les classes (hétérogénéité interclasse). L'homogénéité intra-classe consistera à regrouper les individus présentant les plus fortes similarités. L'hétérogénéité interclasse quant à elle consistera à ce que les classes soient les plus dissemblables possibles.

La CAH est une méthode ascendante et itérative. Elle commence par considérer chaque individu comme une classe et procède à des regroupements successifs d'une ou plusieurs classes en se basant sur un critère de similarité préalablement défini. Le processus itératif s'achève lorsque l'ensemble des individus appartiennent à une seule classe.

Le résultat de ce processus se présente sous la forme d'un arbre de classification appelé aussi dendrogramme.

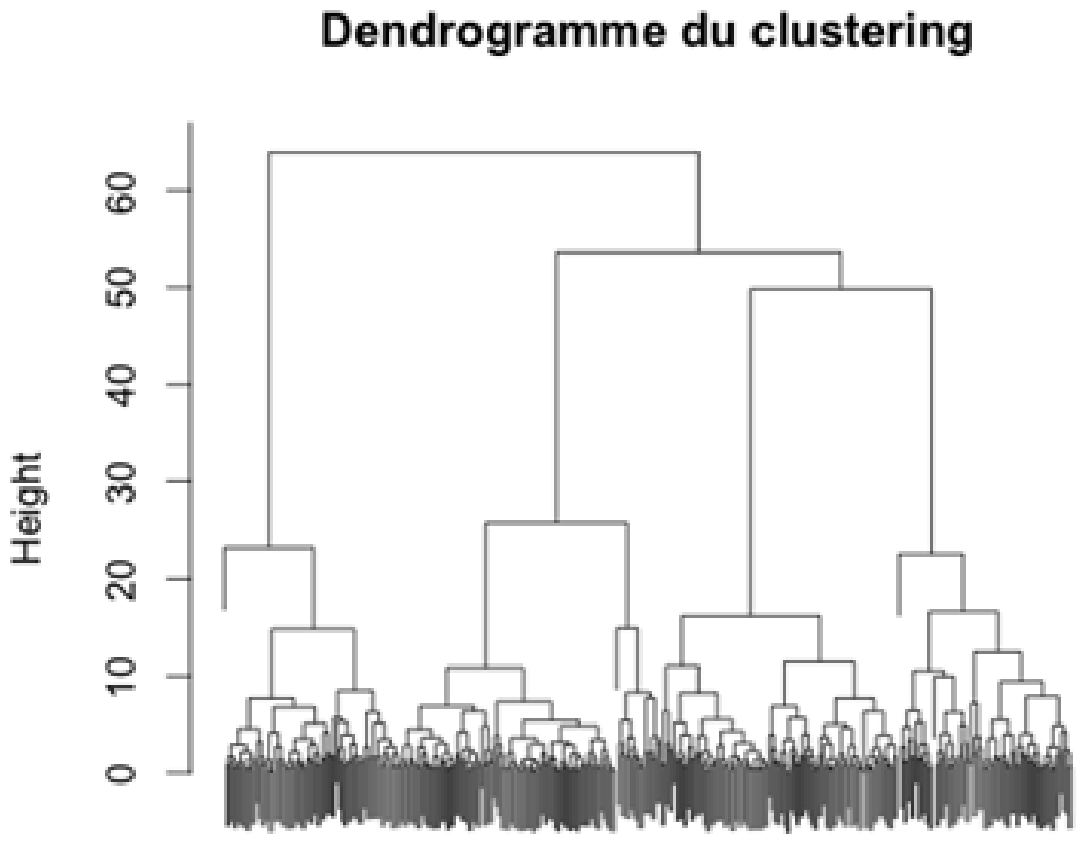


FIGURE 2.14 – Dendogramme

Les regroupements successifs des classes sont réalisés via un algorithme ascendant s'appuyant sur une notion de distance comme critère de mesure de la similarité entre les individus et une règle d'agrégation des classes.

2.2.2.3 La notion de distance

Il existe une multitude de fonction de distance notamment la distance euclidienne, la distance de Manhattan, la distance de Minkowsky, etc.

Distance Euclidienne

Il s'agit de la distance la plus fréquemment utilisée.

$$distance(x, y) = \left\{ \sum_i (x_i - y_i)^2 \right\}^{\frac{1}{2}}$$

Distance Euclidienne au carré

Elle a pour effet de sur pondérer les points atypiques.

$$distance(x, y) = \sum_i (x_i - y_i)^2$$

Distance du City-block (Manhattan).

Cette distance est simplement la somme des différences entre les dimensions. Dans la plupart des cas, cette mesure de distance produit des résultats proches de ceux obtenus par la distance euclidienne simple. En revanche, cet indicateur permet d'atténuer les effets des points atypiques car les distances ne sont pas élevées au carré.

$$distance(x, y) = \sum_i |x_i - y_i|$$

2.2.2.4 Règles d'agrégation

La règle d'agrégation intervient à partir de la seconde étape du processus itératif. Lors de la première étape, chaque classe est représentée par un seul individu et la distance entre les classes correspond à la distance entre chaque individu. A partir de la deuxième étape, les classes peuvent se composer de plusieurs individus, il est donc nécessaire de choisir une règle permettant de déterminer la distance entre les classes. Il existe différentes règles d'agrégation telle que la règle de saut minimum, la règle du diamètre, la méthode WARD...

Saut minimum ou single linkage (distance minimum)

La distance entre 2 classes correspond à la distance entre les deux individus les plus proches de chacune des classes (les plus proches voisins). Cette règle provoque des chaînes d'objets assemblés en classes, et les résultats obtenus ressemblent à de longues chaînes.

Diamètre ou complete linkage (distance maximum)

Dans cette méthode, les distances entre classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes, c'est-à-dire les voisins les plus éloignés. Cette méthode donne souvent de bons résultats lorsque les objets forment déjà naturellement des "groupes" bien distincts. Si les classes ont plutôt une forme allongée, ou sont en forme de chaîne, cette méthode sera mal adaptée.

Moyenne non pondérée des groupes associés

Ici, la distance entre deux classes est calculée comme la distance moyenne entre tous les objets deux à deux dans les deux classes différentes. Cette méthode est efficace lorsque les objets forment déjà naturellement des groupes bien distincts, mais se révèle également bien adaptée dans le cas de classes allongées, de type chaîne.

Moyenne pondérée des groupes associés

Cette méthode est identique à la méthode moyenne non pondérée des groupes associés, à la différence près que la taille des classes respectives, c'est-à-dire le nombre d'objets qu'elle comporte, est utilisée ici comme pondération. Cette méthode est souvent préférée à la précédente lorsque les tailles de classes sont assez inégales.

Barycentre non pondéré des groupes associés

Le barycentre d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. C'est en quelque sorte le centre de gravité de la classe respective. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les barycentres respectifs.

Barycentre pondéré des groupes associés (médiane)

Cette méthode est identique à la précédente, à la différence près qu'une pondération est introduite dans les calculs afin de prendre en compte les tailles des classes.

Méthode de Ward pour distance Euclidienne (méthode du moment d'ordre 2)

Cette méthode s'appuie sur la notion d'inertie, l'inertie étant un indicateur de dispersion correspondant à la variance généralisée au cas multidimensionnel.

$$Inertie = \sum_{w} d^2(X(w), G) \quad avec\ G\ barycentre\ global$$

D'après la relation de Huygens, suite à une partition des observations, l'inertie totale peut se décomposer comme la somme de l'inertie inter-classes et de l'inertie intra-classes.

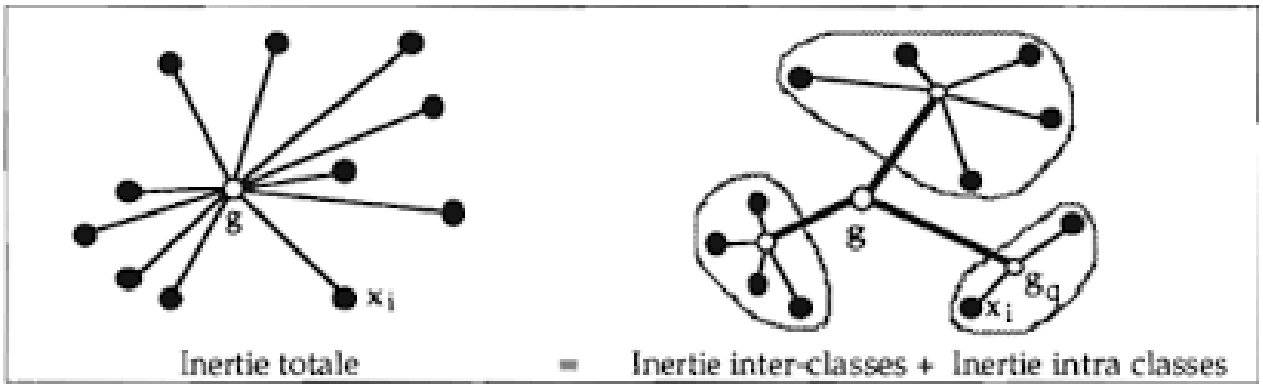


FIGURE 2.15 – Décomposition de l'inertie

$$\sum_{w} d^2(X(w), G) \;=\; \sum_{g} n_g \times d^2(G_g, G) \;+ \sum_{g} \sum_{w \in g} d^2(X(w), G_k)$$

$$Dispersion\ totale\ (T) \;=\; Dispersion\ inter-classes\ (B) \;+\; Dispersion\ intra-classes\ (W)$$

$$Part\ d'inertie\ expliquée\ par\ la\ partition : R^2 \;=\; \frac{B}{T}$$

La méthode de Ward repose sur la détermination d'un indice de dissimilarité entre chaque classe, cet indice correspond à la perte d'inertie inter-classe générée par le regroupement. La stratégie d'agrégation consistera à regrouper les classes qui génèrent la plus faible perte d'inertie inter-classe et par équivalence qui font le moins augmenter l'inertie intra-classe.

Estimation de la perte d'inertie inter-classe lors du regroupement de deux classes A et B :

Soit :

$$\begin{aligned} g_A &= \textit{centre de gravité de la classe } A \textit{ (poids } p_A) \\ g_B &= \textit{centre de gravité de la classe } B \textit{ (poids } p_B) \\ g_{AB} &= \textit{ centre de gravité de leur reunion} \\ g_{AB} &= \frac{p_A.g_A + p_B.g_B}{p_A + p_B} \end{aligned}$$

L'inertie inter-classe étant la moyenne des carrés des distances des centres de gravité des classes au centre de gravité total, la perte d'inertie inter-classe est égale à :

$$p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B) d^2(g_{AB}, g)$$

Ce qui est équivalent à :

$$\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

2.2.2.5 Règle d'élagage de l'arbre

L'élagage de l'arbre consiste à choisir la meilleure partition de l'arbre. Ce choix peut s'effectuer en s'appuyant sur la perte de distance inter-classe, cela revient à découper l'arbre au niveau d'un nœud suivi d'une forte perte de distance (Inertie inter-classe pour Ward). Une autre approche complémentaire à la première reposera sur l'interprétation des classes, il s'agira de choisir un découpage présentant un sens naturel. Le choix du découpage à partir de la perte de distance peut être facilité avec une approche graphique. Le semi-partial R-squared (SPRSQ) mesure la perte d'inertie inter-classe générée par le regroupement de 2 classes.

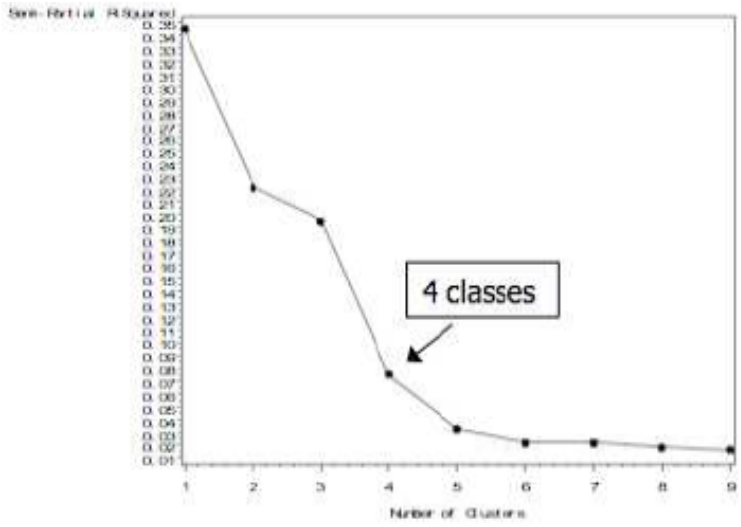


FIGURE 2.16 – SPRSQ en fonction du nombre de classes

Sur le graphique ci-dessus, un découpage en 4 classes semble approprié, cela correspond a un niveau avec un faible SPRSQ suivi d'un fort SPRSQ à l'agrégation suivante en 3 classes.

2.2.2.6 Application au portefeuille

En santé, la zone géographique peut influer sur les deux paramètres de la consommation médicale, à savoir la fréquence et le coût des actes médicaux. Cette influence peut être différemment marquée en fonction du poste de soins considéré. Sur le poste soins courants par exemple, la consommation médicale sera plus importante dans les grands centres urbains en raison notamment des dépassements d'honoraires des médecins plus élevés. En revanche, ce n'est pas le cas pour le poste pharmacie pour lequel les prix des médicaments remboursés sont réglementés. Afin de prendre en compte les spécificités des postes, le zonier sera déterminé pour chacun des 6 grands postes de dépenses (hospitalisation, soins courants, pharmacie, dentaire, optique et bien-être).

Ces zoniers à l'exception du poste hospitalisation seront établis à partir des prestations moyennes annuelles par assuré en considérant les 3 années 2017, 2018 et 2019. Le choix des prestations plutôt que la dépense réelle permet de prendre en compte le coût réel pour l'assureur. Cela permet également de limiter le problème des valeurs extrêmes observé avec la dépense réelle en raison des limites de remboursement de la complémentaire santé.

Concernant le poste hospitalisation, pour un même département il peut parfois y avoir une forte volatilité des prestations moyennes par assuré d'une année à l'autre. Cette instabilité rend difficile l'utilisation de cet indicateur pour la construction du zonier. Afin de palier à cet effet, le choix de l'indicateur s'est porté sur le coût moyen de la chambre particulière beaucoup plus stable et permettant de refléter la composante géographique liée au coût des actes d'hospitalisation.

L'exposition du portefeuille est également très hétérogène en fonction des départements et apparaît très corrélée à la densité d'habitant. L'élaboration du zonier en conservant des départements avec peu d'assurés n'est pas pertinente. En effet, pour ces départements, les indicateurs de consommation médicale sont très volatiles. Ils ont donc été exclus de la classification hiérarchique. Le critères d'exclusion a été fixé à 200 bénéficiaires. La consommation médicale moyenne de ces départements exclus est faible, ils ont donc été affectés à la zone avec l'effet géographique le plus faible.

Les différents zoniers ont été établis selon le même procédé. Seul les résultats de l'élaboration du zonier soins courants seront détaillés. Les cartes des zoniers élaborés sur les autres postes sont disponibles en annexe B.

Élaboration du zonier soins courants

Le poste soins courants est le poste le plus important, environ 40% de la consommation médicale. L'application de la CAH sur le poste soins courants permet d'obtenir les résultats suivants :

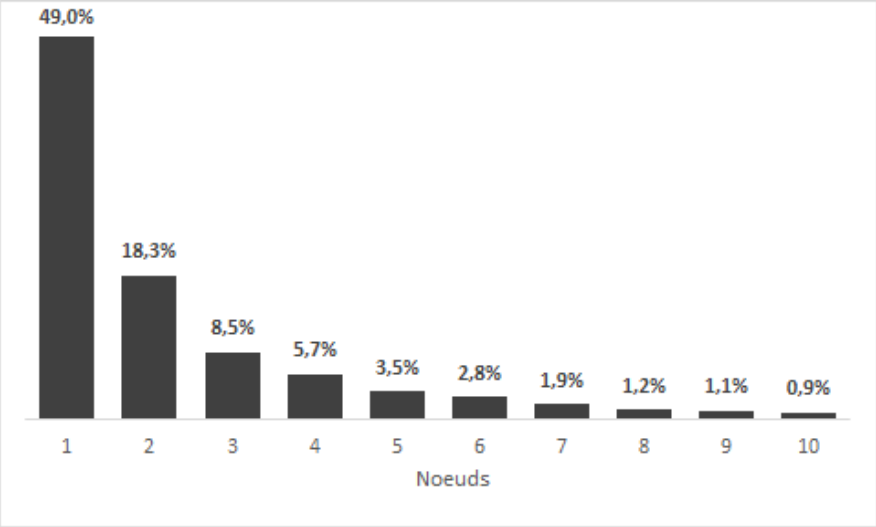


FIGURE 2.17 – Part de l'inertie expliquée par noeud

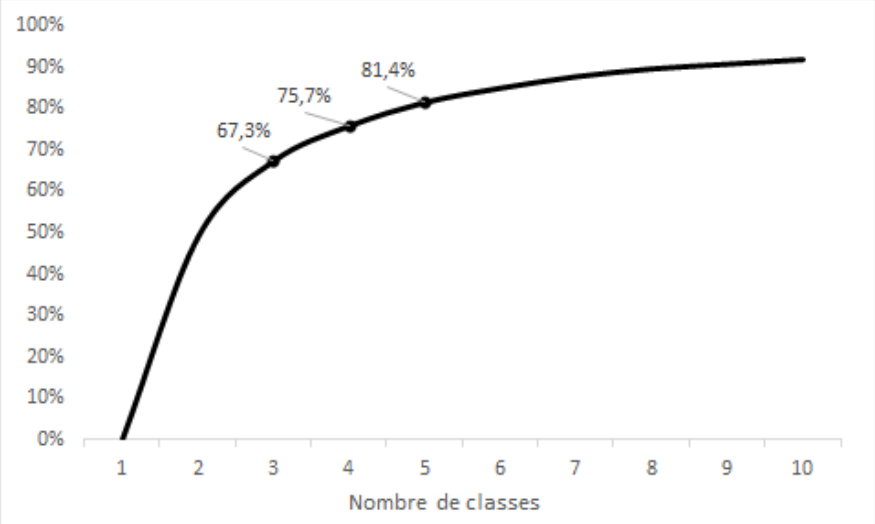


FIGURE 2.18 – Part de l'inertie expliquée en fonction du nombre de classes

L'analyse du graphique de la part de l'inertie expliquée par chacun des nœuds (Figure 2.17) met en évidence un saut important entre le nœud n°2 (18,1% d'inertie) et le nœud n°3 (8,5% d'inertie). Cette zone correspond au coude de la courbe et peut être retenue comme découpage optimal. Avec 3 classes, elle permet de conserver 67,3% de l'inertie.

Cependant, il peut apparaître pertinent de sélectionner un découpage plus fin permettant ainsi de conserver une part plus importante de l'inertie totale. Dans le cas suivant, un découpage en 4 et 5 classes permet de conserver respectivement 75,7% et 81,4% de l'inertie.

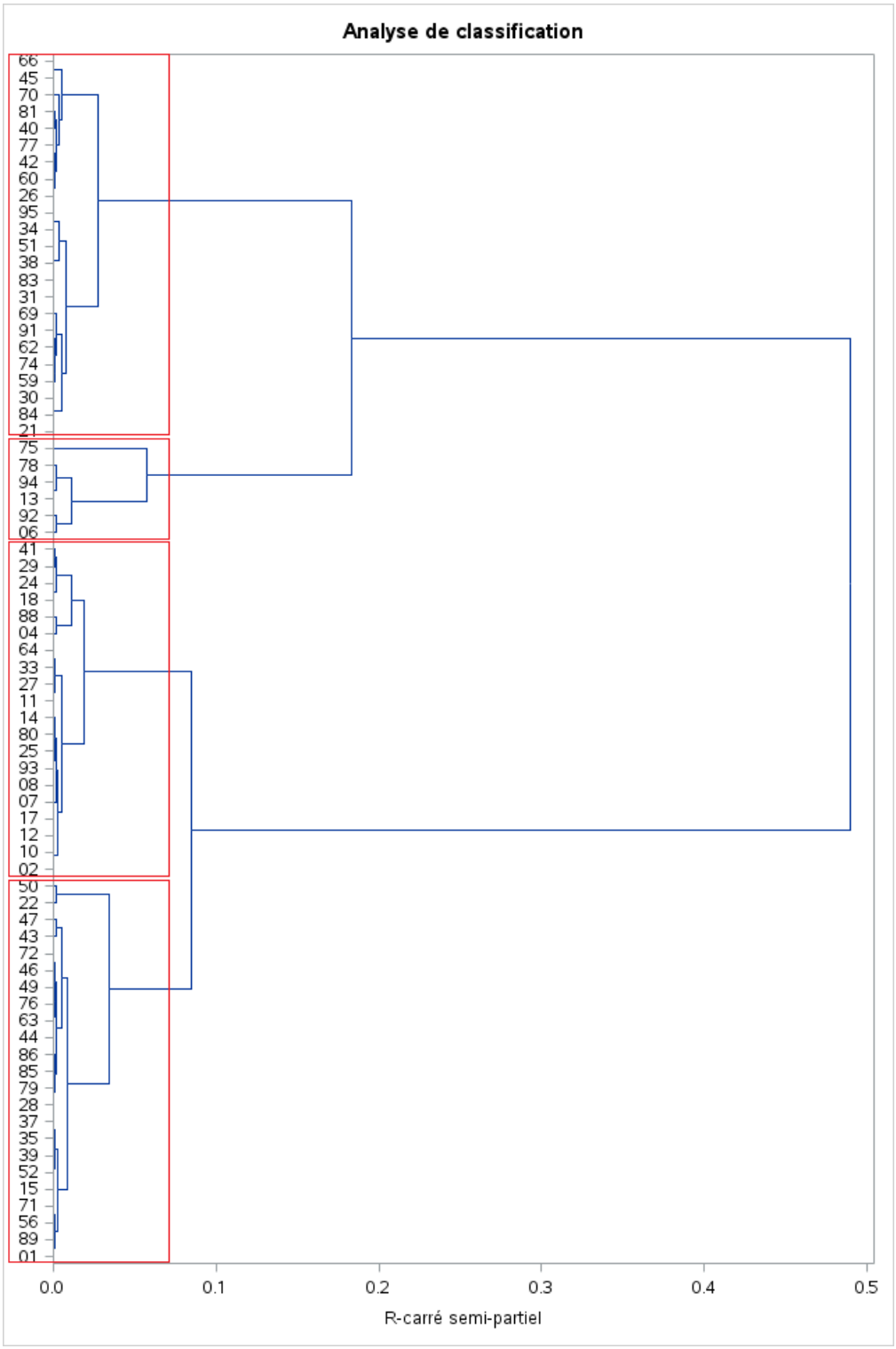


FIGURE 2.19 – Dendogramme sur le poste soins courants

L'analyse du dendogramme montre qu'un découpage en 5 classes conduit à établir une classe avec un seul département correspondant à paris (75). De ce fait, un découpage en 4 classes sera retenu pour le poste soins courants. La pertinence d'un découpage en 4 classes versus le découpage optimal en 3 classes suggéré par l'analyse de la perte d'inertie pourra être testée lors de l'étape de modélisation.

Ci-dessous la carte résultante du découpage en 4 classes :

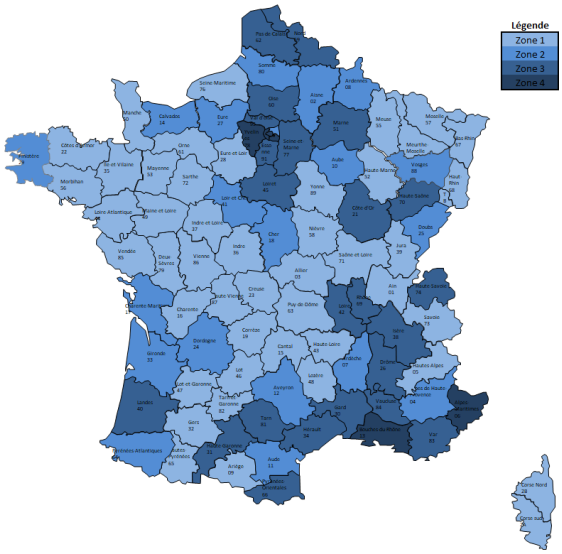


FIGURE 2.20 – Zonier soins courants

La première classe se compose de 6 départements dont 4 départements de l'Ile-de-France (Paris, Yvelines, Hauts-de-Seine, Val-de-Marne et 2 départements du sud est (Bouches-du-Rhône, Alpes-Maritimes). Dans une étude, l'observatoire des pratiques tarifaires fait mention des 2 cartes ci-dessous l'une reflétant le pourcentage de médecins spécialistes installés en secteur 2 et la seconde les taux de dépassement d'honoraire.

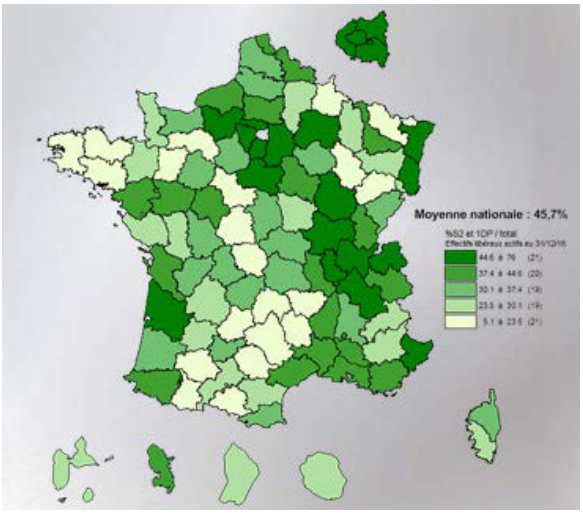


FIGURE 2.21 – % des médecins spécialistes installés en secteur 2 (Assurance maladie, 2017)

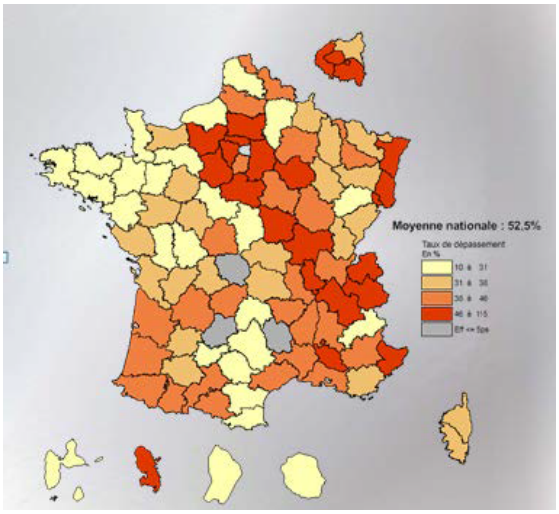


FIGURE 2.22 – Taux de dépassement moyen / département (Assurance maladie, 2017)

Ces deux cartes montrent de fortes concordances avec le zonier soins courants. Les 2 cartes ci-dessus mettent en évidence une part de médecins spécialistes de secteur 2 supérieure à 50% dans les départements d'Ile-de-France, mais aussi dans le Rhône, les Alpes-Maritimes, le Bas-Rhin, la Marne, la Côte-d'Or ou la Haute-Savoie. La carte des dépassements tarifaires est également très liée à celle des médecins secteur 2 avec une diagonale allant de la Normandie en passant par la région parisienne et allant jusqu'à la région Rhône-Alpes.

2.2.3 Étude des corrélations entre variables explicatives

L'étude des corrélations a pour objectif de détecter la présence potentielle de multicolinéarité, c'est à dire d'identifier des variables expliquant le même phénomène. Dans le cadre d'une régression, une forte multicolinéarité peut poser problème lors de l'estimation des paramètres d'un modèle, les coefficients peuvent alors être instables et leur interprétation difficile.

Afin de détecter la présence éventuelle de multicolinéarité, il est nécessaire dans un premier temps d'identifier les variables fortement liées entre elles. Le test du $\chi 2$ permet de tester la présence d'une relation entre deux variables mais il ne donne pas d'information sur l'intensité de cette relation. Pour répondre à cette problématique, le test de Cramer est plus adapté. Ses valeurs sont comprises entre 0 et 1, plus sa valeur est élevée plus la relation entre les variables est forte.

Le V de cramer correspond au rapport du $\chi 2$ sur le χ^2 max. Le χ^2 max correspond au χ^2 maximal qui peut être observé sur le tableau de contingence. Ce χ^2 max théorique est égal à l'effectif multiplié par le plus petit côté du tableau.

$$V \text{ de cramer} : \ V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

$$avec \ \chi^2 = \sum_{ij} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Le tableau ci-dessous des V de Cramer sur les variables explicatives permet d'identifier les relations fortes entre variables.

	Réseau de distribtion	Niveau	Age moyen des salariés	Nombre de salariés	Collège	Zonier	Secteur d'activité	Lien de parenté	Détenteur d'option	Régime
Réseau de distribtion	100%									
Niveau -	21%	100%								
Age moyen des salariés -	4%	8%	100%							
Nombre de salariés-	16%	8%	17%	100%						
Collège-	8%	22%	18%	21%	100%					
Zonier-	12%	9%	3%	2%	5%	100%				
Secteur d'activité-	10%	11%	7%	12%	11%	6%	100%			
Lien de parenté-	2%	11%	12%	3%	8%	2%	4%	100%		
Détenteur d'option-	4%	47%	3%	7%	9%	1%	9%	3%	100%	
Régime-	4%	10%	3%	5%	4%	21%	4%	1%	3%	100%

TABLE 2.6 – V de cramer entre les variables explicatives

Un V de cramer inférieur à 30% entre deux variables indique un lien suffisamment faible entre les variables et ne pose pas de problème pour la modélisation. Sur le tableau, un lien important se dégage entre la détention d'option et le niveau de garantie, le V de Cramer est de 47%. Cependant cette relation forte s'explique par le fait que les assurés qui souscrivent une option sont des assurés avec un niveau de couverture faible.

L'analyse croisée de la consommation médicale entre le niveau de garantie et la détention d'option montre que ces 2 variables n'expliquent pas le même phénomène, la détention d'option explique une surconsommation sur tous les niveaux.

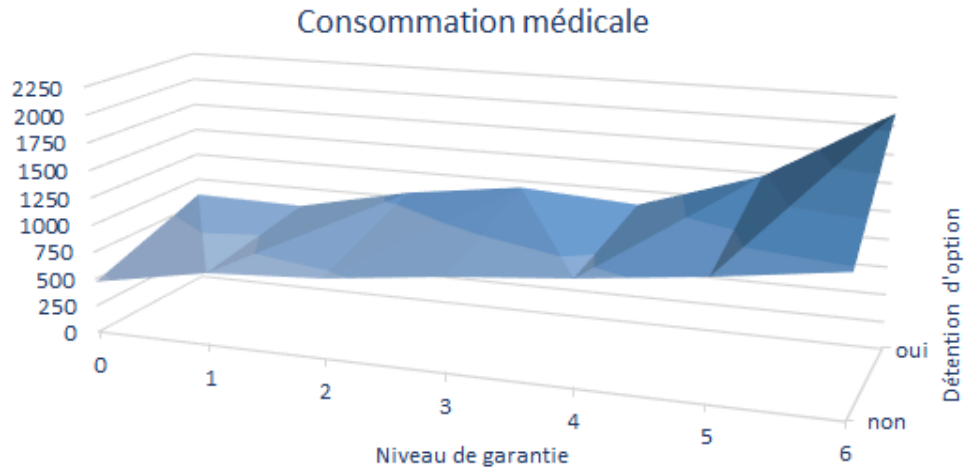


FIGURE 2.23 – Consommation médicale en fonction du niveau de garantie et de la détention d'option

Chapitre 3

Modélisation de la prime pure

La prime pure constitue le socle de l'élaboration du tarif, elle est à l'origine de sa finesse et de sa précision. La modélisation de la prime pure représente donc un enjeu fondamental, la qualité de sa détermination conditionne les souscriptions ainsi que les résultats futurs.

La méthode parfaite de modélisation de la prime pure n'existe pas, il existe un ensemble de techniques disponibles chacune présentant des avantages et des inconvénients. La technique la plus fréquemment utilisée repose sur les modèles linéaires généralisés. D'autres techniques existent notamment celles appartenant à la famille des méthodes dites d'apprentissage. Parmi elles, figurent notamment l'algorithme CART, l'algorithme des forêts aléatoires (Random Forest) et les réseaux de neurones.

De même qu'il existe différentes techniques de modélisation, la modélisation de la prime pure peut être réalisée selon différentes approches. Une première approche consiste à modéliser directement la prime pure et une seconde approche dite modélisation coût-fréquence consiste à modéliser séparément la fréquence et le coût moyen. Dans cette seconde approche sous condition d'indépendance entre la fréquence et le coût moyen, la prime pure modélisée correspond au produit de la fréquence moyenne modélisée par le coût moyen modélisé. Une troisième approche moins fréquente peut consister à modéliser la probabilité de consommer et le coût total.

Dans ce chapitre, trois techniques de modélisation citées précédemment seront abordées, la méthode GLM ainsi que les méthodes CART et RANDOM FOREST. Après une description générale de leur principe, elles feront chacune l'objet d'une application en faisant varier le type d'approche. Une analyse comparative de leurs résultats sera réalisée dans le dernier chapitre.

3.1 Sélection de l'approche pour la modélisation de la prime pure

3.1.1 Vérification de la condition d'indépendance entre la fréquence et le coût moyen

Comme évoqué précédemment, la condition à vérifier lors d'une modélisation de type fréquence-coût moyen est l'indépendance entre la fréquence et le coût moyen. Avant de procéder à cette vérification, quelques rappels théoriques induisant cette condition sont nécessaires.

3.1.1.1 Rappel théorique

Soit S le montant total de la charge de sinistres, alors S peut s'écrire sous la forme :

$$S = \sum_{i=1}^{N} C_i$$

avec :

- N une variable aléatoire représentant le nombre de sinistres survenus pendant la période observée.

- C_i le coût unitaire du sinistre i. Le coût de chaque sinistre est supposé être une variable aléatoire indépendante et identiquement distribuée.

Sous condition d'indépendance de la fréquence et du coût moyen, la prime pure à l'intérieur d'une classe de risque est de la forme :

$$E[S|X] = E[N|X] \ \times \ E[C|X]$$

avec X les caractéristiques de l'individu.

3.1.1.2 Test d'indépendance

Afin d'identifier un lien éventuel entre la fréquence et le coût moyen, le test du khi-deux apparaît bien approprié. Pour le réaliser, les variables doivent être qualitatives, il est donc nécessaire de réaliser un découpage de la fréquence et du coût moyen en classe. Ces classes seront élaborées à partir des déciles.

Ci-dessous le résultat du test pour les 6 postes de consommation :

Poste	χ^2	**Ddl**	**P-value**
Hospitalisation	3099.51	18	<.0001
Soins courants	27118.41	63	<.0001
Pharmacie	7190.42	54	<.0001
Dentaire	7645.89	18	<.0001
Optique	1063.86	9	<.0001
Bien-être	1815.39	14	<.0001

TABLE 3.1 – Résultat du test du khi-deux

La condition d'indépendance entre la fréquence et le coût moyen est vérifiée sur l'ensemble des postes permettant d'envisager une modélisation coût-fréquence.

3.1.2 Choix de l'approche sur les différents postes

L'approche fréquence - coût moyen est possible sur l'ensemble des postes. D'autres approches sont également envisageables, la modélisation directe de la prime pure, la modélisation de la probabilité de consommer et du coût total. Cette dernière approche pourrait apparaître particulièrement adaptée au poste optique. En effet sur ce poste, la fréquence des consommants est souvent égale à un acte, un acte correspondant à un équipement optique complet composé d'une monture et de deux verres. Cependant afin de conserver le caractère multiplicatif du tarif avec la technique GLM, l'approche fréquence-coût moyen a tout de même été retenue pour ce poste.

Dans le cadre de la modélisation GLM, seule l'approche fréquence - coût moyen a été mise en œuvre mais l'approche prime pure aurait été possible avec les modèles GLM et la famille des lois Tweedie.

Concernant les techniques CART et Random Forest, les deux approches fréquence - coût moyen et prime pure seront testées. Pour ces deux techniques, l'approche directe de la modélisation de la prime pure est souvent privilégiée mais l'application avec l'approche fréquence - coût moyen permettra de comparer la performance avec le GLM en évitant un biais lié à l'approche.

3.2 Modélisation de la prime pure par poste avec la méthode GLM

Les modèles linéaires généralisés (Generalized Linear Models, GLM) sont une généralisation des modèles de régression linéaire. Leurs noms a été introduit pour la première fois par Nelder et Wedderburn en 1972. Ces modèles apparaissent particulièrement bien adaptés à la modélisation des risques non vie. Ils permettent de modéliser une variable Y, la fréquence de sinistres, le coût des sinistres en fonction de caractéristiques de l'assuré.

3.2.1 La théorie des modèles linéaires généralisés (GLM)

3.2.1.1 Principe généraux d'un modèle linéaire généralisé

Les modèles GLM sont une extension des modèles de régression linéaires mais ils sont beaucoup moins contraignants. En effet, ils présentent l'avantage de permettre de s'affranchir de certaines hypothèses. Tout d'abord, ils n'imposent pas de relation de la forme linéaire entre la variable à expliquer et les variables explicatives. Ensuite, l'hypothèse de normalité des observations n'est pas requise. Enfin, la variance des observations n'est pas nécessairement constante. C'est le cas des variables aléatoires suivant une loi de poisson, loi fréquemment utilisée pour modéliser la fréquence de sinistres et pour lesquelles la variance est fonction de la moyenne.

Les modèles GLM sont composés de 3 composantes :

- la composante aléatoire
- la composante déterministe
- la fonction de lien

La composante aléatoire

Il s'agit de la variable réponse $\mathrm{Y} = (Y_1, ..., Y_n)$ composée de n observations indépendantes suivant une loi de probabilité de la famille exponentielle. Cette propriété de la variable Y constitue une des hypothèses fondamentales des modèles GLM. Cette condition implique que la variable Y admet une fonction de densité de la forme :

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

avec θ appelé paramètre naturel, ϕ paramètre de dispersion ou d'échelle et a(.), b(.) et c(.) des fonctions réelles. Cette forme de fonction de densité s'applique à la plupart des lois usuelles : Normale, Poisson, Exponentielle, Bernoulli, Binomiale, Binomiale négative,...

Loi	θ	$a(\phi)$	$b(\theta)$	**c(y,ϕ)**
$N(\mu,\sigma^2)$	μ	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2\right)$
$Pois(\lambda)$	$ln(\lambda)$	1	e^{θ}	$-ln(y!)$
Bin(n,p)	$\ln(\frac{p}{1-p})$	1	$nln(1+e^{\theta})$	$\ln\binom{n}{y}$
$Gamma(\nu, \mu)$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$	$-\ln(-\theta)$	$\nu ln(\nu y) - ln(y) - ln(\Gamma(\nu))$

TABLE 3.2 – Tableau de lois de la famille exponentielle

L'espérance et la variance de la variable Y suivant une loi exponentielle sont de la forme :

$$\mathrm{E(Y)} = b'(\theta) = \frac{db(\theta)}{d\theta} \text{ et } \mathrm{Var(Y)} = b''.a(\phi)$$

La composante déterministe

Elle se compose des variables explicatives $X_1, ..., X_k$ utilisées comme prédicteurs sous la forme d'une combinaison linéaire de type :

$$\eta_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$$

ou sous la forme matricielle :

$$\eta = X^t\beta$$

La matrice X comporte les n valeurs des variables explicatives et β est le vecteur des paramètres du modèles.

La fonction de lien

Elle décrit la relation entre l'espérance conditionnelle de la variable Y et la combinaison linéaire des variables explicatives. Soit μ_i l'espérance de Y_i, alors μ_i est liée à η_i par la fonction de lien :

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$$

Cette fonction de lien est monotone et dérivable. Les prédictions de Y sont obtenues en appliquant la fonction inverse de la fonction lien g^{-1} :

$$\mu_i = E[Y_i] = g^{-1}(\eta_i)$$

Les fonctions de lien les plus fréquemment utilisées sont listées dans le tableau ci-dessous.

Fonction de lien	**g(μ)**	**Loi de Y|X**
Identité	μ	Normale
log	$\ln(\mu)$	Poisson
Inverse	$\frac{1}{\mu}$	Exponentielle
Puissance	g(μ)	Gamma
Logit	$\ln(\frac{\mu}{1-\mu})$	Binomiale

TABLE 3.3 – Tableau des fonctions de lien usuelles

Le choix de la fonction de lien est propre à la problématique traitée et notamment à l'échelle de valeur de la variable à modéliser. Dans le cas de la modélisation d'une probabilité, Y est comprise entre 0 et 1, la fonction de lien la plus adaptée sera la fonction de lien logit. En revanche, dans le cas de la modélisation de sinistres, l'échelle de valeur de Y se situe sur R^+, c'est la fonction de lien log qui sera retenue et conduira au modèle multiplicatif suivant :

$$\mathrm{E[Y|x]} = \exp(X^t\beta)$$

3.2.1.2 Estimation des paramètres d'un modèle linéaire généralisé

L'estimation des paramètres d'un modèle GLM est réalisée à l'aide de la méthode du maximum de vraisemblance.

Soit $(y_1, ..., y_n)$ un échantillon de variables aléatoires indépendantes de taille n et identiquement distribuées selon une loi de la famille exponentielle, alors la fonction vraisemblance qui est le produit de densités peut s'écrire sous la forme suivante :

$$L(\theta, \phi, y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i|\theta, \phi)$$

L'objectif est de déterminer les paramètres permettant de maximiser cette fonction. La vraisemblance étant positive et le logarithme étant une fonction croissante, par simplification il est préférable de maximiser le logarithme de celle-ci.

$$ln(L(\theta, \phi, y_1, ..., y_n)) = \sum_{i=1}^{n} \left(\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right)$$

L'estimation des coefficients β permettant de maximiser la fonction de vraisemblance consiste à résoudre l'équation qui annule la dérivée de la fonction par rapport au paramètre recherché et à s'assurer que la dérivée seconde est négative pour être en présence d'un maximum.

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

Les équations qui déterminent les paramètres au sens du maximum de vraisemblance sont souvent non linéaires. La solution pour résoudre ces équations réside dans l'utilisation d'algorithme itératif. L'algorithme de Newton-Raphson est l'algorithme le plus connu permettant de résoudre ce système d'équation.

3.2.1.3 Intervalle de confiance des paramètres

L'estimation des coefficients avec la méthode du maximum de vraisemblance implique les propriétés suivantes :

- ils sont asymptotiquement sans biais
- ils sont asymptotiquement normaux

$$\sqrt{n}\ (\widehat{\beta}_n - \beta)\ \xrightarrow[n\to+\infty]{\text{Loi}}\ N\left(0, I(\beta)^{-1}\right)$$

avec $I(\beta)^{-1}$ la matrice d'information de Fisher.

Cette propriété de convergence en loi permet de déterminer un intervalle de confiance asymptotique de niveau (1-α) pour le coefficient β_j :

$$IC = \left[\widehat{\beta}_j - z_{1-\frac{\alpha}{2}} * \sqrt{\nu_{jj}} \; ; \; \widehat{\beta}_j + z_{1-\frac{\alpha}{2}} * \sqrt{\nu_{jj}}\right]$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile $(1 - \frac{\alpha}{2})$ de la loi normale N(0,1)

et ν_{jj} le j^{eme} terme diagonal de la matrice de Fisher correspondant à la variance de $\widehat{\beta}_j$.

3.2.1.4 Critères de choix du modèle

La sélection du meilleur modèle est une étape importante qui nécessite de s'appuyer sur un critère statistique permettant de comparer la qualité de l'ajustement de différents modèles entre eux. Il n'existe pas de critère universel permettant de définir la notion de meilleur modèle. Le meilleur modèle sera propre au critère choisi, on parlera donc de meilleur modèle au sens d'un critère donné.

Généralement, les critères s'appuient sur la notion de vraisemblance. Cependant, l'utilisation de la vraisemblance comme seul critère de choix du meilleur modèle n'est pas suffisante car cela conduit à sélectionner systématiquement le modèle le plus complexe. Un modèle trop complexe s'avérera souvent moins robuste, ses résultats pourront être excellents sur un échantillon d'apprentissage mais s'avérer beaucoup moins satisfaisants sur un échantillon test.

Afin d'éviter ce phénomène, les critères AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion) ont été élaborés dans les années 70. Il s'agit de critères pénalisant la vraisemblance par la complexité, ils permettent de prendre en compte à la fois la notion de performance et la notion de complexité. Ils reflètent un certain compromis entre une bonne performance et une moindre complexité.

$$AIC \; = \; -2.ln \; L(\hat{\mu}|\, Y) \; + 2p$$

$$BIC \; = \; -2.ln \; L(\hat{\mu}|\, Y) \; + p \; log \; n$$

avec p le nombre de paramètres du modèle.

Le modèle retenu sera celui qui minimise l'AIC ou le BIC, le critère BIC ayant tendance à pénaliser davantage la complexité.

3.2.1.5 Sélection des variables

La sélection des variables dans le modèle est une étape cruciale qui peut s'avérer complexe si le nombre de variables candidates est élevé. En effet, tester l'ensemble des combinaisons possibles de variables dans un modèle revient à tester 2^p combinaisons ce qui peut rapidement s'avérer difficile à mettre en œuvre et nécessiter des temps de traitement très conséquents. L'une des solutions permettant de répondre à cette problématique consiste à utiliser des méthodes de sélection automatique des variables. Ce sont des méthodes pas à pas qui s'appuient sur des critères de comparaison de qualité de modèle tel que l'AIC, le BIC ou la déviance.

La méthode Forward

La méthode *Forward* est une méthode ascendante. Le modèle de départ comporte uniquement la constante, les variables sont ensuite intégrées de manière itérative en fonction de leur capacité à améliorer la qualité du modèle au sens du critère choisi. Les variables induisant la meilleure qualité du modèle seront introduites en premier, le processus s'achevant lorsque l'introduction de nouvelles variables ne permet plus d'améliorer la qualité du modèle. L'inconvénient de cette méthode est qu'une variable introduite dans le modèle ne peut plus être supprimée, ce qui peut induire la présence de variables non significatives.

La méthode Backward

A l'inverse de la méthode *Forward*, la méthode *Backward* est une méthode descendante. Le modèle de départ comporte l'ensemble des variables candidates, les variables sont ensuite retirées une à une en priorisant les variables induisant la plus faible dégradation du modèle et tant que cette dégradation n'est pas statistiquement significative. L'inconvénient de cette méthode est qu'il n'est plus possible de réintroduire une variable qui a été supprimée.

La méthode Stepwise

La méthode *Stepwise* est une méthode alternative qui permet de répondre aux principaux défauts des deux approches citées précédemment. Il s'agit d'une amélioration de la méthode ascendante dans le sens où à chaque nouvelle introduction de variable, un réexamen des variables déjà présentes dans le modèle est effectué. Du fait de corrélation pouvant exister entre les variables, l'ajout ou la suppression d'une variable peut modifier la significativité des variables déjà présentes dans le modèle. La procédure permettra après l'ajout d'une nouvelle variable explicative dans le modèle de vérifier la significativité des variables et si besoin de retirer les variables non significatives en commençant par la moins significative d'entre elles. Le processus s'arrête lorsque l'ajout ou la suppression de variables ne permet plus d'améliorer significativement le modèle.

Les 3 méthodes citées précédemment ne sont donc pas nécessairement optimales car elles ne permettent pas de tester l'ensemble des combinaisons possibles. Par ailleurs, ces approches ne se substituent pas à l'expertise, elles sont complémentaires et constituent une aide à la décision. Dans cette étude, l'approche qui a été retenue pour la sélection des variables est l'approche Stepwise.

3.2.1.6 Qualité d'ajustement d'un modèle linéaire généralisé

La qualité de l'ajustement est mesurée en comparant la différence entre les estimations et les observations. La déviance et le test de Person sont deux critères permettant de mesurer la qualité de l'ajustement.

La déviance

Pour mesurer la qualité de l'ajustement d'un modèle GLM, on utilise souvent la déviance. Le modèle estimé est comparé avec le modèle saturé, c'est à dire le modèle possédant autant de paramètres que d'observations et qui estime donc parfaitement les données.

$$D \;=\; 2 \times (\; ln\; L(Y|\, Y) \;-\; ln\; L(\hat{\mu}|\, Y)\;)$$

Une valeur de D positive et faible traduira une bonne qualité de modèle. Cette statistique suit asymptotiquement une loi du Khi-2 à n-p-1 degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle. L'analyse doit être complétée par une analyse des résidus.

Test de Pearson

La statistique de Pearson est également utilisée pour comparer les valeurs observée y_i avec leur prédiction par le modèle, elle est définie par la formule suivante :

$$\chi^2 \;=\; \sum_{i=1}^{n} \frac{(y_i \;-\; \widehat{\mu}_i)^2}{\widehat{Var}(\widehat{\mu}_i)}$$

Cette statistique suit une loi du Khi-2. L'espérance de la loi du khi-2 est son nombre de degré de liberté, le modèle sera jugé satisfaisant pour un rapport D/ddl plus petit que 1.

Les résidus

Les résidus peuvent être calculés de différentes manières. Les deux principales méthodes sont les résidus de Pearson et les résidus de la déviance.

- Résidus de Pearson $\;r_i^p = \sqrt{\omega_i}\dfrac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$

- Résidus de déviance $\;r_i^d = \varepsilon(y_i - \mu_i)\sqrt{d_i}$

3.2.2 Modélisation de la fréquence

3.2.2.1 Choix de la loi

Généralement, la modélisation de la fréquence en IARD fait intervenir deux types de lois particulièrement adaptées aux données discrètes, la loi de Poisson et la loi Binomiale Négative.

Une première analyse permettant d'identifier la loi présentant la meilleure qualité d'ajustement aux données consiste à comparer la moyenne et la variance de la fréquence observée par poste. En effet, une des propriétés fondamentales de la loi de poisson est l'égalité entre la moyenne et la variance. Ainsi, une moyenne et une variance empirique proche conduira plûtot à opter pour une loi de Poisson. En revanche, une variance beaucoup plus élevée que la moyenne traduira un phénomène de sur-dispersion, et dans ce cas la loi Binomiale Négative sera vraisemblablement la mieux adaptée.

Poste de garantie	**Moyenne (μ)**	**Variance (σ^2)**
Hospitalisation	0.33	0.94
Soins courant	4.95	42.59
Pharmacie	3.57	20.32
Dentaire	0.72	2.56
Optique	0.25	0.29
Bien-être	0.33	0.72

TABLE 3.4 – Tableau des moyennes et variances empiriques de la fréquence par poste

La comparaison de la moyenne et de la variance met en évidence un phénomène de sur-dispersion sur les 5 postes concernés par la modélisation de la fréquence. Par exemple, sur le poste soins courants, la variance est égale à 8 fois la moyenne. Ce constat semble traduire un meilleur ajustement des données avec la loi Binomiale Négative mieux adaptée aux données présentant une sur-dispersion.

Une autre analyse graphique permet de guider le choix de la loi présentant le meilleur ajustement. Elle consiste à comparer la distribution de la loi théorique avec celle des données empiriques. Ci-après, la comparaison de l'ajustement de la loi empirique avec les lois théoriques Poisson et Binomiale Négative sur le poste soins courants. Les graphiques associés aux autres postes sont disponibles en annexe C.

Sur les graphiques 3.1 et 3.2, la courbe en rouge correspond à la loi théorique et l'histogramme en gris aux fréquences observées. Afin de mieux visualiser l'écart d'ajustement, il a été opté pour une représentation de la racine carrée de la fréquence, cela permet de réduire l'écrasement des données sur les fréquences plus faibles. Par ailleurs, l'histogramme des fréquences observées a été ajusté sur la courbe théorique en rouge et non pas sur la ligne des abscisses. L'écart entre la base de l'histogramme et la ligne des abscisses permet de juger de la qualité de l'adéquation.

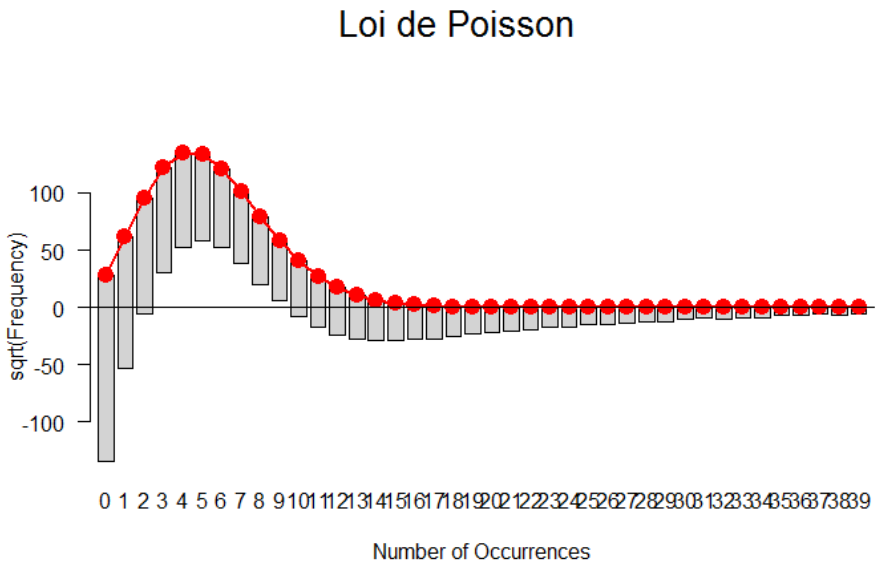


FIGURE 3.1 – Soins Courants - Ajustement de la fréquence par une loi de Poisson

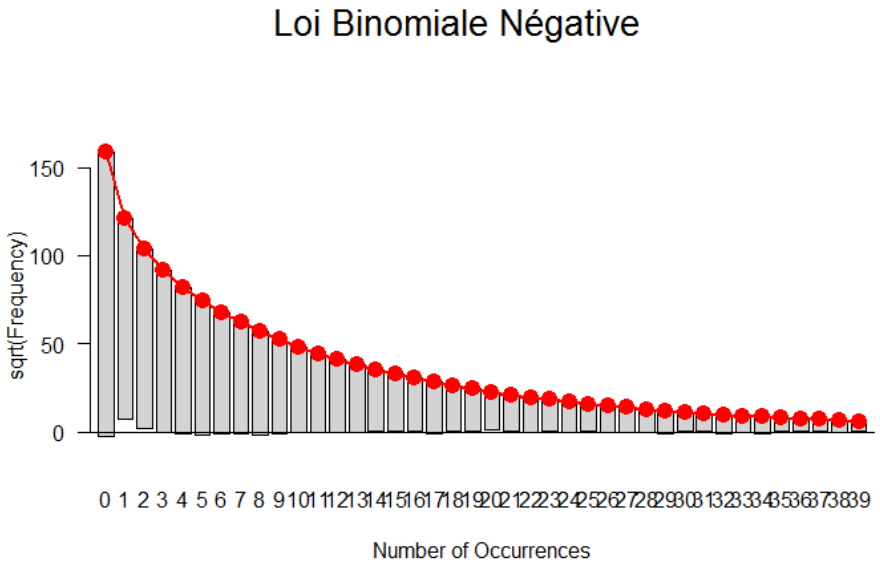


FIGURE 3.2 – Soins Courants - Ajustement de la fréquence par une loi Binomiale Négative

L'analyse graphique montre une meilleure adéquation avec la loi Binomiale Négative. La base de l'histogramme des fréquences observées est proche de la ligne des abscisses.

Le test d'ajustement du khi-deux sur le poste soins courants entre les données empiriques et les lois théoriques rejette l'adéquation des données avec les deux lois.

Statistiques	**Binomiale Négative**	**Poisson**
χ^2	341.04	5.8e6
Df	50	14
P-value	$<$ 2.2e-16	$<$ 2.2 e-16

TABLE 3.5 – Résultats du test du khi-deux

Malgré le rejet du test, les résultats permettent de conclure à une meilleure adéquation des données à la loi Binomiale Négative. Ce constat est également observé sur les 5 autres postes dont les résultats figurent en annexe.

Poste de soins	**Type de modélisation**	**Loi**	**Fonction de lien**
Hospitalisation	Fréquence	Binomiale Négative	Log
Soins Courants	Fréquence	Binomiale Négative	Log
Pharmacie	Fréquence	Binomiale Négative	Log
Dentaire	Fréquence	Binomiale Négative	Log
Optique	Fréquence	Binomiale Négative	Log
Bien-Être	Fréquence	Binomiale Négative	Log

TABLE 3.6 – Lois retenues pour la modélisation de la fréquence par poste

3.2.2.2 Prise en compte de l'exposition

Afin de prendre en compte l'exposition, la fonction de lien retenue étant de type log, une variable offset sous la forme *log(exposition)* a été intégrée dans la modélisation de type fréquence.

3.2.2.3 Sélection des variables

La sélection des variables est réalisée à l'aide de la méthode stepwise présentée précédemment dans la section 3-2-1-5 et en utilisant le critère AIC. Le résultat ci-dessous de la procédure stepwise pour la modélisation de la fréquence sur le poste soins courants montre que l'ensemble des variables ont été retenues. En revanche, sur le poste hospitalisation la sélection stepwise a permis d'exclure la variable du nombre de salariés non significative.

Modèle	**AIC**
Modèle Complet	150 394
Modèle Stepwise	150 390

TABLE 3.7 – Résultat de la procédure stepwise - poste hospitalisation

Modèle	**AIC**
Modèle Complet	555 468
Modèle Stepwise	555 468

TABLE 3.8 – Résultat de la procédure stepwise - poste soins courants

La sélection des variables est une première étape de l'élaboration du modèle. Elle permet de sélectionner les variables significatives mais cette étape n'est pas suffisante. Une seconde étape est nécessaire afin de s'assurer que l'ensemble des modalités des variables retenues sont significatives.

3.2.2.4 Regroupement des modalités

Dans cette seconde étape, les modalités non significatives seront regroupées avec d'autres modalités. Les modalités des variables suivent une loi normale, elle seront retenues si elles sont significativement différentes de 0 au seuil $\alpha = 5\%$. Lors de cette étape, il est également important de s'assurer de la pertinence des effets. Sur le poste soins courants, un certain nombre de modalités sont apparues non significatives ou alors avec des coefficients très proches. Par exemple, en ce qui concerne la zone géographique, les zones 3 et 4 avaient un coefficient très proche, les deux modalités ont donc été regroupées.

3.2.2.5 Résultats des modèles de fréquence

Deux modélisations par poste ont été réalisées. Une première modélisation a consisté à intégrer chacune des variables dans le modèle sans effectuer de retraitement. Cette modélisation correspond à la modélisation actuelle du tarif en vigueur. Une seconde modélisation a également été réalisée afin de neutraliser la sensibilité du tarif des enfants de salariés avec

l'âge moyen des salariés. Les résultats de la modélisation de la fréquence avec la loi binomiale négative permettent d'attribuer des coefficients pour les différentes modalités ainsi que l'écart type et la p-value.

Résultat du modèle sans retraitement

Ci-dessous (tableau 3.9) les résultats de la modélisation sans retraitement de la fréquence sur le poste soins courants.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.106346	0.026270	42.114	< 2e-16	***
niveau 1	0.118407	0.016290	7.269	3.63e-13	***
niveau 2	0.239725	0.015062	15.916	< 2e-16	***
niveau 3	0.289011	0.013941	20.730	< 2e-16	***
niveau 4	0.302317	0.015489	19.519	< 2e-16	***
niveau 5	0.327788	0.018475	17.742	< 2e-16	***
niveau 6	0.342227	0.020042	17.076	< 2e-16	***
Nombre de salariés : 1 salarié	0.063752	0.011924	5.346	8.97e-08	***
Nombre de salariés : 2 salariés	0.046196	0.012486	3.700	0.000216	***
Age : 30-34 ans	0.112435	0.017625	6.379	1.78e-10	***
Age : 35-39 ans	0.150173	0.016592	9.051	< 2e-16	***
Age : 40-44 ans	0.171932	0.016495	10.423	< 2e-16	***
Age : 45-49 ans	0.212038	0.017780	11.926	< 2e-16	***
Age : 50 ans et plus	0.297465	0.018753	15.862	< 2e-16	***
Zone 2	0.061579	0.010616	5.801	6.60e-09	***
Zones 3 et 4	0.097166	0.008919	10.895	< 2e-16	***
Secteur 2	0.124020	0.012922	9.597	< 2e-16	***
Secteur 3	0.306640	0.015714	19.514	< 2e-16	***
Catégorie : ensemble du personnel +Non cadre	0.058563	0.012171	4.812	1.50e-06	***
Lien parenté : Conjoint	0.185125	0.010713	17.281	< 2e-16	***
Lien parenté : Enfant	-0.222972	0.008407	-26.522	< 2e-16	***
Détention Option : oui	0.263975	0.013213	19.979	< 2e-16	***
Régime Alsace-Moselle	0.115349	0.023576	4.893	9.94e-07	***

TABLE 3.9 – Soins Courants - coefficients estimés sur le modèle GLM

La hiérarchie de valeurs des coefficients des différentes modalités des variables est conforme à l'attendu et confirme ce qui a été observé dans le cadre de l'analyse descriptive. Par exemple, les coefficients du modèle sont croissants avec l'âge moyen des salariés, ce qui valide l'observation de l'augmentation de la consommation médicale selon l'âge moyen des salariés. Ce constat se confirme également pour le niveau de garantie.

A partir de l'écart-type de chaque estimateur, il est possible de définir un intervalle de confiance pour chaque coefficient. Ci-dessous, les graphiques des coefficients avec leur intervalle de confiance sur le poste soins courants.

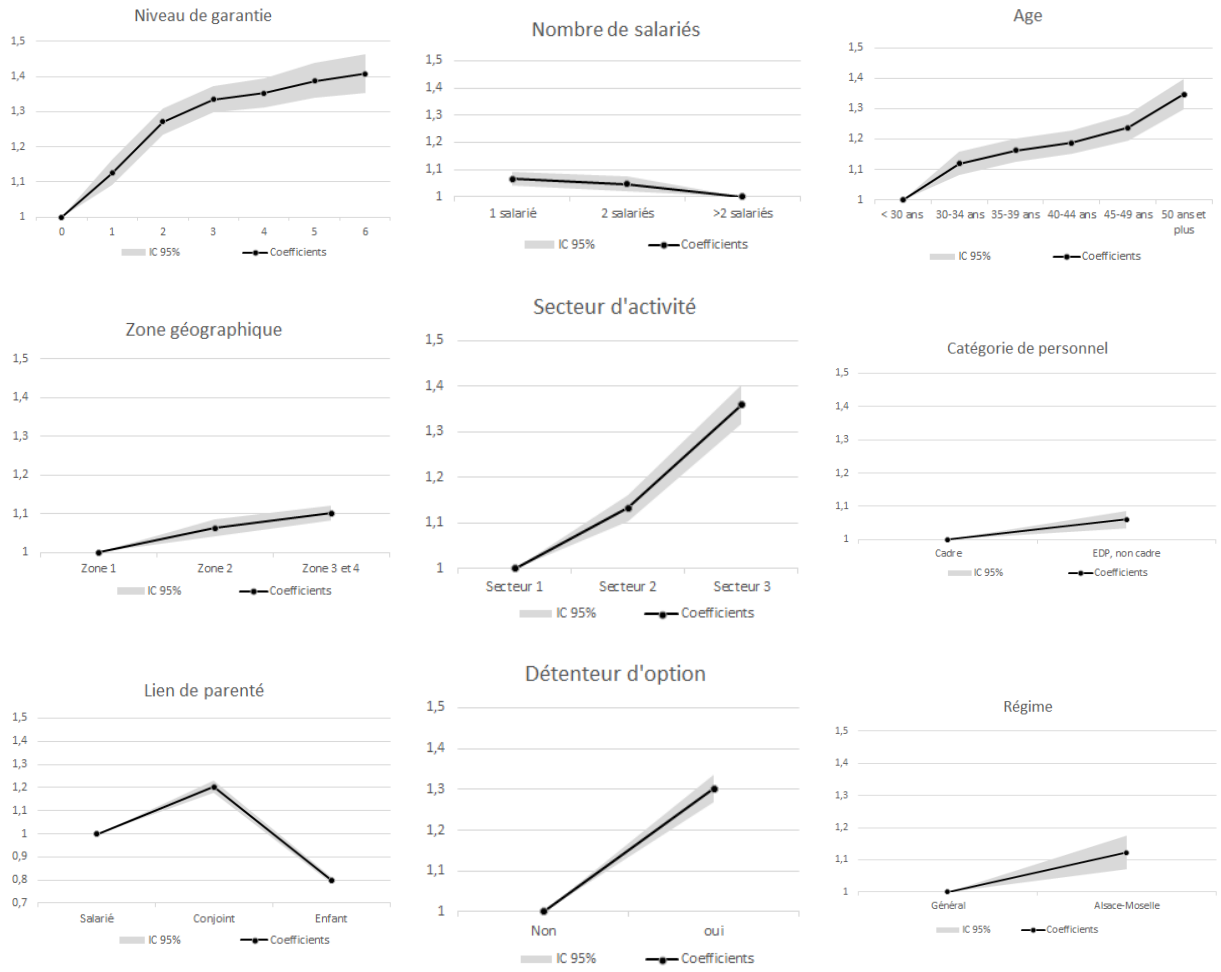


FIGURE 3.3 – Soins Courants - Intervalle de confiance des paramètres du modèle fréquence

Le graphique permet également d'identifier les variables les plus influentes, à savoir l'âge moyen des salariés, le niveau de garantie, le lien de parenté, le secteur d'activité et la détention d'option.

Résultat du modèle avec retraitement sur le lien de parenté

Dans la modélisation précédente, l'intégration des variables « âge moyen des salariés » et « lien de parenté » de manière indépendante introduit un biais sur l'élaboration du tarif des enfants de salariés. En effet, le tarif des enfants était sensible à l'âge moyen des salariés. Ainsi, le tarif des enfants de salariés se trouve majoré dans les entreprises où l'âge moyen des salariés est élevé et minoré dans les entreprises où l'âge moyen des salariés est bas. Afin de palier à ce biais, la modalité « Enfant » du lien de parenté a été intégrée au sein de la variable « âge moyen des salariés » et une indicatrice « Lien de parenté : Conjoint » a été créée.

Le tableau ci-dessous (figure 3.10) donne les résultats de la modélisation de la fréquence sur le poste soins courants. La modalité « Enfant » n'apparaît pas dans le tableau ci-dessous, elle n'est pas significativement différente de la modalité de référence correspondant au moins de 30 ans.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.052905	0.022576	46.638	< 2e-16	***
niveau 1	0.119311	0.016258	7.339	2.16e-13	***
niveau 2	0.240508	0.015031	16.001	< 2e-16	***
niveau 3	0.287902	0.013886	20.733	< 2e-16	***
niveau 4	0.302111	0.015413	19.602	< 2e-16	***
niveau 5	0.326577	0.018403	17.746	< 2e-16	***
niveau 6	0.338295	0.019974	16.936	< 2e-16	***
Nombre de salariés : 1 salarié	0.055444	0.011852	4.678	2.89e-06	***
Nombre de salariés : 2 salariés	0.041604	0.012426	3.348	0.000814	***
Age : 30-34 ans	0.126922	0.012980	9.779	< 2e-16	***
Age : 35-39 ans	0.180937	0.010915	16.578	< 2e-16	***
Age : 40-44 ans	0.244610	0.010634	23.004	< 2e-16	***
Age : 45-49 ans	0.299829	0.013327	22.497	< 2e-16	***
Age : 50 ans et plus	0.424520	0.014543	29.191	< 2e-16	***
Zone 2	0.061221	0.010587	5.783	7.36e-09	***
Zones 3 et 4	0.096723	0.008896	10.872	< 2e-16	***
Secteur 2	0.123128	0.012899	9.546	< 2e-16	***
Secteur 3	0.302548	0.015669	19.309	< 2e-16	***
Catégorie : EDP, non cadre	0.061378	0.012028	5.103	3.35e-07	***
Lien parenté : Conjoint	0.176719	0.010698	16.518	< 2e-16	***
Détention Option : oui	0.258709	0.013173	19.640	< 2e-16	***
Régime Alsace-Moselle	0.113513	0.023540	4.822	1.42e-06	***

TABLE 3.10 – Soins Courants - coefficients estimés sur le modèle GLM avec retraitement

3.2.2.6 Validation des modèles de type fréquence

3.2.2.7 La déviance

Le rapport de la déviance sur le nombre de degré de liberté est de 1,11 sur le modèle finale sans retraitement, cette valeur est proche de 1 ce qui permet de conclure que le modèle est satisfaisant.

Modèle	Déviance	ddl	D/ddl
Modèle final	123 580	110 898	1,11

TABLE 3.11 – Valeur de la déviance des modèles

3.2.2.8 Étude des résidus

L'étude des résidus permet d'identifier d'éventuels problèmes comme la présence d'une tendance ou des problèmes d'hétéroscédasticité.

La figure 3.4 montre que l'essentiel des résidus sont compris entre -2 et 2 correspondant à l'intervalle de confiance à 95% de la loi normale. Les lignes observées sur le graphique s'explique par le caractère discret de la variable modélisée. La ligne la plus basse correspond aux observations pour lesquelles la fréquence observée sur le poste soins courants est de 0. Le graphique 3.5 indique une fréquence importante des résidus à droite de l'abscisse -2, ces résidus s'expliquent par le nombre important de 0. Le modèle demanderait a être amélioré afin de tenir compte de la fréquence importante de non consommants. Les modèles du type zéro inflated négative binomiale (ZINB) peuvent permettre d'améliorer les résultats. Ils permettent de prendre en compte la présence excessive de 0.

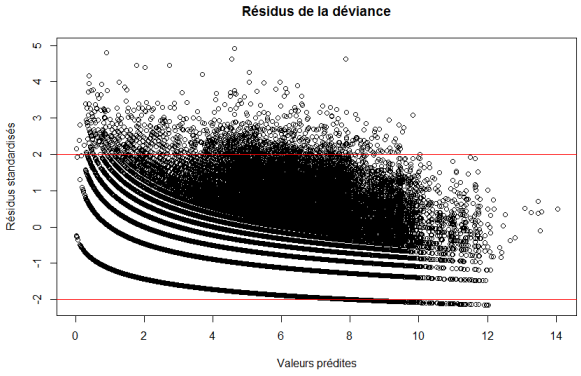


FIGURE 3.4 – Soins Courants - résidus standardisés de la déviance sur le modèle sans retraitement

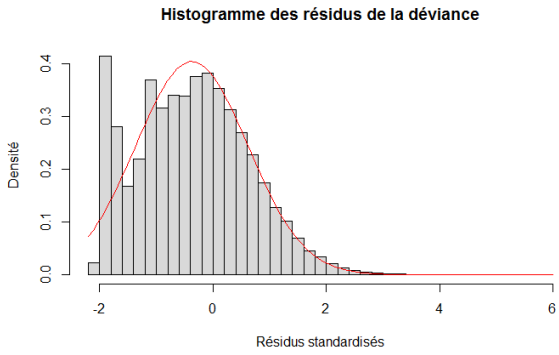


FIGURE 3.5 – Soins Courants - histogramme des résidus de la déviance sur le modèle sans retraitement

3.2.3 Modélisation du coût moyen

3.2.3.1 Choix de la loi

De même que pour la fréquence, il existe également deux lois souvent utilisées pour la modélisation du coût moyen, il s'agit des lois Gamma et Log-Normale. Afin de déterminer celle qui convient le mieux, différentes analyses graphiques ainsi que des tests d'adéquation seront réalisés.

Une première analyse graphique consiste à comparer l'histogramme du coût moyen avec les deux lois théoriques. Sur la figure 3.6 concernant le poste soins courants, il semble difficile d'établir laquelle des deux lois s'ajuste le mieux. La loi Gamma apparaît mieux adaptée sur les coûts moyens inférieurs à 7 euros mais la log-normale semble mieux adaptée sur la queue de distribution.

L'erreur d'ajustement est élevée sur les coûts moyens faibles qui sont sur-représentés par les lois théoriques. Très peu d'actes ont un coût moyen inférieur à sept euros pour la complémentaire santé. Le pic observé à 8 euros correspond principalement au coût moyen des actes de consultation chez un généraliste.

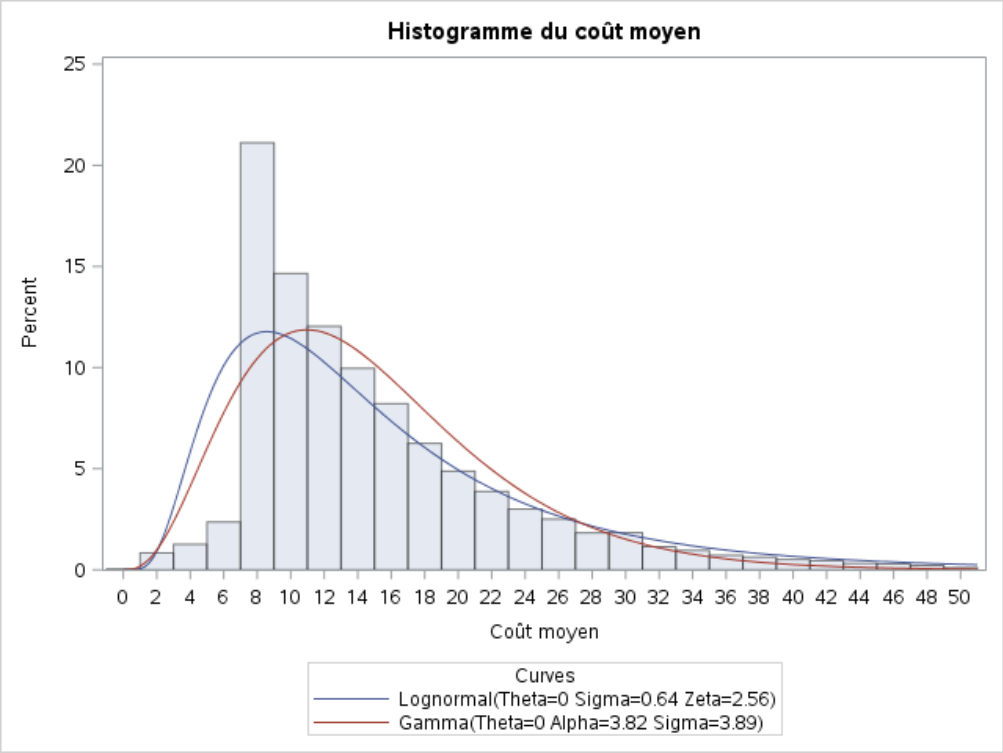


FIGURE 3.6 – Soins Courants - Comparaison de l'ajustement de l'histogramme du coût moyen avec les lois Gamma et Log-Normale

Une autre analyse graphique consiste à comparer les fonctions de répartitions des données observées avec celles des lois théoriques. La loi gamma présente un ajustement légèrement meilleur et a été retenue pour la modélisation.

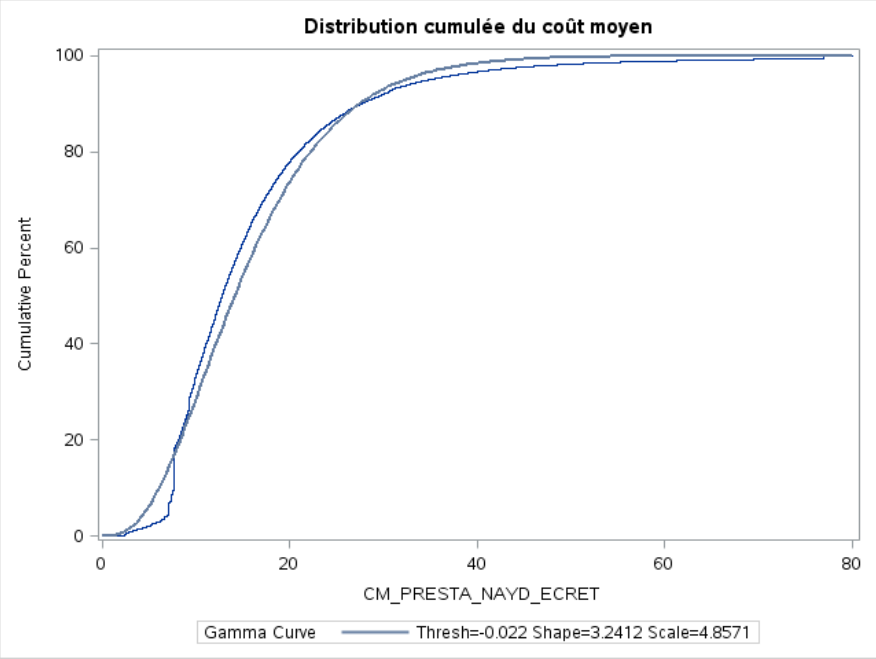


FIGURE 3.7 – Soins Courants - Ajustement de la fonction de répartition du coût moyen avec la loi Gamma

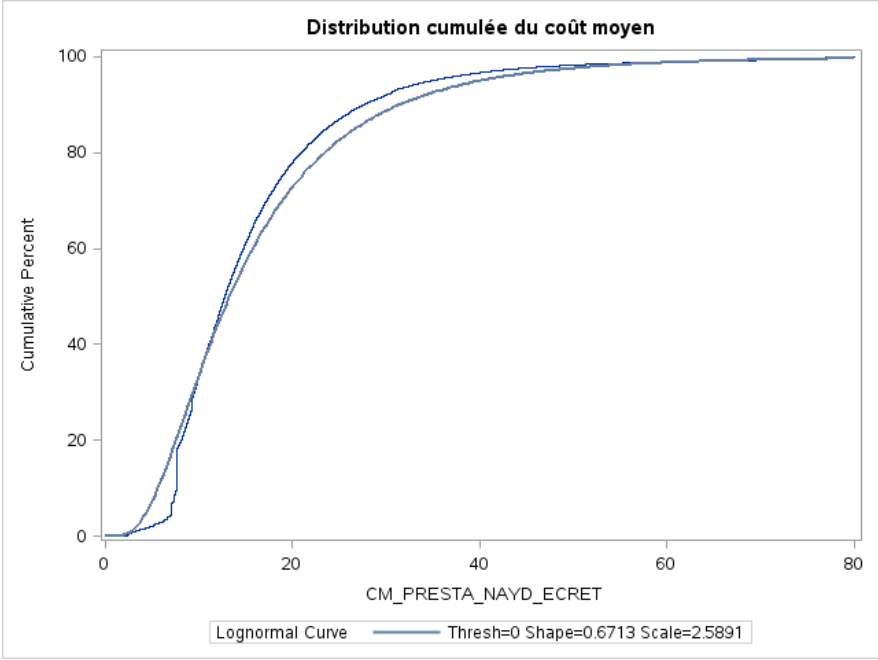


FIGURE 3.8 – Soins Courants - Ajustement de la fonction de répartition du coût moyen avec la loi log-Normale

3.2.3.2 Résultats des modèles de coût moyen

De même que pour la fréquence, deux modélisations ont été réalisées. La première modélisation du coût moyen par poste a été effectuée sans retraitement et la seconde a consisté à intégrer le même retraitement sur la variable « lien de parenté » que précédemment avec la fréquence ainsi qu'une interaction entre le niveau de garantie et la zone géographique.

Résultat du modèle sans retraitement et sans interaction

Ci-dessous (figure 3.12) les résultats de la modélisation sans interaction du coût moyen sur le poste soins courants.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.486539	0.008123	306.127	< 2e-16	***
niveau 1	0.028085	0.009083	3.092	0.00199	**
niveau 2	0.097577	0.008265	11.806	< 2e-16	***
niveau 3	0.233576	0.007732	30.209	< 2e-16	***
niveau 4	0.277812	0.008450	32.878	< 2e-16	***
niveau 5	0.387166	0.009761	39.663	< 2e-16	***
niveau 6	0.430128	0.010403	41.347	< 2e-16	***
Nombre de salariés : 1 salarié	0.048645	0.006024	8.075	6.83e-16	***
Age : 40-49 ans	0.027637	0.005238	5.276	1.32e-07	***
Age : 50 ans et plus	0.080278	0.005939	13.517	< 2e-16	***
Zone 2	0.065797	0.005556	11.843	< 2e-16	***
Zone 3	0.127780	0.004948	25.827	< 2e-16	***
Zone 4	0.321810	0.006039	53.290	< 2e-16	***
Catégorie : cadre	0.031226	0.005934	5.262	1.43e-07	***
Lien parenté : Conjoint	0.036519	0.005052	7.229	4.90e-13	***
Lien parenté : Enfant	-0.086479	0.004525	-19.111	< 2e-16	***
Top Option	0.049962	0.006881	7.261	3.87e-13	***
Régime Alsace-Moselle	-0.547235	0.011738	-46.622	< 2e-16	***

TABLE 3.12 – Soins Courants - coefficients estimés sur le modèle GLM

Dans le cadre de l'élaboration du zonier soins courants avec la méthode CAH, le critère du coude suggérait un découpage en trois classes mais 4 classes avait tout de même été retenues. Les résultats de la modélisation ci-dessus montrent que le découpage en quatre zones est pertinent.

Résultat du modèle avec retraitement sur le lien de parenté et interaction

Une interaction supplémentaire entre le niveau et le zonier a été introduite dans le cadre de la modélisation du coût moyen. Le facteur de la zone géographique intervient de manière différente selon le niveau de garantie. Le coût moyen est beaucoup moins sensible à la zone sur les niveaux d'entrée de gamme que sur les niveaux haute gamme. Ceci s'explique assez naturellement. Sur les niveaux d'entrée de gamme, le niveau de remboursement de la complémentaire santé atteint souvent le plafond de la garantie quelque soit la zone géographique. Ainsi dans les zones de plus forte sinistralité, la dépense totale et le reste à charge pour l'assuré seront plus importants sur les garanties basses mais cela ne sera pas nécessairement le cas pour le coût moyen qui sera limité par le plafond de la garantie.

Afin de prendre en compte cette spécificité, une variable croisée niveau-zonier a été introduite dans la modélisation. Par ailleurs, comme dans le cas de la modélisation de la fréquence avec retraitement, la modalité « Enfant » de la variable « lien de parenté » a été intégrée au sein de la variable « âge moyen des salariés ».

Le tableau ci-dessous (tableau 3.13) donne les résultats de la modélisation avec interaction du coût moyen sur le poste soins courants. La modalité de référence initiale sur le croisement niveau-zone est le niveau 0 et la zone 1. Cependant certaines modalités n'apparaissaient pas significatives, elles ont donc été regroupées avec la modalité de référence.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.587575	0.005537	467.291	< 2e-16	***
niveau 0 - zone 3	0.033416	0.010849	3.080	0.002071	**
niveau 0 - zone 4	0.046535	0.013419	3.468	0.000525	***
niveau 1 - zone 3	0.042233	0.011942	3.537	0.000406	***
niveau 1 - zone 4	0.086810	0.018367	4.726	2.29e-06	***
niveau 2 - zone 2	0.075754	0.010264	7.380	1.59e-13	***
niveau 2 - zone 3	0.133115	0.009466	14.062	< 2e-16	***
niveau 2 - zone 4	0.239769	0.016755	14.310	< 2e-16	***
niveau 3 - zone 1	0.116562	0.007943	14.675	< 2e-16	***
niveau 3 - zone 2	0.181299	0.008462	21.425	< 2e-16	***
niveau 3 - zone 3	0.277768	0.007668	36.226	< 2e-16	***
niveau 3 - zone 4	0.462006	0.010265	45.007	< 2e-16	***
niveau 4 - zone 1	0.161778	0.010479	15.438	< 2e-16	***
niveau 4 - zone 2	0.207916	0.012112	17.167	< 2e-16	***
niveau 4 - zone 3	0.303870	0.008772	34.641	< 2e-16	***
niveau 4 - zone 4	0.537453	0.011722	45.848	< 2e-16	***
niveau 5 - zone 1	0.215971	0.015718	13.741	< 2e-16	***
niveau 5 - zone 2	0.315466	0.017488	18.039	< 2e-16	***
niveau 5 - zone 3	0.363800	0.012203	29.813	< 2e-16	***
niveau 5 - zone 4	0.711142	0.012702	55.985	< 2e-16	***
niveau 6 - zone 1	0.256054	0.019687	13.006	< 2e-16	***
niveau 6 - zone 2	0.342505	0.024462	14.001	< 2e-16	***
niveau 6 - zone 3	0.415424	0.013236	31.386	< 2e-16	***
niveau 6 - zone 4	0.722691	0.013006	55.568	< 2e-16	***
Nombre de salariés : 1 salarié	0.046341	0.006007	7.714	1.23e-14	***
Age : Enfant	-0.073051	0.004726	-15.456	< 2e-16	***
Age : 45-49 ans	0.032242	0.006049	5.330	9.82e-08	***
Age : 50 ans et plus	0.081115	0.006448	12.580	< 2e-16	***
Catégorie : cadre	0.034231	0.005880	5.821	5.86e-09	***
Lien parenté : Conjoint	0.036791	0.005035	7.307	2.75e-13	***
Top Option	0.045336	0.006504	6.971	3.18e-12	***
Régime Alsace-Moselle	-0.519697	0.012151	-42.769	< 2e-16	***

TABLE 3.13 – Soins Courants - coefficients estimés sur le modèle GLM avec interaction

3.2.3.3 Etude des résidus

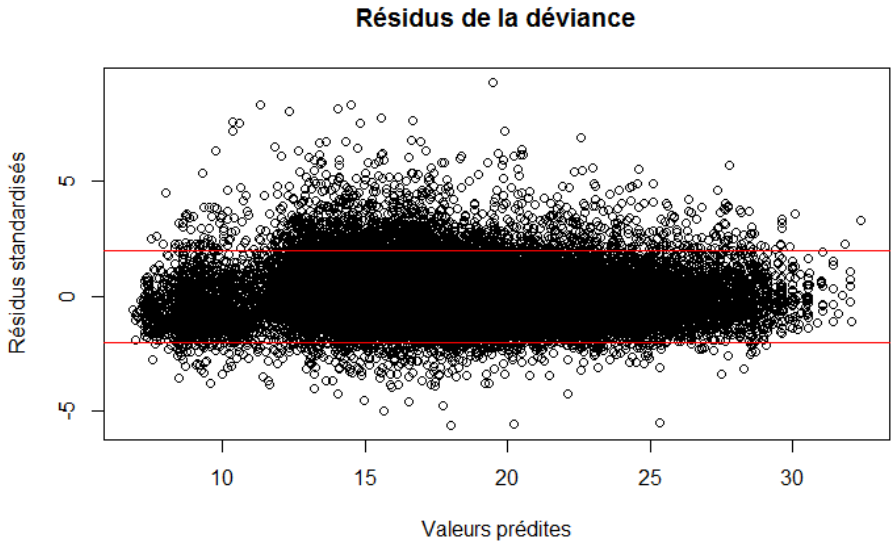


FIGURE 3.9 – Soins Courants - Résidus de la déviance en fonction des valeurs prédites sur le modèle sans interaction

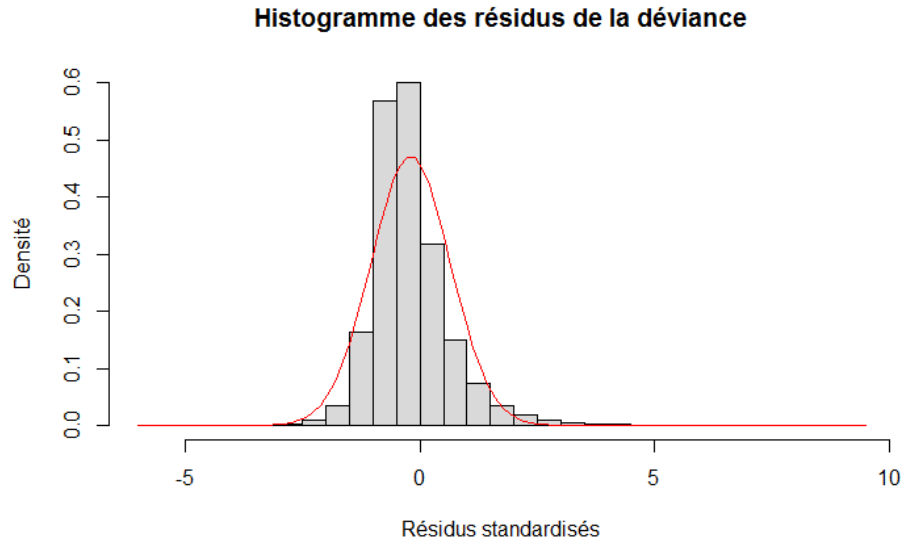


FIGURE 3.10 – Soins Courants - Histogramme de la répartition des résidus sur le modèle sans interaction

Sur le graphique 3.9, les résidus en fonction des valeurs prédites sont concentrés sur l'intervalle [-2,2] et restent centrés autour de 0, cela traduit une bonne adéquation du modèle. L'histogramme des résidus standardisés sur le graphique 3.10 illustre également la concentration des résidus autour de 0, ils suivent une distribution proche de la loi normale (courbe en rouge).

3.3 Modélisation de la prime pure avec la méthode CART

Introduction

La méthode CART, acronyme de Classification And Regression Tree (Arbre de classification et de régression) a été introduite par Breiman en 1984. Il s'agit d'une méthode non paramétrique permettant de construire des classes de risque et d'associer à chaque classe la valeur moyenne de la fréquence et du coût des sinistres. Elle constitue une alternative aux modèles linéaires généralisés dans le cadre de travaux de tarification et présente l'avantage d'avoir des résultats relativement faciles à interpréter.

3.3.1 Présentation de l'algorithme CART

Tout d'abord, CART est un algorithme consistant à élaborer un arbre de décision. Cette méthode appartient à la famille des méthodes de classification supervisées, c'est à dire qu'elle fait intervenir des variables d'entrées, les explicatives X = $(X_1, ...X_p)$ et une variable de sortie, la variable à expliquer Y. L'apprentissage supervisé consiste à construire une fonction de prédiction de la variable à modéliser Y à partir des variables explicatives $(X_1, ...X_p)$. En fonction de la nature de la variable à expliquer, qualitative ou quantitative, l'arbre de décision sera qualifié d'arbre de classification ou de régression. Les arbres de classification ont pour objet de prédire l'appartenance d'un individu à une classe et les arbres de régression vont chercher à prédire les valeurs de la variable réponse en fonction des variables explicatives. Dans le cadre d'une problématique de modélisation de prime pure, le type d'arbre utilisé avec la méthode CART est l'arbre de régression.

La particularité de la méthode CART est de procéder au partitionnement de manière binaire et récursive. L'ensemble des observations sont regroupées à la racine de l'arbre et constituent le nœud initial de l'arbre. Ensuite à chaque étape de découpage, l'arbre se divise en deux branches. Chaque nœud fait ainsi l'objet d'une division en deux nœuds fils composés chacun d'observations plus homogènes au sens d'un critère donné. L'algorithme segmente chaque nœud à partir des variables explicatives afin de différencier le mieux possible les deux groupes au sens de la variable à expliquer. Le processus de partitionnement s'achève lorsque la condition d'arrêt de l'algorithme est vérifiée. La condition d'arrêt peut être un nombre minimum d'observations au sein du nœud à découper, un nombre d'observations minimum au sein d'un nœud final. Les nœuds terminaux ainsi constitués correspondent aux feuilles de l'arbre.

De manière générale, l'arbre de décision se caractérise par les éléments suivants :

- La racine : il s'agit du nœud initial, il contient l'ensemble des observations (point de départ)

- Des branches : elles contiennent les règles de division qui permettent de segmenter les nœuds

- Des nœuds : ils forment des sous-ensembles d'observations

- Des feuilles : elles correspondent aux nœuds terminaux de l'arbre.

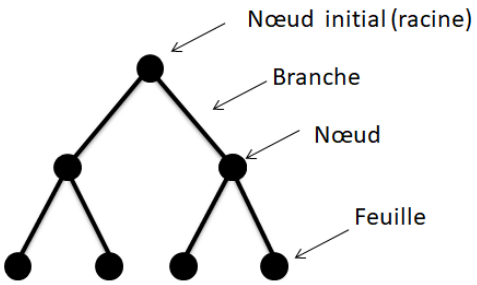


FIGURE 3.11 – Schéma d'un arbre CART

L'arbre maximal n'est pas nécessairement optimal en terme de prévision. En effet, malgré la présence de critères d'arrêt du processus de division, l'arbre maximal peut présenter des problèmes de sur-apprentissage. Afin de déterminer l'arbre optimal, il est nécessaire de procéder à une phase d'élagage de l'arbre maximal.

3.3.1.1 Principe d'élaboration de l'arbre binaire maximal

La construction de l'arbre binaire maximal appelé également arbre saturé consiste à déterminer une séquence de nœuds. Chaque nœud est élaboré à partir du choix d'une variable explicative à laquelle est associée une règle de division générant ainsi une partition en deux sous-ensembles. La division est définie par une valeur seuil dans le cas d'une variable quantitative et un groupe de modalités dans le cas d'une variable qualitative.

A chaque étape de division, il existe un ensemble de divisions admissibles. Le rôle de l'algorithme consistera à déterminer parmi cet ensemble quelle est la meilleure division au sens d'un critère de mesure d'homogénéité préalablement défini. Le processus de division de chaque nœud s'achève lorsque le critère d'arrêt est atteint. Ce critère d'arrêt peut être par exemple un nombre d'observations minimum au sein d'un nœud. En l'absence de critère d'arrêt, le processus se poursuit jusqu'à ce que chaque nœud terminal ne contienne plus qu'une observation.

La figure 3.12 ci-dessous présente un exemple d'arbre saturé. Cet arbre très fin s'avère illisible et présente un risque fort de sur apprentissage d'où la nécessité de procéder à une seconde étape nommée phase d'élagage présentée à la section 3.3.1.3.

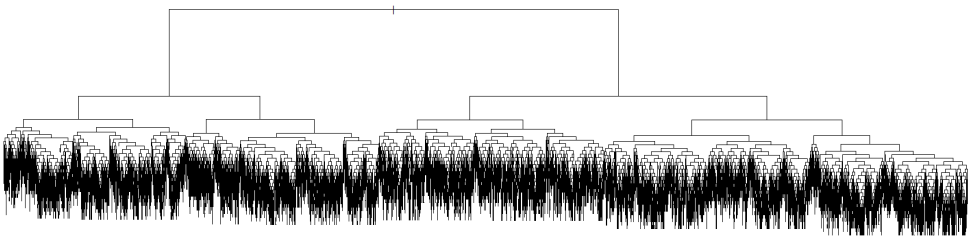


FIGURE 3.12 – Schéma d'un arbre saturé

3.3.1.2 Critère de division

A chaque étape de division, il existe un grand nombre de critères de division des nœuds possibles. Ce nombre est tout d'abord lié au nombre de variables explicatives mais également à leur nature. Si la variable explicative est quantitative et comporte n valeurs distinctes alors il existe n-1 possibilités de découpage. Dans le cas d'une variable explicative qualitative comportant m modalités, il existe 2^{m-1} divisions possibles.

Il existe donc une multiplicité de critères de division admissibles à chaque étape, d'où la nécessité d'introduire une règle permettant de sélectionner le critère le plus optimal. Le choix du critère de division parmi les divisions admissibles s'appuie sur la notion d'hétérogénéité. L'objectif est de sélectionner le critère qui permet de segmenter un nœud en deux groupes les plus homogènes possibles au sens de la variable à expliquer.

Traduction théorique

- Soit N un nœud parent et une règle de division générant deux nœuds fils gauche et droit notés N_g et N_d.

- Soit D un indicateur de mesure de l'hétérogénéité avec D_N la mesure de l'hétérogénéité du nœud N, D_{NG} l'hétérogénéité du nœud fils gauche N_g et D_{ND} l'hétérogénéité du nœud fils gauche N_d

alors l'algorithme sélectionnera la règle de division qui permet de minimiser la somme des hétérogénéités des deux nœuds fils $D_{ND} + D_{NG}$.

Par équivalence, l'algorithme déterminera à chaque étape le critère de division qui maximise l'écart entre l'hétérogénéité du nœud père et la somme des hétérogénéités des deux nœuds fils :

$$\underset{\{division\ de\ X^j\ ;\ j=1,p\}}{Max} \quad D_N \ - \ (D_{ND} + D_{NG})$$

Dans le cas d'une régression avec une variable à expliquer Y quantitative, l'hétérogénéité du nœud k est définie par la variance :

$$D_k \ = \ \frac{1}{n_k} \sum_{i\epsilon \ k} (y_i - \overline{y}_k)^2$$

où n_k est l'effectif du nœud k.

Maximiser l'écart entre l'hétérogénéité du nœud père K et la somme des hétérogénéités des deux nœuds fils L et M revient à minimiser la variance intra-classe :

$$\frac{1}{n_k} \sum_{i\epsilon \ l} (y_i - \overline{y}_l)^2 + \frac{1}{n_k} \sum_{i\epsilon \ m} (y_i - \overline{y}_m)^2 = \frac{n_l}{n_k} D_l + \frac{n_m}{n_k} D_m$$

3.3.1.3 Phase d'élagage

L'arbre maximal ne constitue généralement pas l'arbre optimal. De par son extrême complexité, l'arbre maximal a tendance à présenter un surajustement aux données de l'échantillon d'apprentissage. Ce surapprentissage le rendra souvent peu robuste et faiblement performant en terme de prévision.

Afin de déterminer l'arbre optimal, il est nécessaire de procéder à une phase d'élagage de l'arbre maximal. Cette phase d'élagage consiste à sélectionner un sous-arbre élagué de l'arbre maximal. Ce sous-arbre de même racine que l'arbre maximal pourra être identifié en estimant la performance des différents sous-arbres sur un échantillon de test. Il s'agira donc de construire l'ensemble des modèles d'arbres possibles pour chaque nombre de feuilles k et de tester chacun d'eux sur un échantillon test. Le nombre de feuille k varie de 1 à $|T_{max}|$ avec $|T_{max}|$ le nombre de feuilles de l'arbre maximal.

Cependant le nombre de candidats potentiels pour être l'arbre optimal s'avère rapidement très important, le nombre de combinaisons évolue de manière exponentielle avec la complexité de l'arbre maximal. La détermination de l'arbre optimal devient donc algorithmiquement insolvable. Afin de résoudre ce problème, Breiman (1984) a élaboré une nouvelle approche consistant à construire une suite emboitée de sous-arbres de l'arbre maximal et à choisir comme arbre optimal celui qui minimise l'erreur sur un échantillon test. La construction de la séquence d'arbres emboités repose sur l'introduction d'un critère de pénalisation.

L'objectif est donc de construire une suite d'arbres emboités $T_1, ..., T_K$ où chaque T_k minimise l'erreur d'ajustement pénalisée. Pour chaque sous arbre T de T_{max} composé de $|T|$ nœuds terminaux k, avec k=1,...,$|T|$, la qualité d'ajustement correspond à l'expression suivante :

$$D(T) = \frac{1}{n} \sum_{k=1}^{|T|} \sum_{i\epsilon\ k} (y_i - \overline{y}_k)^2$$

Le critère de pénalisation de la complexité de l'arbre est introduit sous la forme d'une fonction linéaire du nombre de feuilles. La mesure de la qualité d'ajustement pénalisée prend la forme suivante :

$$C(T) = D(T) + \gamma \times |T|$$

Le processus de construction de la séquence d'arbres emboités est initialisé avec γ=0. Pour cette valeur du paramètre, l'arbre qui minimise l'erreur d'ajustement pénalisé C(T) correspond à l'arbre maximal T_{max}, ce qui ce traduit par $T_K = T_{max}$. Ensuite en faisant croître le paramètre γ, il existe une valeur seuil de γ à partir de laquelle le gain d'ajustement de l'une des K-1 divisions de l'arbre T_K ne compense pas la pénalisation γ. Ainsi les deux feuilles issues de cette division sont regroupées et permettent de former l'arbre T_{K-1}.

Le processus est itéré jusqu'à l'obtention de la séquence complète d'arbres emboités :

$$T_{max} = T_K \supset T_{K-1} \supset ... \supset T_1$$

A partir de cette séquence d'arbres emboités, il est possible de construire un graphique de la décroissance de la qualité d'ajustement D(1), ...,D(K) en fonction du nombre de feuilles de l'arbre ou du coefficient de pénalisation γ.

3.3.1.4 Identification de l'arbre optimal

L'identification de l'arbre optimal constitue l'étape ultime de l'algorithme CART. Elle consiste à choisir parmi la séquence d'arbres emboités l'arbre qui minimise l'erreur sur un échantillon de validation. Cependant, si l'échantillon de données n'est pas de taille suffisante pour permettre la constitution d'un échantillon de validation, il est possible d'adopter une approche en validation croisée.

L'approche par une validation croisée s'appuie sur la segmentation de l'échantillon en N sous échantillons. Chaque sous échantillon possède sa propre séquence d'arbres emboités. L'objectif de cette approche est de déterminer l'erreur moyenne pour chaque valeur du paramètre γ issu de la séquence d'arbres emboités sur l'échantillon global et d'en déduire le paramètre γ optimal, c'est à dire celui qui minimise l'erreur moyenne de validation sur les N sous échantillons. Plus précisément, à chaque valeur du paramètre γ issu de la séquence d'arbres emboités de l'échantillon global correspond un arbre pour chaque sous échantillon. A partir de ces N arbres dont le nombre de feuilles peut varier, une erreur moyenne est déterminée. Ainsi pour les K valeurs du paramètre γ une erreur moyenne est associée.

La sélection de l'arbre optimal pourra être réalisée en sélectionnant l'arbre correspondant à la valeur γ qui minimise l'erreur moyenne de validation ou en appliquant la règle de l'écart type consistant à retenir l'arbre le moins complexe pour lequel l'erreur moyenne est inférieure à un écart type de l'erreur minimale.

3.3.2 Étapes préliminaires à la modélisation

3.3.2.1 Paramétrage de l'algorithme CART

L'algorithme CART fait intervenir un certain nombre de paramètres qu'il est nécessaire de définir en amont de la modélisation. Le choix de la valeur de ces paramètres est essentiel car ils peuvent influer sur les résultats de l'algorithme.

Voici le paramétrage qui a été retenu dans le cadre de la modélisation de la prime pure :

- le nombre d'observations minimum au sein d'un nœud terminal a été fixé à 1000. Ce nombre doit être suffisamment grand afin que l'estimation de la prime pure au sein des nœuds terminaux soit fiable.

- la profondeur maximale de l'arbre a été fixée à 11. La limite de la profondeur de l'arbre a pour objectif d'éviter le surapprentissage. Après une analyse des différents arbres issus des modélisations CART, les découpages au delà de ce seuil n'apparaissaient pas toujours pertinents.

- le nombre d'échantillons de validation croisée a été fixé à 10, ce qui correspond à la valeur par défaut. Ce nombre ne doit pas être top élevé afin que le calcul de l'erreur par validation croisée porte sur un échantillon suffisamment grand. En fixant la valeur à 10, l'erreur de validation sera déterminée à chaque étape sur un échantillon composé de 10% de la base d'apprentissage.

- il est possible de fixer une valeur minimale du paramètre de complexité, cela peut permettre d'optimiser les temps de traitement. Dans le cadre de cette étude, la valeur 0 a été retenue afin de conserver l'arbre maximal.

3.3.2.2 Prise en compte de l'exposition

Une des particularités à prendre en compte lors de la modélisation de la prime pure en assurance est la notion d'exposition. Précédemment dans le cadre de la modélisation de la fréquence via l'approche GLM, celle-ci avait été prise en compte sous la forme d'une variable offset.

Afin de prendre cette spécificité liée à la tarification, dans le cadre de l'algorithme CART, Christmann propose en 2004 une adaptation de l'algorithme consistant à diviser les sinistres par leur exposition. Cependant cette méthode conduit à une erreur d'estimation. La charge de sinistres estimée pour un groupe d'assurés ne correspond pas à la somme des sinistres de ce groupe.

$$\begin{aligned}\sum_{i=1}^{n} t_i \times \overline{Y} &= \sum_{i=1}^{n} t_i \times \left(\frac{1}{n}\sum_{i=1}^{n} Y_i^*\right) \\ &= \sum_{i=1}^{n} t_i \times \left(\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{t_i}\right) \\ &\neq \sum_{i=1}^{n} Y_i\end{aligned}$$

avec n le nombre d'individus du groupe, $\overline{Y}$ la prime pure payée par chaque assuré, Y_i le montant des sinistres, t_i la durée d'exposition et Y_i^* le montant des sinistres divisé par l'exposition.

Afin d'éviter ce biais, il est nécessaire de prendre en compte le poids de chaque assuré, poids w_i équivalent à son exposition t_i lors de l'estimation de la prime pure moyenne du groupe, ce qui permet d'obtenir l'égalité entre la charge de sinistres estimée et la somme des sinistres.

$$\begin{aligned}\sum_{i=1}^{n} t_i \times \overline{Y} &= \sum_{i=1}^{n} t_i \times \left(\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} Y_i^* \times w_i\right) \\ &= \sum_{i=1}^{n} t_i \times \left(\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} \frac{Y_i}{t_i} \times w_i\right) \\ &= \sum_{i=1}^{n} t_i \times \left(\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} Y_i\right) \\ &= \sum_{i=1}^{n} Y_i\end{aligned}$$

Dans le cas de la modélisation avec l'approche fréquence-coût moyen, la prise en compte de l'exposition sera donc réalisée en divisant la fréquence observée par l'exposition et en retenant l'exposition comme facteur de poids pour chaque observation. Dans le cas de la modélisation avec l'approche coût total, ce procédé sera appliqué en considérant le montant total des sinistres de chaque assuré.

3.3.2.3 Seuil d'élagage

Le seuil d'élagage est une étape clé de la modélisation CART. A la section 3.3.1.4, deux méthodes ont été évoquées, une première consistant à sélectionner l'arbre correspondant à la valeur γ qui minimise l'erreur moyenne de validation et une seconde nommée règle de l'écart type consistant à retenir l'arbre le moins complexe pour lequel l'erreur moyenne est inférieure à un écart type de l'erreur minimale.

La première méthode aura tendance à sélectionner des arbres relativement complexes. A partir d'un certain seuil de complexité, l'erreur continue à diminuer mais son amélioration devient très faible et présente des oscillations (cf exemple tableau 3.14). La sélection d'un arbre moins complexe que celui qui minimise l'erreur pourrait réduire légèrement la performance mais améliorer l'interprétabilité. La seconde méthode est beaucoup plus robuste et conduit à sélectionner des arbres beaucoup plus simples. Toutefois, cette méthode n'apparaît pas très adaptée en tarification, elle peut conduire à une segmentation trop faible et à un manque de finesse du tarif.

Il n'existe donc pas de méthode parfaite d'élagage de l'arbre, c'est pourquoi deux seuils d'élagage seront testés. Le premier seuil d'élagage retenu sera celui qui minimise l'erreur moyenne de validation. Afin de limiter la complexité de l'arbre tout en conservant une certaine finesse du tarif, un second seuil nommé « seuil intermédiaire » figurant comme un compromis entre les deux méthodes sera également testé. L'arbre sera élagué au seuil de complexité pour lequel l'erreur moyenne est inférieure à 2,5% de l'écart type de l'erreur minimale afin de réduire l'effet de surapprentissage lié au choix de l'arbre minimisant l'erreur de validation.

Le tableau 3.14 de la page suivante issu de la modélisation de la fréquence du poste soins courants montre l'évolution de l'erreur de validation en fonction de la complexité de l'arbre. Les erreurs en rouge correspondent à des découpages pour lesquels l'erreur de validation ne s'améliore pas par rapport à l'arbre précédent comportant un nœud de moins. Sur ce tableau, l'arbre minimisant l'erreur de validation est composé de 55 feuilles. Cependant, le gain sur l'erreur de validation apporté par les dix précédentes division est relativement faible, de l'ordre de 0,3%. De plus, parmi les douze divisions précédentes, trois n'ont pas généré d'amélioration de l'erreur de validation. Ce constat tend à montrer que l'arbre minimisant l'erreur risque de comporter une part de surapprentissage.

L'arbre correspondant à l'erreur minimale plus 2,5% de l'écart type est beaucoup plus simple avec 35 feuilles et la dégradation de l'erreur de l'ordre de 0,3% reste mesurée. Ce compromis d'élagage permet donc de réduire nettement la complexité tout en limitant la dégradation de l'erreur de validation.

size of tree	CP	nsplit	rel error	xerror	xstd	
1	9,82E-03	0	1	1,00003	0,01318	
2	5,18E-03	1	0,99018	0,99022	0,013025	
3	2,62E-03	2	0,985	0,98505	0,012965	
4	2,47E-03	3	0,98238	0,98261	0,012943	
5	2,35E-03	4	0,97991	0,98047	0,012925	
7	1,80E-03	6	0,97521	0,97548	0,012896	Erreur minimal + un écart type
8	1,43E-03	7	0,97341	0,97399	0,012858	
9	1,37E-03	8	0,97197	0,97288	0,012854	
10	8,03E-04	9	0,97061	0,97094	0,012837	
11	5,69E-04	10	0,9698	0,97057	0,012832	
12	5,46E-04	11	0,96923	0,9702	0,012821	
13	4,80E-04	12	0,96869	0,96994	0,012816	
14	3,61E-04	13	0,96821	0,96878	0,012805	
15	3,55E-04	14	0,96785	0,96853	0,012806	
16	3,20E-04	15	0,96749	0,96836	0,012805	
17	3,19E-04	16	0,96717	0,96827	0,012804	
19	2,99E-04	18	0,96653	0,96808	0,012803	
20	2,56E-04	19	0,96624	0,96778	0,012796	
21	1,94E-04	20	0,96598	0,96734	0,012795	
22	1,78E-04	21	0,96579	0,96731	0,012793	
23	1,69E-04	22	0,96561	0,9673	0,012792	
25	1,63E-04	24	0,96527	0,96726	0,012793	
26	1,50E-04	25	0,96511	0,96729	0,012794	
27	1,49E-04	26	0,96496	0,96719	0,012788	
28	1,29E-04	27	0,96481	0,96706	0,012785	
29	1,29E-04	28	0,96468	0,96685	0,012784	
30	1,22E-04	29	0,96455	0,96677	0,012784	
31	1,16E-04	30	0,96443	0,96668	0,012782	
32	9,50E-05	31	0,96431	0,96657	0,012779	
33	8,85E-05	32	0,96422	0,96648	0,012775	
34	8,73E-05	33	0,96413	0,96642	0,012774	
35	7,17E-05	34	0,96404	0,96631	0,012776	Erreur minimal + 0,025 écart type
36	7,07E-05	35	0,96397	0,96628	0,012776	
37	6,76E-05	36	0,9639	0,96632	0,012776	
38	6,43E-05	37	0,96383	0,96628	0,012775	
39	6,40E-05	38	0,96377	0,96628	0,012775	
40	6,24E-05	39	0,9637	0,96628	0,012775	
41	5,90E-05	40	0,96364	0,96633	0,012775	
42	4,86E-05	41	0,96358	0,96632	0,012775	
43	4,39E-05	42	0,96353	0,96616	0,012773	
48	3,81E-05	47	0,9633	0,96614	0,012773	
51	2,84E-05	50	0,96319	0,96605	0,012774	
52	2,80E-05	51	0,96316	0,96602	0,012774	
53	1,84E-05	52	0,96313	0,96603	0,01277	
54	1,48E-05	53	0,96311	0,96602	0,01277	
55	4,63E-06	54	0,9631	0,966	0,01277	Erreur minimale
56	0,00E+00	55	0,96309	0,966	0,01277	

TABLE 3.14 – Fréquence soins courants - Tableau de l'évolution de l'erreur de validation en fonction de la complexité de l'arbre avec un seuil à 1000 observations

3.3.2.4 Retraitement de variables

Certaines variables correspondantes à un découpage en classes doivent faire l'objet d'un retraitement afin que les critères de division des nœuds soient cohérents. L'âge moyen des salariés illustre ce besoin. La consommation médicale est monotone croissante avec l'âge moyen des salariés. Le fait de conserver la variable de l'âge moyen des salariés en classes peut générer des discontinuités au sein d'un critère de découpage. Par exemple sans retraitement, il est possible d'observer un critère de division de ce type : les segments des [40-44 ans] et [50 ans et +] affectés au premier nœud fils et les autres modalités au second nœud fils. Il y a donc une incohérence sur le segment des [45-49 ans] qui est affecté au second nœud.

Afin de palier à ce problème, la variable classe d'âge en classe est transformée en variable numérique en affectant la valeur centrale du segment.

Classe d'âge moyen des salariés	Valeur numérique attribuée
< 30 ans	27,5
30-34 ans	32,5
35-39 ans	37,5
40-44 ans	42,5
45-49 ans	47,5
50 ans et plus	52,5

TABLE 3.15 – Retraitement de la classe d'âge moyen des salariés

Ainsi, l'incohérence précédente est rendue impossible du fait de la nature de la variable qui est numérique. De plus, l'attribution des valeurs centrales permet d'afficher la bonne règle sur l'arbre. Par exemple, pour un découpage des moins de 30 ans, l'algorithme affichera la valeur centrale entre 27,5 et 32.5, soit 30 ce qui correspond bien à la valeur attendue.

Ce retraitement sera également appliqué sur la variable du nombre de salariés moyen dans l'entreprise avec les valeurs ci-dessous.

Classe du nombre de salariés	Valeur numérique attribuée
1 salarié	1
2 salariés	2
3 salariés	3
4-10 salariés	4
11-20 salariés	18
21-30 salariés	25
31 salariés et plus	37

TABLE 3.16 – Retraitement de la classe du nombre de salariés dans l'entreprise

Les variables niveau de garantie et zonier seront considérées sous leur forme numérique.

3.3.3 Application de la modélisation CART avec la même approche que pour le GLM

Afin de pouvoir comparer les résultats de la méthode CART avec la méthode GLM, une première modélisation a été réalisée avec une approche fréquence-coût moyen en utilisant les mêmes variables que dans le cadre de la modélisation GLM et en considérant la règle d'élagage du « seuil intermédiaire » défini à la section 3.3.2.3. Les résultats seront détaillés uniquement sur le poste soins courants comme dans le cas de la modélisation GLM.

D'autres modélisations avec l'approche prime pure et la modification des valeurs des paramètres de l'algorithme seront également testées et présentées à la section suivante.

3.3.3.1 Modélisation de la fréquence

Avant d'exposer les résultats de la modélisation de la fréquence, une présentation de l'étape d'élagage s'avère importante.

Elagage de l'arbre

La phase d'élagage détaillée à la section précédente est une des étapes essentielles de la modélisation CART. Le graphique de l'erreur de validation croisée en fonction de la taille de l'arbre permet d'illustrer le résultat de cette étape. La figure 3.13 ci-dessous représente l'évolution de l'erreur par validation croisée dans le cadre de la modélisation de la fréquence sur le poste soins courants.

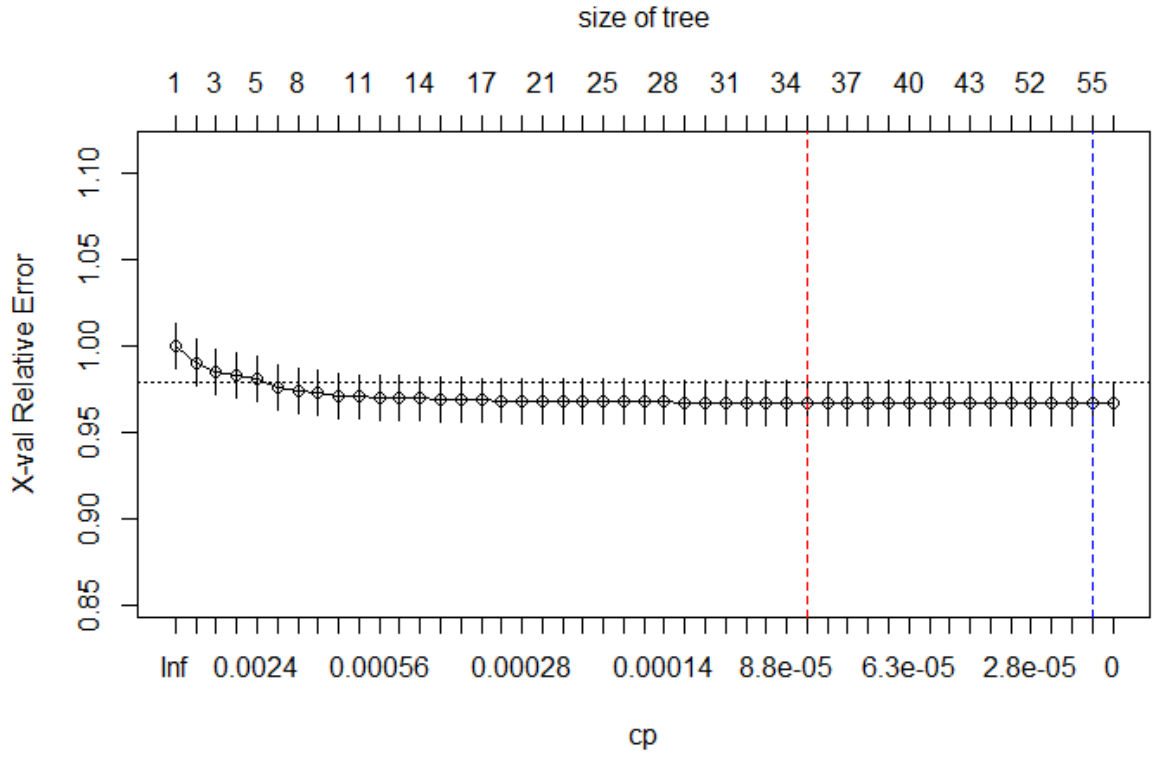


FIGURE 3.13 – Soins courants - Courbe d'évolution de l'erreur sur la fréquence en fonction de la complexité de l'arbre

Les deux traits verticaux en rouge et en bleu sur le graphique correspondent respectivement aux seuils d'élagage avec la règle du « seuil intermédiaire » et la règle du « seuil optimal ». Sur ce graphique, l'erreur décroit assez rapidement sur les premiers nœuds de l'arbre et beaucoup plus lentement ensuite. Enfin, à partir d'un certain seuil, l'erreur n'évolue plus et tend à stagner.

L'application de la règle d'élagage du « seuil optimal » qui minimise l'erreur de validation (seuil en bleu sur la figure 3.13) conduit à un élagage au 55^{eme} nœud de l'arbre. Dans ce cas l'arbre optimal est très proche de l'arbre complet, ce qui est dû vraisemblablement à une règle d'arrêt trop restrictive. Un test sera réalisé dans la partie suivante en abaissant le nombre d'observations minimales sur les nœuds terminaux de 1000 à 500. En utilisant la règle d'élagage du « seuil intermédiaire » (seuil en rouge sur la figure 3.13), l'arbre obtenu s'avère nettement moins complexe avec 35 feuilles.

Résultat de la modélisation

L'arbre résultant de la modélisation CART permet de visualiser les critères ayant permis d'élaborer les différents segments obtenus. La figure 3.14 ci-après représente l'arbre issu de la modélisation de la fréquence soins courants, celui-ci est composé de 35 segments (nœuds terminaux). Les résultats sur les autres postes sont disponibles en annexe D.

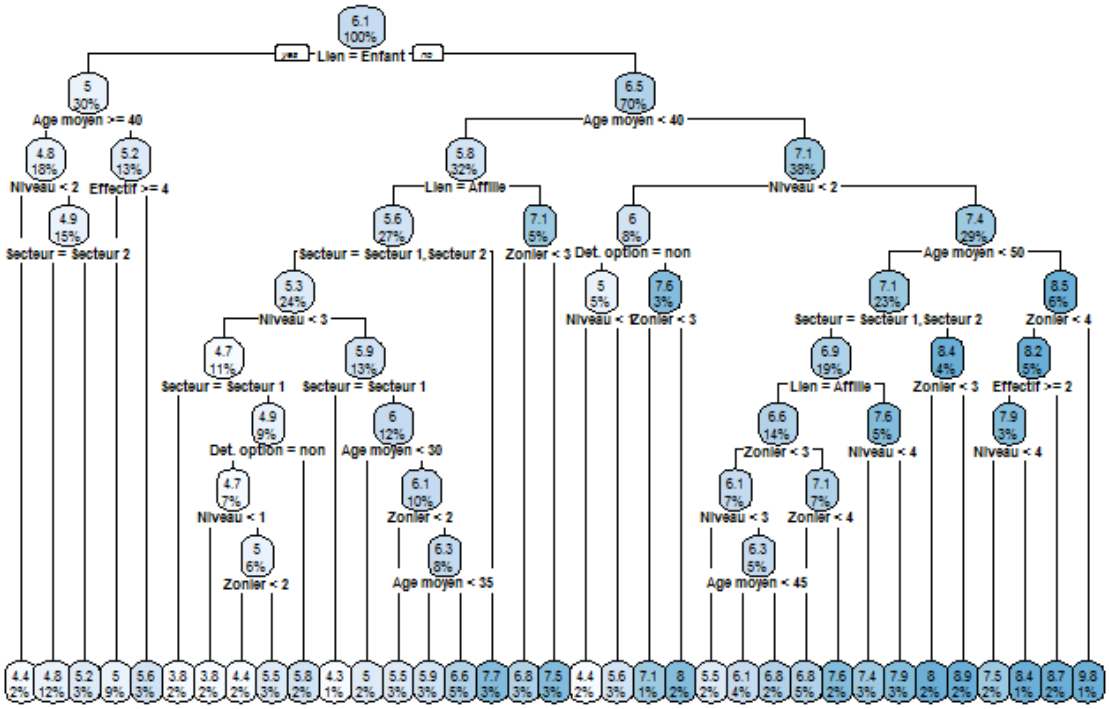


FIGURE 3.14 – Soins courants - Arbre CART sur la fréquence

Le lien de parenté et l'âge moyen des salariés sont les deux premières variables qui interviennent dans le processus de division. Ce constat n'est pas très surprenant, ces variables ayant été identifiées comme étant très explicatives de la consommation médicale dans le cadre de l'analyse descriptive (section 2.2.1).

Sur l'arbre, les segments en bleu foncé correspondent à des segments où la fréquence de consommation en soins courants est la plus importante. Ils se situent sur la partie droite de l'arbre. La partie gauche de l'arbre concentre les segments avec une faible fréquence de consommation, il s'agit majoritairement des enfants de salariés.

Le segment le plus à droite correspond au segment où la fréquence est la plus élevée avec 9,8 actes par an et par assuré. Il s'agit de salariés ou conjoints rattachés à une entreprise située en zone 4 (Ile-de-France / PACA) dont la moyenne d'âge des salariés est supérieure à 50 ans avec un niveau de garantie moyen haute-gamme (niveaux 2 et +).

Les segments avec les plus petits consommants se situent sur la partie gauche du sous bloc adulte. Leur fréquence de consommation est de 3,8 actes par an et par assuré. Il s'agit de salariés ($\neq$ ayant droit) rattachés à une entreprise du secteur agricole ou de la construction (secteur 1) dont la moyenne d'âge des salariés est inférieure à 40 ans avec niveau de garantie d'entrée et de moyenne gamme (niveau <3).

De manière générale, l'analyse des arbres issus de la modélisation de la fréquence montre que les résultats sont cohérents avec les constats réalisés dans le cadre de l'analyse descriptive.

Le graphique 3.15 ci-dessous spécifiant l'importance relative des variables confirme le constat précédent. Les deux variables les plus importantes sont le lien de parenté et l'âge moyen des salariés, viennent ensuite le niveau de garantie et le secteur d'activité. L'importance de la variable régime est faible, cette variable impactant plutôt le coût moyen.

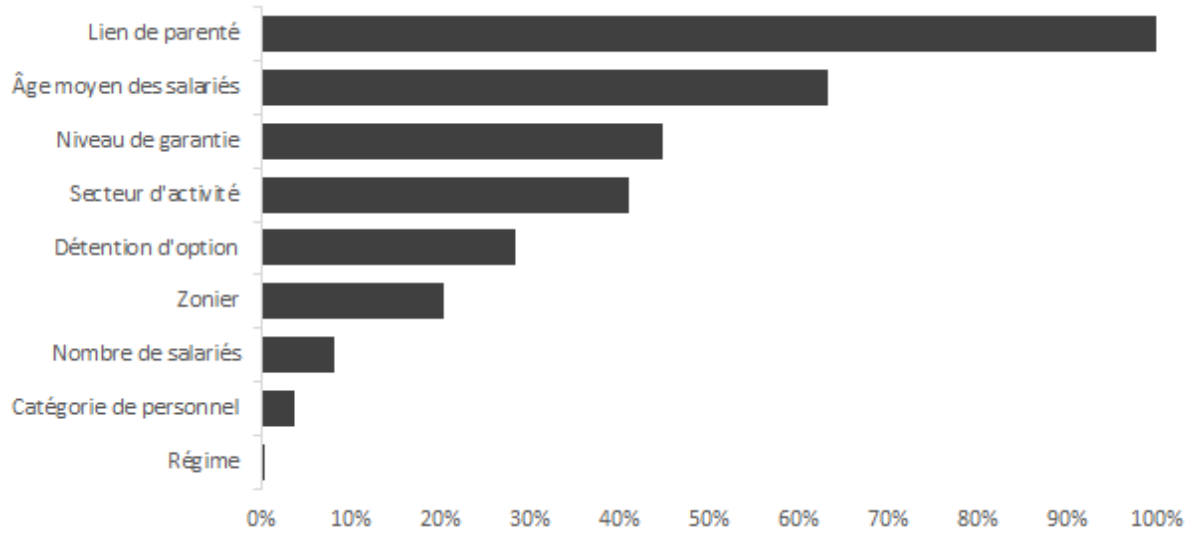


FIGURE 3.15 – Soins courants - Importance relative des variables du modèle fréquence

3.3.3.2 Modélisation du coût moyen

Les résultats de la modélisation du coût moyen se présentent de la même manière que pour la modélisation de la fréquence.

Élagage de l'arbre

Le processus d'élagage de l'arbre maximal pour le coût moyen est le même que pour la fréquence. La figure 3.16 ci-dessous représente l'évolution de l'erreur par validation croisée dans le cadre de la modélisation du coût moyen sur le poste soins courants. Les barres sur le graphique représentent l'écart type de l'erreur de validation croisée qui apparaît relativement élevée.

L'application de la règle d'élagage « seuil intermédiaire » (seuil en rouge sur la figure 3.16) dans le cadre de la modélisation du coût moyen des soins courants conduit à un arbre moins complexe que pour la fréquence, celui-ci est composé de 20 feuilles alors que celui de la fréquence contenait 35 feuilles.

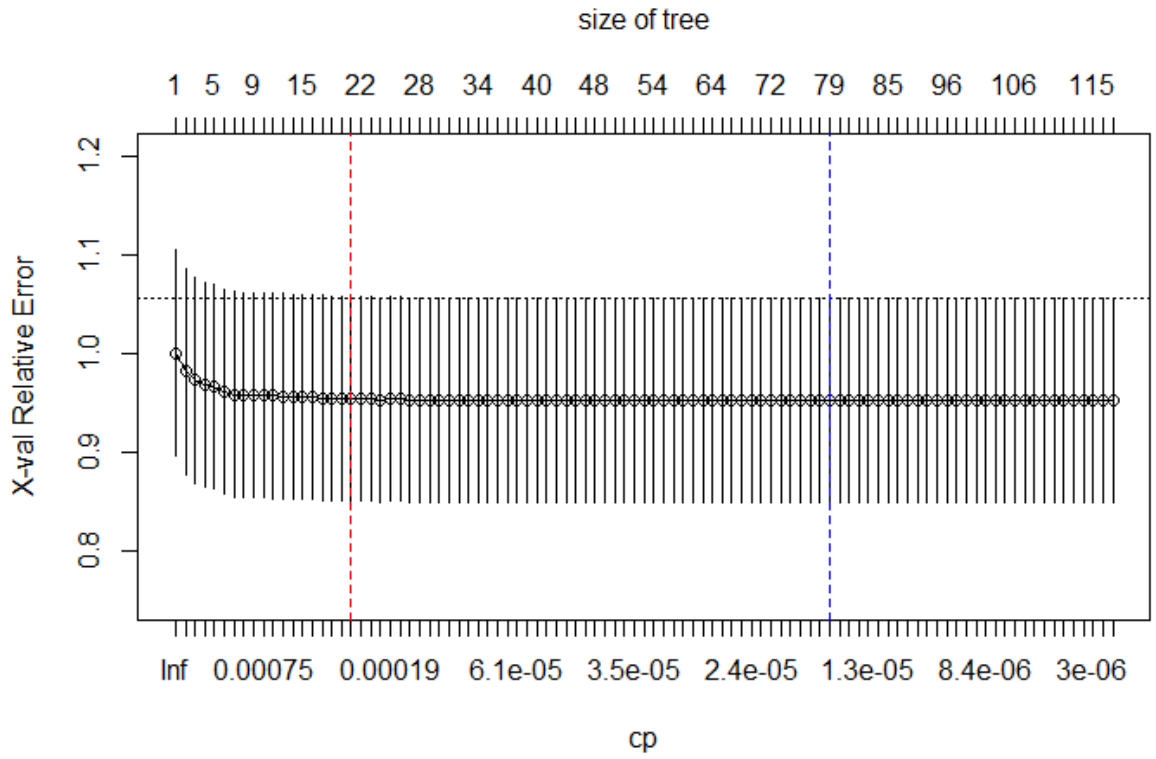


FIGURE 3.16 – Soins courants - Courbe d'évolution de l'erreur sur le coût moyen en fonction de la complexité de l'arbre

Résultat de la modélisation

De même que pour la fréquence, 6 arbres ont été élaborés dans le cadre de la modélisation du coût moyen, chacun correspondant à un poste de garantie.

La figure 3.17 présente le résultat de la modélisation du coût moyen sur le poste soins courants, les résultats sur les autres postes étant disponibles en annexes.

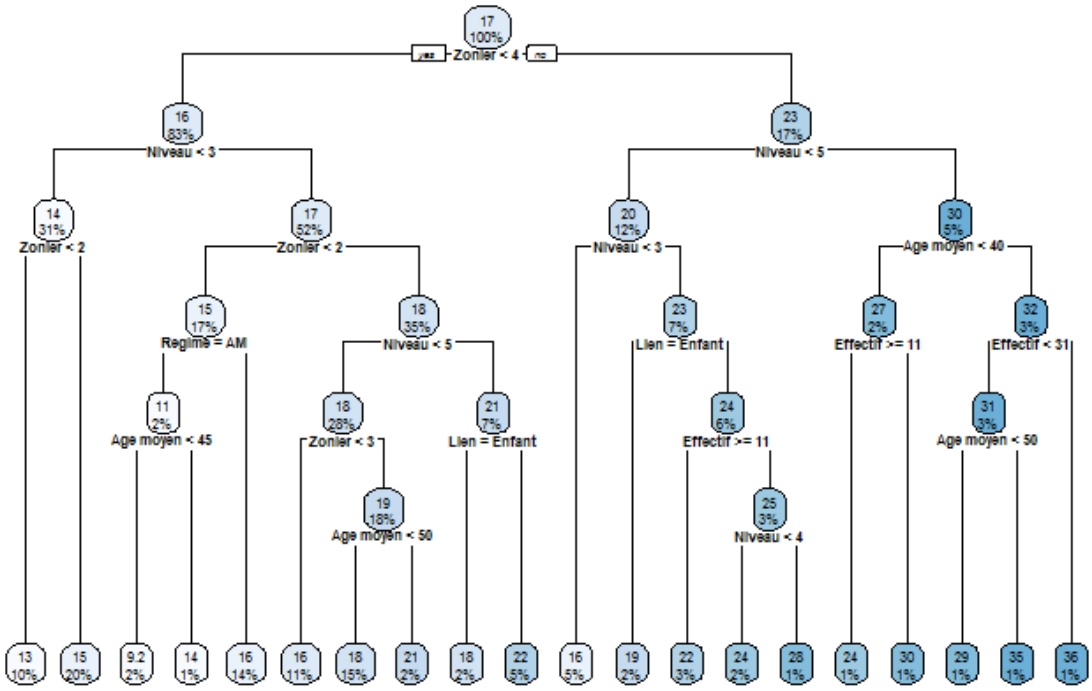


FIGURE 3.17 – Soins courants - Arbre CART sur le coût moyen

Les deux premières variables qui interviennent dans le processus de division ne sont pas les mêmes que pour la fréquence, il s'agit ici du zonier et du niveau de garantie.

Le zonier est par construction le reflet de l'hétérogénéité géographique de la consommation médicale de soins courants. Il traduit des pratiques de dépassement d'honoraires différentes selon les zones. La zone 4 correspond à une zone où les dépassements d'honoraires des médecins, notamment des médecins spécialistes sont particulièrement importants. Ceci se traduit par un montant de dépense par consultation plus élevé, ce qui se répercute sur le coût moyen pour l'assureur.

Le niveau de garantie correspond au niveau de prise en charge de la dépense par l'assureur, plus le niveau de garantie est important plus la part de la dépense par acte prise en charge par l'assureur est importante et donc plus le coût moyen est élevé.

Comme pour l'arbre de la fréquence, les segments en bleu foncé sur l'arbre correspondent aux segments avec les valeurs les plus importantes. L'échelle de valeur du coût moyen par segment est comprise entre 9,20 euros et 36 euros. Le coût moyen de 9,20 euros correspond à un coût moyen observé sur des bénéficiaires affiliés au régime Alsace-Moselle dans une entreprise dont l'âge moyen des salariés est inférieur à 45 ans et située sur la zone 1 avec un niveau de garantie supérieur à 2. A l'opposé, le coût moyen de 36 euros correspond au

segment composé de bénéficiaires rattachés à des entreprises de moins de 30 salariés de la zone 4 (zonier élevé) avec un niveau de garantie haute de gamme (niveau 5 et 6) et dont l'âge moyen des salariés est supérieur à 40 ans.

Sur le graphique 3.18 ci-dessous, les deux variables citées précédemment, le zonier et le niveau de garantie se détachent nettement des autres variables en terme d'importance. La troisième variable, le nombre de salariés a une importance relative à la variable principale inférieure à 20%.

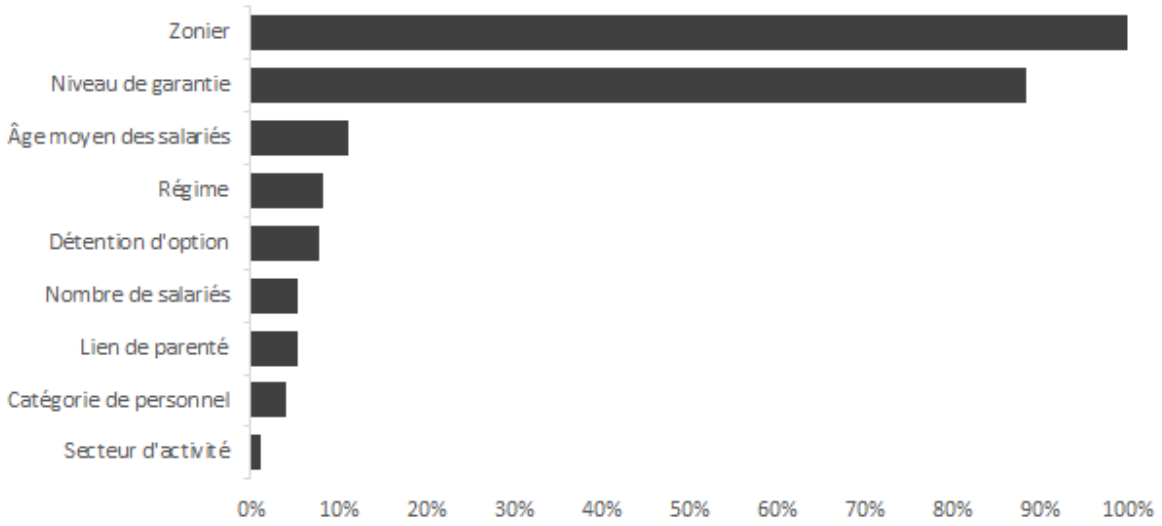


FIGURE 3.18 – Soins courants - Importance relative des variables du modèle coût moyen

Les critères de division déterminés au niveau des différents nœuds de l'arbre montrent que les modalités retenues pour le découpage influence la prime pure dans le sens attendu.

3.3.3.3 Qualité d'ajustement

L'erreur d'ajustement globale reste mesurée à la fois sur l'échantillon d'apprentissage et sur l'échantillon test. La charge de sinistres globale présente une légère sur-estimation de l'ordre de 1,8% liée principalement à une surestimation du coût moyen (Tableau 3.17 ci-après).

Les modèles élaborés pour la fréquence et le coût moyen diffèrent, les deux arbres présentés précédemment sur la fréquence et le coût moyen concernant le poste soins courants en sont la démonstration. Ainsi, même si le coût moyen modélisé est parfaitement ajusté sur la base de modélisation, cette différence de modèle est à l'origine de l'erreur d'ajustement observée sur l'échantillon d'apprentissage. En effet, le modèle de fréquence est parfaitement ajusté au global mais il ne l'est pas au niveau de chacun des segments définis par le modèle du coût moyen. Cet écart est à l'origine de l'erreur d'ajustement sur l'échantillon d'apprentissage et cette erreur se reproduit sur l'échantillon test. Cette erreur générée par l'approche fréquence-coût peut être corrigée par une recalibration.

Le tableau ci-dessous des erreurs d'ajustement par poste montre quelques variations selon les postes sur l'échantillon d'apprentissage.

	Echantillon d'apprentissage			Echantillon test		
	Erreur fréquence	Erreur coût moyen	Erreur Prestations	Erreur fréquence	Erreur coût moyen	Erreur Prestations
Hospitalisation	0,00%	3,24%	3,24%	0,50%	-2,39%	-1,90%
Soins courants	0,00%	2,40%	2,40%	-0,08%	2,47%	2,39%
Pharmacie	0,00%	2,94%	2,94%	-0,59%	3,28%	2,67%
Dentaire	0,00%	0,57%	0,57%	0,60%	2,07%	2,68%
Optique	0,00%	0,21%	0,21%	0,28%	1,27%	1,56%
Bien-être	0,00%	4,54%	4,54%	1,01%	3,84%	4,89%
Total	**0,00%**	**1,84%**	**1,84%**	**-0,15%**	**1,98%**	**1,82%**

TABLE 3.17 – Tableau des erreurs d'ajustement par poste

Le poste où l'erreur d'ajustement est la plus forte est le poste bien-être, viennent ensuite les postes hospitalisation, soins courants et pharmacie. Plus précisément, l'erreur d'ajustement sur le poste bien-être (+4.54%) traduit le fait que le modèle de fréquence sur estime la prévision du nombre d'actes sur les segments où le coût moyen modélisé du bien-être est le plus élevé et sous estime la prévision du nombre d'actes sur les segments où le coût moyen modélisé du bien-être est plus faible.

Concernant le poste hospitalisation, le sens de l'erreur diffère entre l'échantillon d'apprentissage (+3.24%) et l'échantillon test (-1.90%). Cet écart provient du coût moyen observé plus élevé sur l'échantillon test que sur l'échantillon d'apprentissage. Sur ce poste, il y a une forte volatilité des coûts, les prestations peuvent correspondre à une simple nuitée de type lit accompagnant à quelques euros à des frais de séjour de plusieurs milliers d'euros liés à une hospitalisation de plusieurs semaines.

3.3.4 Test de sensibilité

Le choix de l'approche, la valeur de certains paramètres de l'algorithme CART sont des éléments de nature à faire varier les estimations. Afin de mesurer comment ces choix peuvent influencer les résultats, un certain nombre de modélisations ont été testées. Cette section a pour objectif de présenter les résultats de ces différentes modélisations.

3.3.4.1 Présentation des modélisations réalisées

Parmi les éléments évoqués précédemment et susceptibles de faire varier les résultats, figure la valeur de certains paramètres de l'algorithme CART. Deux paramètres de l'algorithme feront l'objet de ces simulations, le seuil d'élagage et le nombre minimal d'observations au sein des nœuds terminaux. Concernant les seuils, le seuil intermédiaire et le seuil qui minimise l'erreur de validation croisée seront comparés. Concernant le nombre d'observations au sein des nœuds terminaux, les valeurs 500 et 1000 seront simulées. Par ailleurs, les deux approches modélisation fréquence-coût et modélisation directe de la prime pure seront testées. L'ensemble de ces variations conduit à la réalisation de 8 modélisations en incluant la modélisation détaillée précédemment.

Afin de comparer les résultats de ces modélisations, il est nécessaire de s'appuyer sur des indicateurs qui seront présentés dans le paragraphe suivant.

3.3.4.2 Critères de comparaison des modélisations

Les différentes modélisations évoquées précédemment seront comparées en considérant 2 critères, le RMSE et le MAE.

RMSE (Root Mean Square Error)

Le RMSE correspond à la racine carré de l'erreur quadratique moyenne, c'est-à-dire à la racine carré de la moyenne arithmétique des carrés des écarts entre les prévisions du modèle et les observations. Cet indicateur a tendance à renforcer l'effet des observations présentant des écarts importants.

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

MAE (Mean Absolute Error)

Le MAE correspond à la moyenne arithmétique des valeurs absolues des écarts entre les prévisions du modèle et les observations.

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

3.3.4.3 Résultats des modélisations

Les différentes modélisations montrent une légère variabilité des résultats selon les hypothèses et approches considérées. Le tableau 3.18 ci-après synthétisant les résultats indique que la simulation n^{o}8 donne les meilleurs résultats sur l'échantillon test avec des valeurs pour le RMSE et le MAE respectivement égales à 1829 et 883. Il s'agit d'une modélisation avec une approche type prime pure, un nombre minimal d'observations par nœud final de 500 et la règle d'élagage du seuil optimal.

N° de modélisation	Type de modélisation	Nombre minimal d'observations dans les nœuds terminaux	Seuil d'élagage	Echantillon d'apprentissage		Echantillon test	
				RMSE	MAE	RMSE	MAE
1	Fréquence * coût moyen	1000	Seuil intermédiaire	1812	873	1875	889
2	Fréquence * coût moyen	500	Seuil intermédiaire	1786	863	1859	884
3	Fréquence * coût moyen	1000	Seuil optimal	1793	866	1872	887
4	Fréquence * coût moyen	500	Seuil optimal	1764	857	1867	886
5	Prime pure	1000	Seuil intermédiaire	1858	884	1868	894
6	Prime pure	500	Seuil intermédiaire	1819	875	1842	888
7	Prime pure	1000	Seuil optimal	1814	871	1847	888
8	Prime pure	500	Seuil optimal	1769	861	1829	883

TABLE 3.18 – Résultats modélisation selon les paramètres

L'approche modélisation de la prime pure versus l'approche de modélisation fréquence - coût moyen influence les résultats de manière hétérogène selon les échantillons et l'indicateur considéré. Les erreurs importantes de prévisions sont réduites avec cette approche, le RMSE de l'échantillon test est systématiquement meilleur que sur les modélisations homologues avec l'approche fréquence coût moyen, ce qui n'est pas toujours le cas pour le MAE. L'approche prime pure permet également une réduction du surapprentissage, les écarts sur le RMSE entre l'échantillon d'apprentissage et l'échantillon test sont plus faibles.

Concernant le nombre minimal d'observations dans le nœud final, un abaissement de la valeur de ce paramètre de 1000 à 500 permet une légère amélioration de la performance du modèle de l'ordre de 0,7% pour le RMSE. De même, le choix du seuil optimal améliore également légèrement la performance. Cependant ces gains s'accompagnent d'une forte augmentation de la complexité des arbres.

Après l'analyse des résultats des différentes modélisations du tableau précédent 3.18 et en s'appuyant sur les critères de performance RMSE et MAE définis précédemment, deux modélisations ont été retenues. La première modélisation retenue correspond à la modélisation ayant obtenu la meilleure performance avec l'approche fréquence coût moyen. Il s'agit de la modélisation n^{o}2 du tableau 3.18, modélisation avec l'application de la règle d'élagage du seuil intermédiaire et une valeur du paramètre du nombre d'observations minimal par nœud final fixée à 500. Cette modélisation sera comparable à la modélisation GLM car elle utilise les mêmes variables et le même type d'approche. La seconde modélisation retenue est celle ayant obtenu les meilleurs résultats parmi l'ensemble des modélisations testées avec l'approche CART. Elle correspond à la modélisation n^{o}8 du tableau 3.18, il s'agit d'une modélisation avec une approche prime pure, avec les mêmes variables que pour le GLM, avec une valeur du paramètre du nombre minimal d'observations par nœud final de 500 et l'application de la règle d'élagage du seuil optimal. Ces deux modélisations feront l'objet d'une analyse comparative avec les autres méthodes dans le cadre de la quatrième partie.

3.3.5 Conclusion sur la modélisation avec CART

L'un des atouts majeurs de cette méthode est la lisibilité et la facilité d'interprétation de ses résultats. Présentés sous forme graphique, leur interprétation s'avère relativement aisée quelque soit le type d'interlocuteur. Cette méthode présente également l'avantage d'avoir des temps de calcul relativement rapide. L'autre atout majeur réside dans le caractère non paramétrique de cette méthode, elle ne nécessite pas d'hypothèses sur les distributions des variables.

Cependant, cette méthode comporte quelques inconvénients. L'algorithme CART est une méthode qui suit une stratégie pas à pas hiérarchisée, elle ne conduit donc pas à un optimum global mais à un optimum local. Un autre inconvénient, le principal de cette méthode et qui concerne toutes les méthodes d'arbres de décision, est son manque de stabilité et son irrégularité. Une faible fluctuation de l'échantillon peut générer une modification importante des résultats.

Afin de palier à ces faiblesses de la méthode CART, des solutions existent aux travers des méthodes d'agrégation de modèles connus sous les nom bagging et boosting. La prochaine section sera dédiée à la modélisation à l'aide de la méthode des forêts aléatoires.

3.4 Modélisation de la prime pure avec la méthode des forêts aléatoires

La méthode des forêts aléatoires ou Random Forest est une technique d'agrégation de modèles qui a été introduite par Breiman en 2001. L'avantage de ce type de technique est d'améliorer l'ajustement en agrégeant les résultats d'un grand nombre de modèles tout en évitant ou contrôlant le surajustement. La particularité de la méthode des forêts aléatoires par rapport aux autres méthodes d'agrégation est d'associer une technique d'agrégation de modèles de type bagging avec l'introduction d'une composante aléatoire. Plus précisément, l'objectif est d'apporter plus de singularité entre les différents modèles en ajoutant de l'aléa sur la sélection des variables intervenant dans les modèles.

3.4.1 Principe de la méthode des forêts aléatoires

Tout d'abord, la technique des forêts aléatoires fait référence à la stratégie d'agrégation de modèles de type Bagging introduite également par Léo Breiman en 1996. Avant de développer les particularités de cette méthode, il est donc nécessaire de présenter ce en quoi consiste le bagging.

3.4.1.1 Bagging

Le terme bagging provient de l'association des termes **B**ootstrap **Agg**regat**ing**. Il s'agit d'une technique visant à améliorer la classification notamment celle des arbres de décision réputés instables. L'objectif est de diminuer la variance des estimateurs en moyennant les résultats de sous-échantillons obtenus par bootstrap, c'est à dire par tirage aléatoire avec remise.

Principe d'agrégation de modèles

Soit Y=$(Y_1, ..., Y_n)$ une variable à expliquer quantitative ou qualitative composées de n observations, $X^1, ..., X^k$ les variables explicatives et f(x) un modèle fonction de x = $x^1, ..., x^k$ ε R^k.

On note Z = $\{(X_1, Y_1)$,...,$(X_n, Y_n)\}$ un échantillon de loi F.

En considérant B échantillons indépendants notés $\{Z_b\}$ avec b=1,..,B, il est possible d'obtenir une prévision par agrégation de modèles en fonction de la nature de la variable à expliquer Y. Dans le cas où Y est quantitative, le résultat de la prévision correspondra à la moyenne des résultats des modèles obtenus sur chacun des sous-échantillons et sera déduit de l'équation suivante :

$$\widehat{f}_B(.) \ = \ \frac{1}{B} \sum_{b=1}^{B} \widehat{f}_{Zb}(.)$$

Si Y est une variable qualitative, le résultat de la prévision correspondra à la modalité qui a obtenu la majorité des votes au sein de l'ensemble des modèles, ce qui est équivalent au résultat de la formule suivante :

$$\widehat{f}_B(.) \ = \ arg\ max_j\ card\{\ b|\widehat{f}_{Zb}(.) = j\}$$

Le principe de cette méthode est de moyenner la prévision de plusieurs modèles indépendants afin de réduire la variance et l'erreur de prévision. Cependant, la constitution de B échantillons nécessite de disposer d'un volume de données très important et rarement disponible. La solution à ce problème consiste à déterminer les échantillons avec la technique du bootstrap.

Principe du bootstrap

Le principe du bootstrap consiste à construire les B sous échantillons à partir de n tirages aléatoires avec remise selon la distribution empirique $\widehat{F}$.

Algorithme 1 : Bagging

Soit x_0 à prévoir et
Z = $\{(X_1, Y_1)$,...,$(X_n, Y_n)\}$ un échantillon.

Pour b = 1 à B
Tirer un échantillon bootstrap Z_b
Estimer $\widehat{f}_B(x_0)$ sur l'échantillon bootstrap.

Calculer l'estimation moyenne $\widehat{f}_B(x_0) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_{Zb}(x_0)$

Erreur out of bag

A partir de ces estimations, il est possible de calculer l'erreur de prévision comme pour la validation croisée dans le cadre de l'algorithme CART. Pour chaque observation (y_i, x_i), seuls les modèles estimés à partir des échantillons bootstrap ne contenant pas l'observation i seront considérés. Cela représente 1/3 des modèles environ. En considérant ce sous-ensemble de modèles, il est possible de déterminer une prévision $\widehat{y}_i$ et de calculer l'erreur de prévision associée à cette observation.

3.4.1.2 Forêts aléatoires

La méthode des forêts aléatoires fait référence au cas spécifique des modèles d'arbres de décision binaire CART et constitue une amélioration du bagging par l'ajout d'une composante aléatoire. Cette composante aléatoire se traduit par une sélection au hasard des variables permettant d'élaborer les différents modèles. L'objectif de l'introduction de cet aléa est de rendre les différents modèles d'agrégation plus indépendants et réduire ainsi la variance. En effet, la variance de la moyenne de B variables indépendantes, identiquement distribuées, chacune de variance σ^2 est égale à σ^2/B. Ainsi plus le nombre B de sous échantillons est important et plus la variance diminue.

Cependant, ce constat peut s'avérer plus nuancé en présence de corrélation entre les variables. En 2001, Breiman montre que si les variables sont identiquement distribuées mais

présentent une corrélation deux à deux ρ, alors la variance de la moyenne devient de la forme suivante :

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Le second terme décroît bien avec B mais le premier terme est indépendant de B, ce qui réduit l'intérêt du bagging si ρ est grand, c'est-à-dire en présence d'une forte corrélation. L'objectif de la randomisation introduite dans le cadre des forêts aléatoires est de réduire la corrélation ρ entre les prévisions issues des différents modèles. Les modèles ainsi construits avec des prédicateurs différents sont plus indépendants.

Algorithme 2 : Forêts aléatoires

Soit x_0 à prévoir et
Z = $\{(X_1, Y_1)$,...,$(X_n, Y_n)\}$ un échantillon.

Pour b = 1 à B
Tirer un échantillon bootstrap Z_b
Tirer aléatoirement un sous-ensemble de m prédicteurs
Estimer $\widehat{f_B}(x_0)$ sur l'échantillon bootstrap à partir du sous-ensemble de m prédicteurs.

Calculer l'estimation moyenne $\widehat{f_B}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_{Zb}(x_0)$

3.4.2 Paramétrage de l'algorithme des forêts aléatoires

L'algorithme des forêts aléatoires fait intervenir des paramètres identiques à ceux de l'algorithme CART ainsi que deux paramètres qui lui sont propres : le nombre d'échantillons bootstrap B et le nombre m de prédicateurs tirés aléatoirement.

Concernant les paramètres similaires à l'algorithme CART, les mêmes valeurs que précédemment seront appliquées. Le nombre d'observations minimales par feuille sera fixé à 1000 et la profondeur de l'arbre sera limitée à 11.

Sélection du nombre d'arbres

Le choix du nombre d'arbres peut être déterminé à partir d'une analyse graphique de l'évolution de l'erreur OOB (out of bag) en fonction du nombre d'arbres. Il s'agira de choisir le nombre d'arbres à partir duquel l'erreur se stabilise.

Le graphique 3.19 ci-dessous illustre le cas de la modélisation de la fréquence pour le poste soins courants. Au delà de 200 arbres, l'erreur ne s'améliore plus. Le nombre d'échantillons bootstrap B a donc été fixé à 200. Il a été procédé à la même analyse pour l'ensemble des postes.

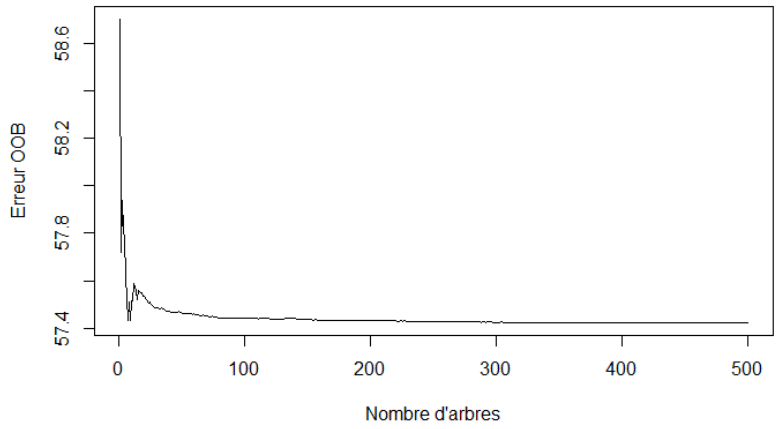


FIGURE 3.19 – Poste soins courants - Modèle fréquence - Erreur OOB en fonction du nombre d'arbres

Sélection du nombre de prédicateurs

La valeur par défaut de ce paramètre est généralement égale à $\sqrt{p}$ avec p le nombre de variables explicatives totales. Une autre manière de déterminer ce paramètre consiste à choisir la valeur du paramètre qui minimise l'erreur OOB. Le graphique ci-dessous de l'évolution de l'erreur en fonction du nombre de prédicateurs sur le poste soins courants suggère une valeur du nombre de prédicateurs optimale m=4.

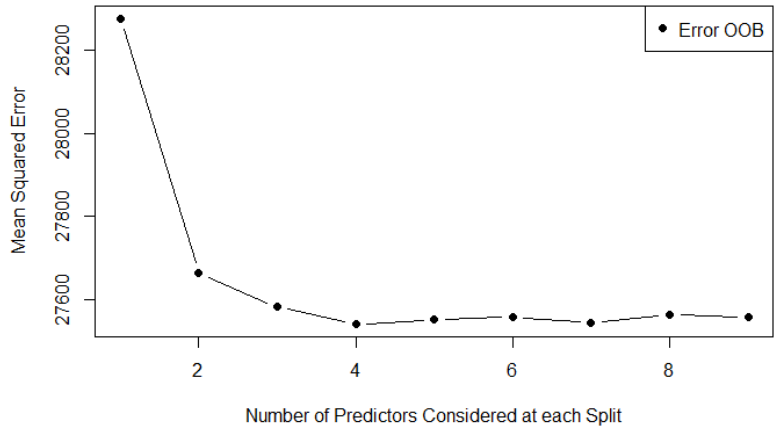


FIGURE 3.20 – Poste soins courants - Modèle fréquence - Erreur OOB en fonction du nombre de prédicateurs

3.4.3 Application de l'algorithme des forêts aléatoires avec la même approche que pour le GLM

La démarche dans le cadre de la modélisation de la prime pure avec l'algorithme des forêts aléatoires est identique à celle employée avec la méthode CART. Une première modélisation avec l'approche fréquence-coût moyen sera réalisée en considérant le nombre d'observations minimal à 1000.

3.4.3.1 Résultats de la modélisation

L'algorithme des forêts aléatoires ne permet pas comme pour les deux précédentes méthodes une restitution des résultats du modèle aussi explicite, d'où le qualificatif de boîte noire parfois employé. Les valeurs prédites proviennent de l'agrégation de plusieurs centaines d'arbres, il n'est donc pas possible de visualiser l'ensemble de ces arbres.

Toutefois, il est possible comme pour la méthode CART de visualiser l'importance relative des variables intervenues dans la modélisation. Les deux graphiques suivants, figures 3.21 et 3.22, représentent l'importance relative des variables dans le cadre de la modélisation de la fréquence et du coût moyen sur le poste soins courants.

L'ordre et le degré d'importance relative des variables sont concordant avec ce qui a été observé avec la méthode CART. Le lien de parenté est la variable la plus importante sur le modèle de fréquence. Le zonier et le niveau de garantie le sont sur le modèle du coût moyen.

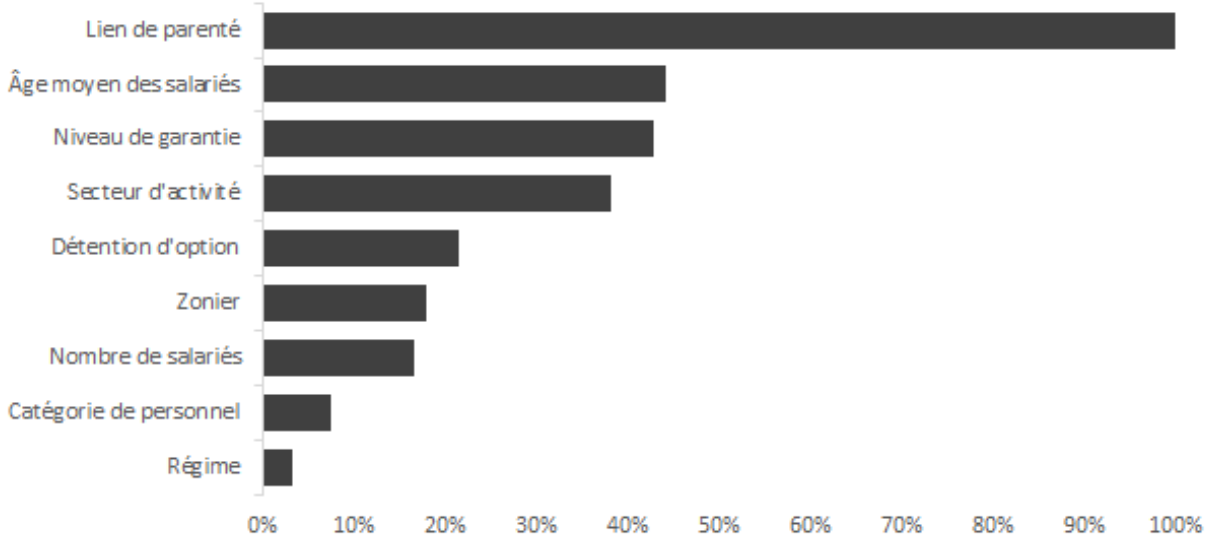


FIGURE 3.21 – Soins courants - Importance relative des variables du modèle fréquence

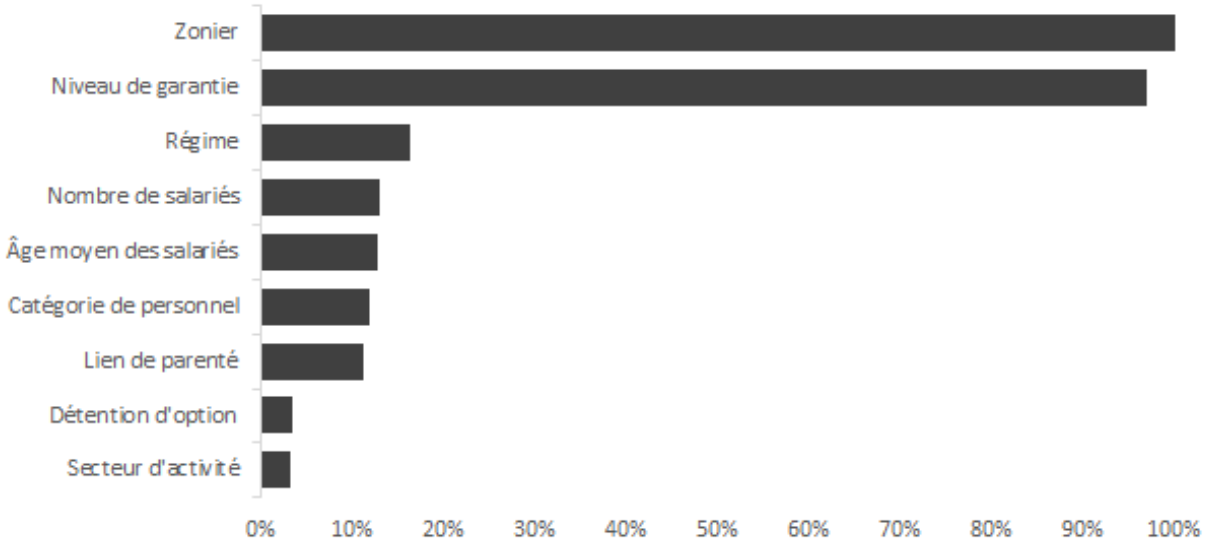


FIGURE 3.22 – Soins courants - Importance relative des variables du modèle coût moyen

3.4.3.2 Qualité d'ajustement

L'erreur d'ajustement globale reste faible à la fois sur l'échantillon d'apprentissage (1,94%) et sur l'échantillon test (1,69%). Les écarts par poste sont similaires à ce qui a été observé avec CART et sont dûs aux mêmes effets.

	Echantillon d'apprentissage			Echantillon test		
	Erreur fréquence	Erreur coût moyen	Erreur Prestations	Erreur fréquence	Erreur coût moyen	Erreur Prestations
Hospitalisation	0,02%	3,22%	3,24%	0,04%	-2,16%	-2,12%
Soins courants	-0,02%	2,23%	2,22%	-0,23%	2,28%	2,05%
Pharmacie	0,03%	2,91%	2,94%	-0,61%	3,28%	2,66%
Dentaire	0,04%	0,90%	0,94%	0,50%	1,57%	2,08%
Optique	-0,05%	0,51%	0,46%	0,71%	1,26%	1,98%
Bien-être	-0,01%	4,74%	4,73%	1,09%	3,89%	5,02%
Total	**0,00%**	**1,94%**	**1,94%**	**-0,23%**	**1,93%**	**1,69%**

TABLE 3.19 – Tableau des erreurs d'ajustement par poste

3.4.4 Test de sensibilité

De la même manière que précédemment avec l'algorithme CART, différentes modélisations ont été testées en faisant varier l'approche et la valeur de certains paramètres de l'algorithme. Les résultats de ces modélisations seront présentés dans cette section.

3.4.4.1 Présentation des modélisations réalisées

Comme pour l'algorithme CART, l'approche et le nombre minimal d'observations au sein des nœuds terminaux, 500 et 1000 feront l'objet de ce test. La valeur du paramètre du nombre de prédicateurs m a été fixé à 3 ou 4 selon les postes de soins. Un test sera réalisé en considérant l'ensemble des variables, soit m=9. Dans ce dernier cas, il n'y a plus de tirage aléatoire des prédicateurs ce qui correspond à un modèle BAGGING. Au total, 8 modélisations seront testées en incluant la modélisation détaillée précédemment.

3.4.4.2 Résultats des modélisations

Les résultats des différentes modélisations sont comparés en considérant à nouveau les indicateurs RMSE et MAE. Le tableau 3.20 ci-après montre que les résultats des différentes modélisations sont relativement proches, les écarts de performance sont inférieurs à 1% entre les différents modèles.

N° de modélisation	Type de modélisation	Nombre minimal d'observations dans les nœuds terminaux	Nombre de prédicateurs	Echantillon d'apprentissage		Echantillon test	
				RMSE	MAE	RMSE	MAE
1	Fréquence * coût moyen	1000	3-4	1736	849	1837	878
2	Fréquence * coût moyen	500	3-4	1695	838	1837	878
3	Prime pure	1000	3-4	1708	843	1829	878
4	Prime pure	500	3-4	1664	831	1831	880
5	Fréquence * coût moyen	1000	Toutes	1717	842	1832	878
6	Fréquence * coût moyen	500	Toutes	1677	831	1841	881
7	Prime pure	1000	Toutes	1686	837	1820	879
8	Prime pure	500	Toutes	1638	824	1839	883

TABLE 3.20 – Résultats modélisation selon les paramètres

Les modèles de type forêt aléatoire (modèles 1 à 4) fournissent une performance très proche voire légèrement meilleure au sens du MAE que les modèles BAGGING homologues (modèles 5 à 8). En ce qui concerne le RMSE, les résultats sont variables. Ce faible écart de performance entre les deux types de modélisation est vraisemblablement dû au nombre de prédicateurs disponibles trop faible.

Le modèle permettant d'obtenir la meilleure performance selon le critère RMSE est le modèle BAGGING avec l'approche prime pure et un nombre minimal d'observations par nœud égal à 1000 (modèle n°7). Trois autres modélisations seront retenues pour l'analyse des résultats, les n°1, 3 et 5 qui sont les meilleurs modèles selon le type de modélisation, bagging ou forêt aléatoire et le type d'approche.

Chapitre 4

Analyse des résultats et impact du 100% santé

Une diversité de techniques de modélisation, GLM, CART, Forêts aléatoires et bagging ont été exposées dans le cadre du chapitre précédent. Leur mise en application sur un portefeuille d'assurés santé collective a donné lieu à l'élaboration d'un ensemble de modèles chacun caractérisé par un niveau de performance propre. La première partie de ce chapitre sera dédiée à l'analyse des résultats de ces différents modèles. Il s'agira de comparer leur performance globale en s'appuyant sur différents critères de mesure. Cependant, la mesure de la performance globale n'est pas suffisante pour préjuger de la qualité d'un modèle, des analyses plus fines sont également nécessaires. Elles consisteront à mesurer la qualité de l'ajustement des modèles sur les différents critères tarifaires.

Dans une seconde partie, les impacts liés à la réforme du 100% santé seront abordés. Cette réforme modifie le comportement des assurés sur certains postes de soins, ce qui n'est pas sans conséquence sur le montant des prestations à la charge des complémentaires santé. L'essentiel des mesures de cette réforme ont été mises en œuvre au 1^{er} janvier 2020, il est donc possible de mesurer ses premiers effets même avec le contexte de crise sanitaire qui bouleverse également le comportement des assurés. A partir de la sinistralité observée sur les dix premiers mois de l'année 2020, ce chapitre tentera d'apporter une première estimation des impacts engendrés par cette réforme.

4.1 Analyse des résultats

Dans le cadre de la partie modélisation de la prime pure, un ensemble de modèles ont été déterminés. Pour chacune des méthodes et types d'approches employées, le modèle présentant le meilleur niveau de performance au sens de RMSE sur l'échantillon test a été retenu. Un ensemble de huit modèles listés ci-dessous ont été sélectionnés et feront l'objet de l'analyse comparative menée dans ce chapitre.

- un modèle GLM avec l'approche fréquence - coût moyen
- un modèle GLM avec interaction (et retraitement du lien de parenté) et l'approche fréquence - coût moyen
- un modèle CART avec l'approche fréquence - coût moyen
- un modèle CART avec l'approche prime pure

- un modèle BAGGING avec l'approche fréquence - coût moyen
- un modèle BAGGING avec l'approche prime pure
- un modèle RANDOM FOREST avec l'approche fréquence - coût moyen
- un modèle RANDOM FOREST avec l'approche prime pure

Les résultats des différents modèles seront également comparés avec le tarif actuel.

4.1.1 Étude de la performance des modèles

L'objectif de cette section est de comparer la performance des différents modèles entre eux mais également avec le tarif en vigueur. L'évaluation de la performance sera réalisée à l'aide des critères RMSE et MAE permettant de refléter l'erreur moyenne de prévision. La qualité de l'ajustement sur l'échantillon test sera également abordée dans un second temps.

4.1.1.1 Critères RMSE et MAE

Les critères du RMSE et du MAE ne seront pas détaillés ici, ils ont déjà été présentés à la section 3.3.4.2. Le tableau 4.1 ci-dessous met en évidence un niveau de performance très proche entre les différents modèles selon ces deux critères. La meilleure performance sur l'échantillon test au sens du RMSE est obtenue avec le modèle BAGGING et une approche prime pure (RMSE=1820). Avec l'indicateur MAE, trois modèles donnent des résultats équivalents. En considérant le MAE comme premier critère et le RMSE comme second critère, c'est le modèle Random Forest avec l'approche prime pure qui fournit la meilleure performance (MAE=878 / RMSE 1829).

Méthode de modélisation	Type de modélisation	Nombre minimal d'observations dans les nœuds terminaux	Seuil d'élagage	Echantillon d'apprentissage		Echantillon test	
				RMSE	MAE	RMSE	MAE
Tarif actuel				1971	920	1969	923
GLM	Fréquence * coût moyen			1830	874	1834	889
GLM avec interactions	Fréquence * coût moyen			1817	869	1841	888
CART	Fréquence * coût moyen	500	Seuil intermédiaire	1786	863	1859	884
	Prime pure	500	Seuil optimal	1769	861	1829	883
BAGGING	Fréquence * coût moyen	1000		1717	842	1832	878
	Prime pure	1000		1686	837	1820	879
RANDOM FOREST	Fréquence * coût moyen	1000		1736	849	1837	878
	Prime pure	1000		1708	843	1829	878

TABLE 4.1 – Résultats des modèles

En considérant l'approche fréquence-coût moyen et l'indicateur RMSE, la méthode GLM donne des résultats similaires aux méthodes BAGGING et RANDOM FOREST sur l'échantillon test. Avec le MAE, le GLM est légèrement mois performant que les autres méthodes, l'écart est de l'ordre de 1.25%. Par ailleurs, les écarts de résultats entre l'échantillon d'apprentissage et l'échantillon test sont plus réduits avec le modèle GLM, le RMSE est de 1830 sur l'échantillon d'apprentissage et de 1834 sur l'échantillon test. Le modèle GLM apparaît plus stable et moins sensible au risque de sur-apprentissage. Le modèle GLM avec la prise en compte des interactions est meilleur sur l'échantillon d'apprentissage mais cette tendance ne se confirme pas sur l'échantillon test.

L'écart de performance des huit modèles avec le tarif actuel est significatif. Le gain de performance des modèles élaborés est de l'ordre de 7% pour le RMSE et 4,5% sur le MAE.

4.1.1.2 Qualité prédictive des modèles

La comparaison sur un échantillon test de la prime pure modélisée avec la charge de sinistres observée permet de juger du pouvoir prédictif des modèles. L'indicateur S/P net correspondant au rapport de la charge de sinistres observée sur le montant de la prime pure modélisée a été retenu pour illustrer la qualité de l'adéquation.

	Echantillon test								
	Tarif actuel	GLM	GLM avec interaction	CART FCM*	CART PP**	Bagging FCM*	Bagging PP**	Random Forest FCM*	Random Forest PP**
Hospitalisation	106,23%	105,05%	105,15%	105,15%	105,59%	105,59%	105,81%	105,59%	105,84%
Soins courants	104,46%	100,52%	100,55%	100,32%	100,10%	100,12%	100,22%	100,22%	100,36%
Pharmacie		100,34%	100,36%	100,18%	100,57%	100,24%	100,19%	100,36%	100,31%
Dentaire	88,03%	98,33%	98,33%	98,08%	97,15%	98,67%	98,41%	98,89%	98,44%
Optique	98,61%	98,92%	98,71%	98,51%	98,30%	98,31%	98,49%	98,51%	98,42%
Bien-être	102,24%	99,39%	99,34%	99,16%	99,03%	99,87%	98,86%	99,95%	99,26%
Total	**99,52%**	**100,23%**	**100,20%**	**100,02%**	**99,81%**	**100,15%**	**100,12%**	**100,28%**	**100,19%**

TABLE 4.2 – S/P net par modèle et poste de soins
(* : Approche fréquence-coût moyen / ** : Approche prime pure)

Le S/P net global est proche de 100% pour les sept modèles ainsi que pour le tarif en vigueur, l'écart est inférieur à 0,5% ce qui traduit une bonne précision globale. En dehors du poste hospitalisation, la qualité de la précision est également confirmée au niveau des postes de soins. Sur le poste hospitalisation, la prime modélisée est inférieure de 5% à la charge de sinistres observée. Cet écart s'explique par la sinistralité réelle légèrement supérieure sur l'échantillon test que sur l'échantillon d'apprentissage, à la fois sur la fréquence et le coût moyen. Ce poste présente une certaine volatilité.

4.1.2 Capacité de segmentation des modèles

L'indice de GINI (ou coefficient de GINI) est un indicateur synthétique permettant de refléter la répartition d'une variable au sein d'une population. Il a été créé à l'origine pour mesurer le degré d'égalité ou d'inégalité de la répartition des richesses dans une société. Sa valeur est comprise entre 0 et 1, une valeur de 0 correspondant à une société très égalitaire et à l'inverse une valeur de 1 correspondant à une société très inégalitaire dans laquelle un seul individu détiendrait l'ensemble de la richesse. Cet indicateur est également très utilisé pour évaluer la capacité d'un modèle à segmenter le risque. Il peut être illustré graphiquement à l'aide de la courbe de Lorenz.

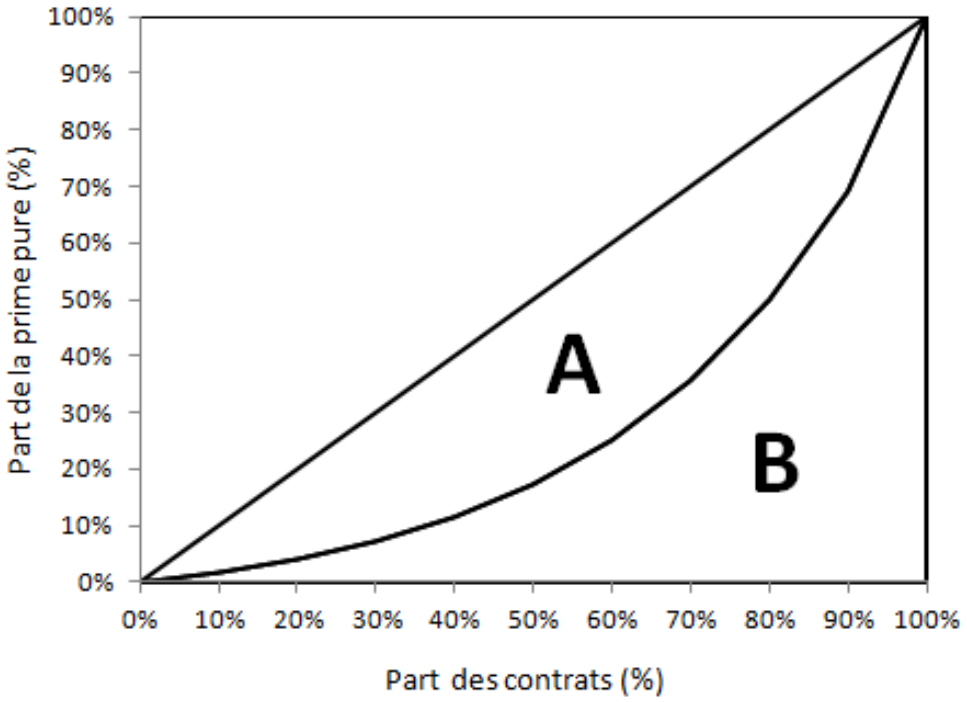


FIGURE 4.1 – Courbe de Lorenz

L'indice de GINI correspond au rapport de l'aire entre la courbe et la bissectrice (A) et l'aire sous la bissectrice (A+B).

$$Indice\ de\ GINI\ =\ \frac{A}{A+B}\ =\ 2A\ \ car\ A+B=1/2$$

La valeur optimale de l'indice est obtenue à partir de la courbe de Lorenz de la charge de sinistres, soit 72,5 % sur le portefeuille étudié. Une valeur proche de l'indice optimal illustre une bonne capacité d'un modèle à segmenter la prime pure.

Ce critère n'est pas suffisant pour juger de la qualité d'un modèle, une analyse complémentaire doit être menée afin de s'assurer que la segmentation du risque définie par le modèle est en adéquation avec le risque observé.

Les courbes de Lorenz des différents modèles apparaissent superposées sur le graphique 4.2. Ceci suggère de faibles écarts de l'indice de GINI entre les modèles. Les courbes des modèles sont également légèrement en dessous de la courbe du tarif ce qui indique une plus forte segmentation de la prime sur les modèles élaborés.

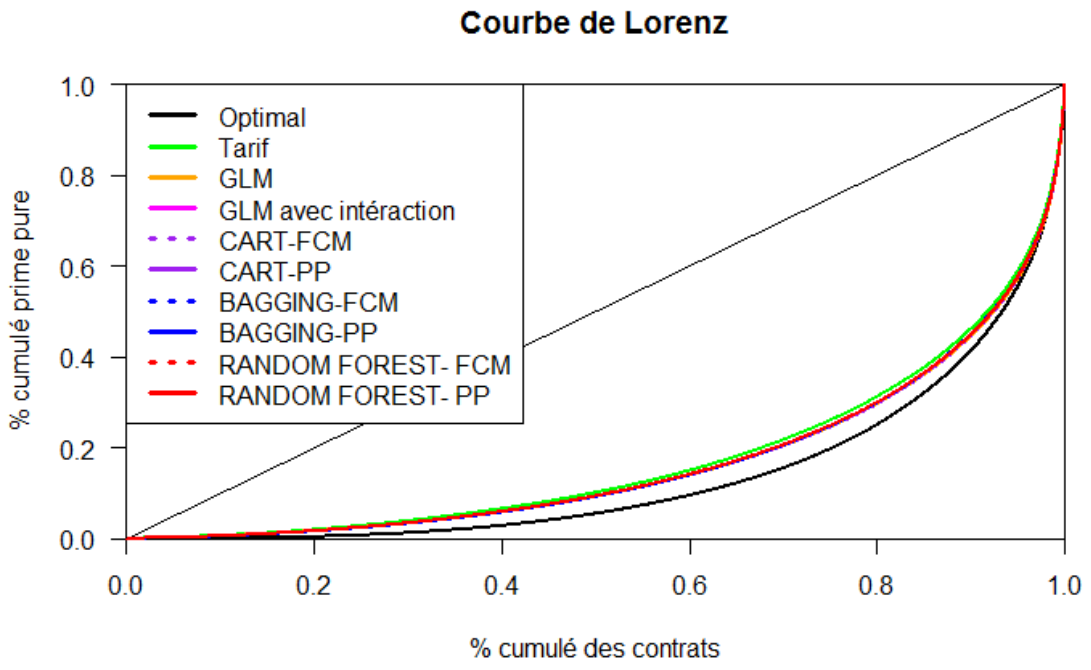


FIGURE 4.2 – Courbes de Lorenz des modèles sur l'échantillon test

(*) : Approche fréquence-coût moyen
(**) : Approche prime pure

Le modèle GLM avec interaction est le modèle sur lequel l'indice de GINI est le plus élevé, 66.93% sur l'échantillon test. Comme déjà identifié graphiquement, les écarts sont extrêmement faibles entre les modèles. L'indice de GINI des modèles est en moyenne supérieur de 2,6% à celui du tarif actuel.

Méthode de modélisation	Type de modélisation	Nombre minimal d'observations dans les nœuds	Seuil d'élagage	**Echantillon d'apprentissage**	**Echantillon test**
Optimal				72,55%	72,48%
Tarif actuel				65,23%	65,16%
GLM	Fréquence * coût moyen			67,16%	66,85%
GLM avec interaction	Fréquence * coût moyen			67,21%	66,93%
CART	Fréquence * coût moyen	500	Seuil intermédiaire	67,12%	66,88%
	Prime pure	500	Seuil optimal	66,96%	66,78%
BAGGING	Fréquence * coût moyen	1000		67,04%	66,75%
	Prime pure	1000		67,03%	66,74%
RANDOM FOREST	Fréquence * coût moyen	1000		66,87%	66,51%
	Prime pure	1000		66,91%	66,59%

TABLE 4.3 – Indice de GINI par modèle

4.1.3 Analyse des résultats selon les critères tarifaires

4.1.3.1 Niveau de garantie

Les modèles GLM et Bagging présentent la meilleure qualité d'ajustement par niveau de garantie, les écarts avec le S/P net d'équilibre sont inférieurs à 5 points sur l'échantillon test. Le modèle Random Forest est quant à lui le moins bien ajusté. Sur les niveaux de garantie d'entrée de gamme (niveaux 0 et 1), il surévalue le niveau de la prime et particulièrement sur le niveau 0 où le S/P net est de 93%, soit une surtarification de 7%. Sur le niveau 6, une légère sous-tarification des 4 modèles est observée. Concernant le tarif en vigueur, une surtarification sur le niveau 1 est à noter.

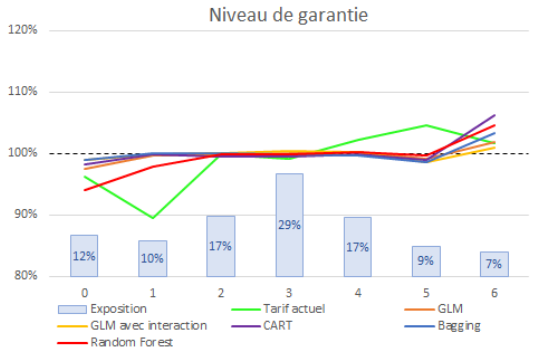


FIGURE 4.3 – S/P net sur l'échantillon d'apprentissage

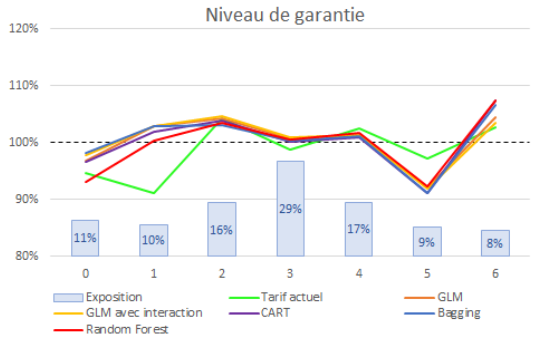


FIGURE 4.4 – S/P net sur l'échantillon test

4.1.3.2 Âge moyen des salariés

Sur l'échantillon d'apprentissage, les modèles sont tous bien ajustés par tranche d'âge excepté sur le segment des moins de 30 ans où les modèles CART, Bagging et Random Forest surestiment la prime pure. Cet écart d'ajustement peut s'expliquer par la faible part de ce segment dans l'exposition totale. Sur l'échantillon test, les S/P net des modèles sont proches de 100%. Avec le tarif en vigueur, la tranche d'âge des 45-49 ans semble surtarifée, ceci peut s'expliquer par la segmentation des classes d'âge moins fines sur celui-ci. Les 45-49 ans et les 50 ans et plus ne sont pas dissociés avec le tarif en vigueur, ils sont dans une même classe des 45 ans et plus.

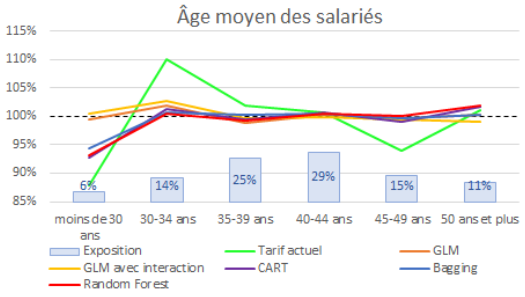


FIGURE 4.5 – S/P net sur l'échantillon d'apprentissage

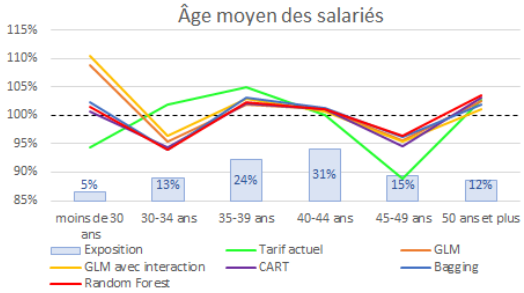


FIGURE 4.6 – S/P net sur l'échantillon test

4.1.3.3 Lien de parenté

Les modèles CART et Bagging sont parfaitement ajustés sur les deux échantillons. Le modèle Random Forest est bien ajusté sur les affiliés, en revanche il surestime légèrement la prime pure des conjoints et sous-estime celle des enfants. Les modèles GLM apparaissent comme les modèles les moins bien ajustés sur cette variable avec une sous-estimation de la prime pure des salariés et une surestimation pour les enfants. Le tarif en vigueur ne fait pas de distinction tarifaire entre les salariés et les conjoints. Or, le niveau de consommation médicale des conjoints est plus important que celui des salariés, ce qui génère une sous-estimation de la prime pure sur les conjoints et une surestimation sur les salariés.

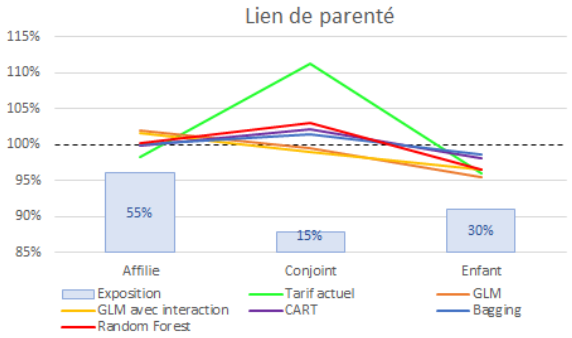


FIGURE 4.7 – S/P net sur l'échantillon d'apprentissage

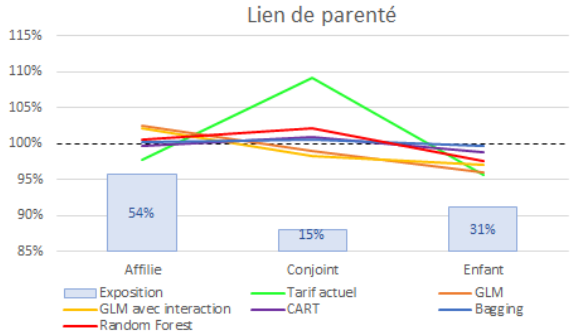


FIGURE 4.8 – S/P net sur l'échantillon test

Le caractère multiplicatif du modèle GLM est de nature à expliquer son moins bon ajustement sur les enfants. L'application de certains coefficients tarifaires notamment ceux liés à l'âge moyen des salariés de l'entreprise dans le cadre du modèle GLM sans interaction ne sont pas très pertinents sur cette population. Il n'y a pas de raison pour que les enfants de salariés rattachés à des entreprises où l'âge moyen est plus important consomment plus. Le graphique 10.12 du S/P net par classe d'âge des salariés sur le segment des enfants ci-dessous illustre ce problème. Le GLM sans interaction sous-tarifie les enfants sur les entreprises où l'âge moyen des salariés est inférieur à 35 ans et surtarifie lorsque l'âge est supérieur à 45 ans. Le modèle GLM avec interaction permet de corriger ce biais.

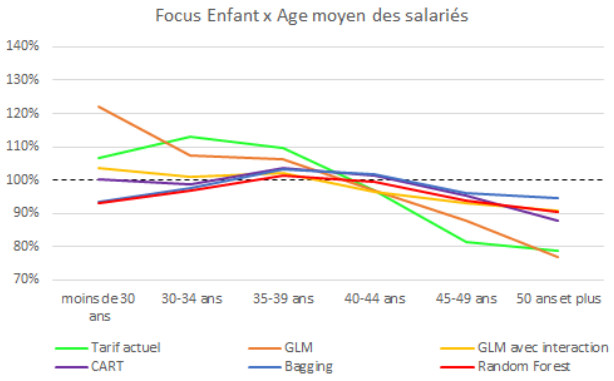


FIGURE 4.9 – Indice de GINI par modèle

4.1.3.4 Nombre de salariés

Les modèles BAGGING et Random Forest figurent comme les deux modèles apportant le meilleur ajustement. Sur le tarif actuel, une surestimation de la prime pure sur les entreprises avec un seul salarié et une sous-estimation sur les entreprises de plus de 30 salariés est observée.

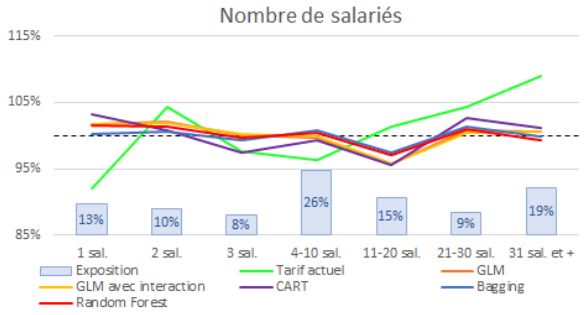


FIGURE 4.10 – S/P net sur l'échantillon d'apprentissage

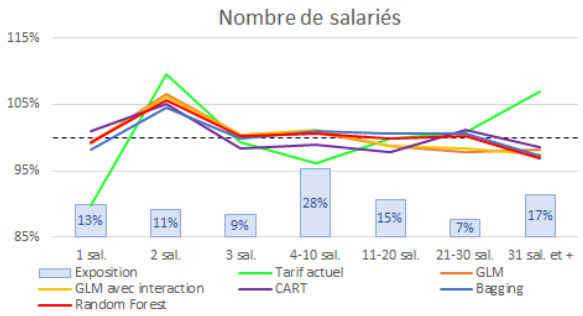


FIGURE 4.11 – S/P net sur l'échantillon test

4.1.3.5 Catégorie de personnel

Les cinq modèles apparaissent parfaitement ajustés sur les différentes modalités de la catégorie de personnel. Concernant le tarif en vigueur, le S/P net des cadres est à 92% sur les deux échantillons traduisant une légère surestimation de la prime pure. Sur les non cadres, c'est le phénomène inverse qui est constaté.

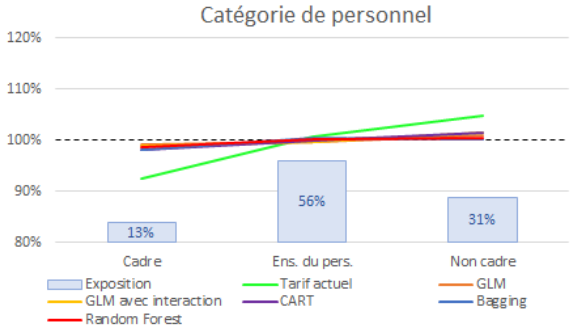


FIGURE 4.12 – S/P net sur l'échantillon d'apprentissage

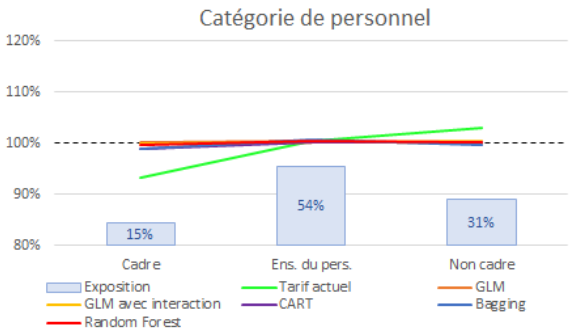


FIGURE 4.13 – S/P net sur l'échantillon test

4.1.3.6 Secteur d'activité

Les modèles GLM sont bien ajustés sur l'échantillon d'apprentissage. Les autres modèles ont tendance à surestimer la prime pure sur la modalité secteur 1 et à la sous-estimer légèrement sur la modalité secteur 3. Sur l'échantillon test, les modèles GLM apparaissent moins bien ajustés mais cela semble s'expliquer par les données de l'échantillon. La surtarification des modèles CART, Bagging et Random forest sur l'échantillon d'apprentissage n'est plus observée sur l'échantillon test. Le secteur d'activité n'intervient pas dans le tarif en vigueur, son ajout pourrait permettre d'affiner le tarif.

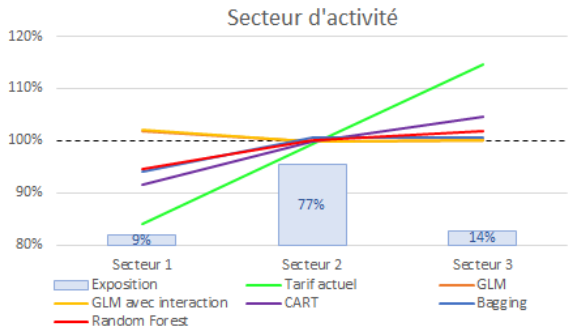


FIGURE 4.14 – S/P net sur l'échantillon d'apprentissage

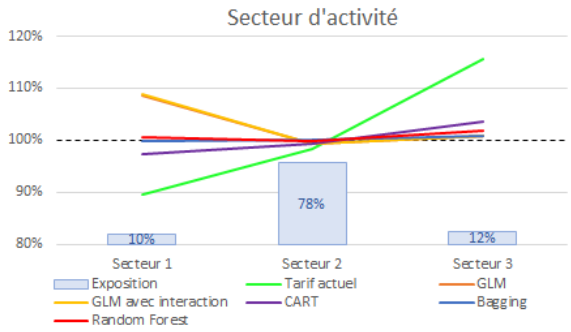


FIGURE 4.15 – S/P net sur l'échantillon test

4.1.3.7 Régime

Deux groupes de modèles se distinguent, un premier groupe avec les deux modèles GLM et un second groupe avec les autres modèles. Les modèles CART, Bagging et Random Forest ont tendance à surestimer la prime pure sur le régime Alsace-Moselle. Les deux graphiques révèlent une forte sous-tarification du régime Alsace-Moselle sur le tarif en vigueur.

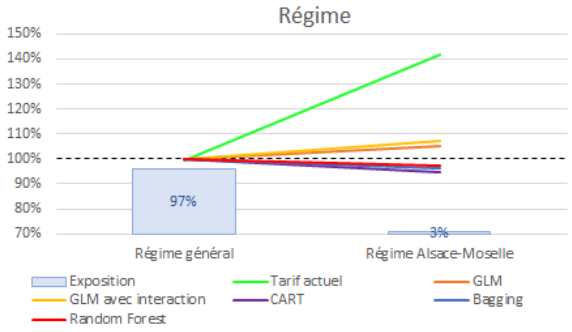


FIGURE 4.16 – S/P net sur l'échantillon d'apprentissage

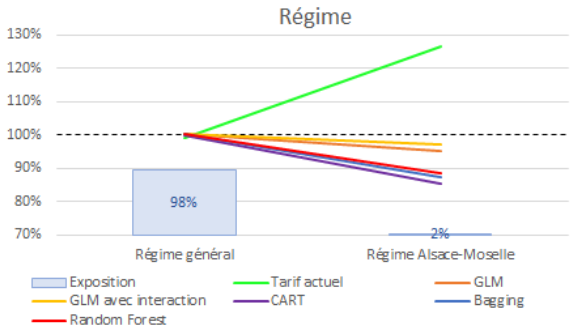


FIGURE 4.17 – S/P net sur l'échantillon test

4.1.3.8 Zonier

Pour chacun des postes, un zonier a été défini mais seul le zonier soins courants fera l'objet d'une analyse dans ce paragraphe, le même type d'analyse pouvant être réalisé sur les autres postes. De manière générale, les ajustements sur les échantillons d'apprentissage et de test sont très bons pour l'ensemble des modèles. Les S/P nets par zone sont compris entre 95% et 105%.

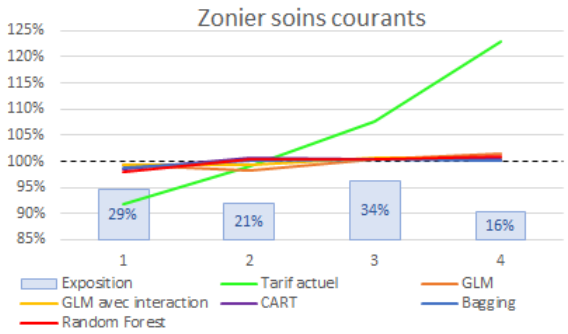


FIGURE 4.18 – S/P net sur l'échantillon d'apprentissage

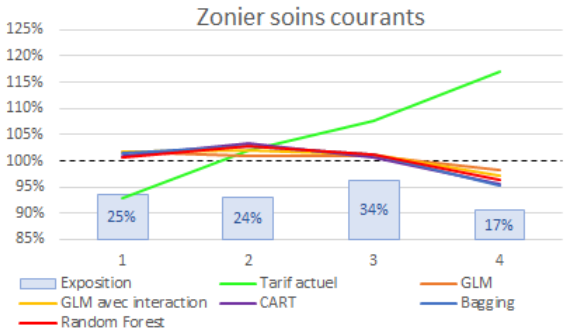


FIGURE 4.19 – S/P net sur l'échantillon test

En revanche, une analyse croisée du zonier et du niveau de garantie met en évidence les faiblesses du modèle GLM sans interaction. Les deux graphiques ci-dessous montrent la qualité d'ajustement par niveau en considérant les zones 1 et 2 de plus faible consommation médicale et les zones 3 et 4 avec une plus forte consommation.

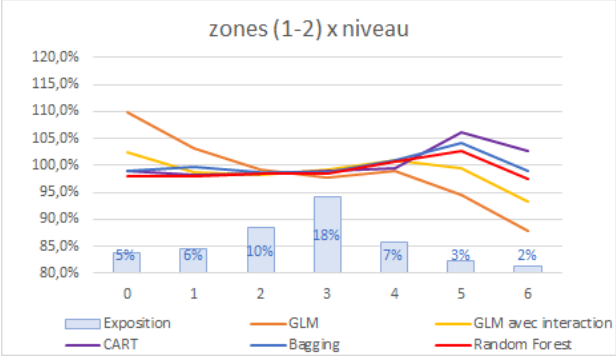


FIGURE 4.20 – S/P net sur l'échantillon test sur les zones 1 et 2

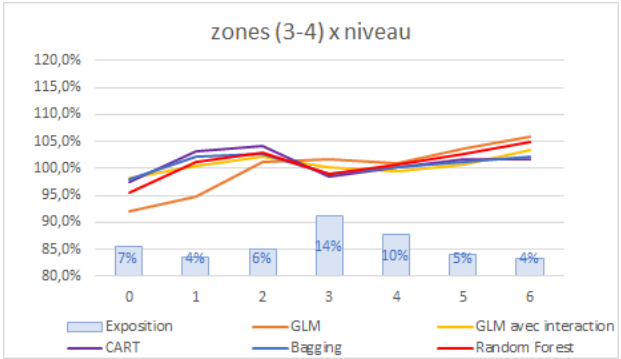


FIGURE 4.21 – S/P net sur l'échantillon test sur les zones 3 et 4

Sur le premier graphique (4.20) avec les zones 1 et 2, le modèle GLM sans interaction sous-tarifie les niveaux d'entrée de gamme et surtarifie les niveaux haut de gamme. Sur le second graphique, figure 4.21, la tendance inverse est observée. Ce phénomène avait été expliqué dans le cadre de la modélisation GLM du coût moyen avec interaction. La prime pure sur les niveaux d'entrée de gamme est moins sensible à l'effet géographique car le niveau de remboursement de la complémentaire santé atteint souvent le plafond de la garantie. Le modèle GLM avec interaction permet de corriger ce biais.

4.1.4 Synthèse des résultats

La performance des différents modèles est relativement proche sur l'échantillon test, les écarts observés sur le RMSE et le MAE entre les modèles sont inférieurs à 2%. Avec le RMSE, c'est le modèle BAGGING et une approche prime pure qui apporte la meilleure performance sur les données du portefeuille étudié. Avec le MAE, c'est le modèle Random Forest qui est le plus performant. De manière générale, les trois méthodes d'apprentissage permettent un léger gain de performance par rapport à la modélisation GLM.

Les analyses graphiques des S/P net par variable sont un peu perturbées par une certaine volatilité des résultats entre l'échantillon d'apprentissage et l'échantillon test. Ces modèles demanderaient à être testés sur un échantillon de taille plus importante, via un backtesting par exemple. Cependant, il est tout de même possible de tirer certaines conclusions lorsque les constats concordent entre les deux échantillons. Ces analyses montrent que le modèle BAGGING est globalement le modèle qui est le mieux ajusté sur les différents critères tarifaires. Les modèles GLM présentent une qualité d'ajustement légèrement meilleure sur les variables âge moyen des salariés et secteur d'activité mais ils apparaissent moins biens ajustés sur la variable lien de parenté que les autres modèles. Le caractère multiplicatif du modèle GLM se traduit par l'application des mêmes coefficients tarifaires sur des population différentes, or les effets des variables ne sont pas les mêmes entre un salarié, un conjoint ou un enfant. Dans le modèle GLM sans interaction, l'application du facteur multiplicatif de l'âge moyen des salariés de l'entreprise pour déterminer la prime pure des enfants de salariés n'est pas très pertinent.

Par ailleurs, les modèles non paramétriques permettent de mieux prendre en compte les interactions entre les variables. Par exemple, les effets du zonier ne sont pas uniformes sur l'ensemble des niveaux de garantie, l'effet de la zone est plus important sur les garanties haute gamme que sur les garanties d'entrée de gamme. L'analyse de l'ajustement par niveau en considérant uniquement les zones 1 et 2 puis les zones 3 et 4 a mis en évidence des problèmes d'ajustements des modèles GLM sur les niveaux d'entrée ou de haute gamme.

Les méthodes d'agrégation des modèles BAGGING et Random Forest permettent d'améliorer la qualité d'ajustement. Ils apparaissent souvent mieux ajustés que le modèle CART.

En ce qui concerne le tarif en vigueur, différents écarts d'ajustement ont été mis en évidence par cette étude. Tout d'abord, la répartition de la prime pure par poste montre une surestimation de la prime pure sur le poste dentaire et une sous-estimation sur les postes soins courants et bien-être. Ensuite, des écarts d'ajustement ont été identifiés sur certaines variables. Une surestimation de la prime pure est constatée sur le niveau de garantie 1 et sur les entreprises avec un seul salarié. Par ailleurs, une sous-estimation sur les entreprises de plus de 30 salariés est observée. Concernant l'âge moyen et le lien de parenté, des écarts d'ajustement sont constatés, ils s'expliquent par les critères tarifaires actuelles qui comportent une segmentation moins fines que les découpages effectués dans cette étude. Le tarif en vigueur ne fait pas de distinction entre le salarié et le conjoint, or le conjoint consomme davantage. Concernant les classes d'âges, une segmentation plus fine des moins de 35 ans et des plus de 45 ans dans cette étude expliquent les écarts. Sur le régime, une sous-tarification du régime Alsace-Moselle assez importante est apparue. Sur le zonier, une sous-estimation du tarif sur le poste soins courants est observée sur les zones 3 et 4 mais le tarif en vigueur s'appuie

sur un zonier identique pour tous les postes de soins avec des coefficients tarifaires de zones identiques. L'effet zone est donc identique quelque soit le poste de soins or il est plus marqué sur le poste soins courants que sur les autres postes. Cela peut expliquer ce constat ainsi que le déséquilibre par poste. Le tarif pourrait aussi être affiné avec l'intégration de la classification des activités définie dans cette étude. Enfin, le tarif en vigueur est construit sur un tarif multiplicatif comme le modèle GLM sans interaction testé dans cette étude. Il présente donc les faiblesses de la méthode telles que l'application identique des coefficients tarifaires sur les enfants de salariés et les salariés, la non prise en compte de l'interaction entre le niveau de garantie et la zone géographique.

4.1.5 Avantages et inconvénients des modèles

Les méthodes d'apprentissage CART et Random Forest présentent l'avantage de ne pas imposer de condition sur la loi de distribution de la variable à expliquer. Elles permettent également de prendre en compte naturellement les interactions entre les variables explicatives contrairement aux modèles GLM où il est nécessaire de les préciser.

Néanmoins, la méthode CART peut présenter un caractère instable. Une faible variation dans l'échantillon des données peut conduire à l'élaboration d'un arbre totalement différent. Chaque étape de découpage de l'arbre conditionne les découpages suivants. Une modification de l'ordre d'intervention des variables dans le processus d'élaboration de l'arbre modifie la suite du processus de construction de l'arbre. Les modèles Random Forest offrent une solution à ce problème d'instabilité en associant une stratégie d'agrégation de modèles et une dimension aléatoire dans l'élaboration de chacun des modèles.

La performance n'est pas le seul critère qui guide le choix de l'utilisation d'une méthode plutôt qu'une autre. La facilité d'implémentation, l'explicabilité du tarif sont aussi des éléments de décision. Les modèles GLM et CART restent relativement lisibles et apportent cette facilité de lecture du tarif. Sur cet aspect, l'avantage est peut être au modèle GLM avec les valeurs des différents coefficients associés aux modalités des variables tarifaires. Avec la méthode CART, une importante complexité de l'arbre pourra rendre difficile l'analyse. En revanche l'explicabilité du tarif est souvent le point faible des méthodes d'agrégation qualifiées parfois de boîte noire. Des algorithmes sont développés afin d'apporter cette explicabilité. Ces modèles plus complexes peuvent présenter un risque opérationnel plus important. L'implémentation du tarif est également plus complexe avec les algorithmes d'agrégation. Des contraintes opérationnelles, informatiques notamment peuvent freiner leur mise en œuvre.

Le tarif doit aussi respecter une certaine cohérence. Quelque soit le profil de l'entreprise qui souscrit, des garanties plus élevées doivent être associées à un tarif plus élevé toute chose étant égale par ailleurs. Ce besoin de cohérence implique la mise en place de lissage et ajustement de la prime pure. Cette étape essentielle dans l'élaboration du tarif s'avère plus complexe à mettre en œuvre avec des méthodes telles que le Bagging et Random Forest. Le tarif final n'est pas simplement l'addition d'une prime pure et de frais de chargement, il intègre également certains ajustements propres à la stratégie commerciale de l'assureur. Ces ajustements s'avéreront d'autant plus complexes à mettre en œuvre si le tarif construit repose lui même sur une forte complexité.

4.2 Impact du 100% santé

Les aspects de la réforme du 100% santé ont été présenté dans le premier chapitre. Cette réforme majeure est mise en œuvre progressivement jusqu'en 2022. Elle touche trois postes de soins sur les six postes modélisés précédemment, à savoir le poste optique, le poste dentaire et le poste soins courants qui intègre les aides auditives. L'objectif premier de cette réforme est de réduire le renoncement aux soins en proposant certains appareillages et prothèses sans reste à charge. Son impact n'est donc pas neutre pour les complémentaires santé.

Le comportement de consommation des assurés pourra être sensiblement modifié. Plus précisément, cette réforme pourra impacter à la fois la fréquence de consommation de certains actes en raison de la baisse du renoncement au soins mais également leur coût moyen du fait de leur prise en charge intégrale par le régime obligatoire et la complémentaire santé.

Cette section sera donc consacré à l'impact de cette réforme sur la primes pure. Sur les postes de soins optique et dentaire, l'essentiel des mesures de cette réforme ont été mises en œuvre au 1^{er} janvier 2020. Il est donc possible de mesurer les premières tendances avec une certaine prudence compte tenu du contexte de crise sanitaire qui bouleverse le comportement des assurés. Concernant l'audio-prothèse, le panier 100% santé sera mis en place à partir du 1^{er} janvier 2021. Il n'est donc pas possible d'observer son impact mais il devrait être plutôt réduit sur la santé collective puis que les assurés sont relativement jeunes. Le poids des prestations en audio-prothèse est inférieur à 0,5% de la charge de sinistres totale.

A partir de la sinistralité observée sur les dix premiers mois de l'année 2020, ce chapitre tentera de fournir une mesure des premiers effets de cette réforme en s'appuyant sur une modélisation GLM. Les impacts seront évalués au travers de la déformation des coefficients tarifaires des modèles GLM entre deux périodes, une période de référence 2017-2019 et une période 2020 avec la réforme.

4.2.1 Méthode de mesure d'impact

La mesure des effets de la réforme sur les postes optique et dentaire va être réalisée à l'aide d'une modélisation GLM afin d'identifier les segments tarifaires les plus impactés. Cette modélisation sera réalisée selon le même procédé que précédemment (chapitre 7) avec la modélisation de la prime pure sans interaction. Les lois utilisées seront donc la loi binomiale négative pour la fréquence et la loi gamma pour le coût moyen. Deux échantillons de données seront considérés, un premier échantillon, échantillon sans la réforme, correspondant aux prestations survenues et réglées sur les dix premiers mois des années 2017 à 2019 et un second échantillon, échantillon avec la réforme, composé des prestations survenues et réglées sur les dix premiers mois de l'année 2020.

L'impact de la réforme sera mesuré au travers de la déformation des coefficients des modèles GLM entre les deux périodes. Deux types de modélisations vont être réalisées sur la période 2020.

Une première modélisation va être effectuée en utilisant les mêmes variables explicatives entre les deux périodes afin de mesurer la déformation des coefficients de chacune d'elles.

Une seconde modélisation va consister à mesurer les effets de déformation uniquement sur le niveau de garantie. Les déformations des coefficients associés aux autres variables seront neutralisées. Cette neutralisation consistera à appliquer sur la période 2020 les valeurs des coefficients obtenue sur la période 2017-2019 pour l'ensemble des variables, excepté le niveau de garantie, sous la forme d'une variable offset.

Application d'une correctif de l'effet COVID-19 sur la fréquence

La crise COVID-19 a bouleversé les comportements des assurés en 2020. Une forte baisse de la fréquence a été observée sur les postes optique et dentaire en raison notamment du confinement du printemps. Afin de pouvoir comparer la valeur des coefficients entre les deux périodes, un redressement des coefficients sera appliqué.

Ce coefficient de redressement a été déterminé différemment entre l'optique et le dentaire. Sur le dentaire, il existe un grand nombre d'actes de soins contrairement à l'optique où les actes concernent essentiellement une paire de lunettes. Cette diversité d'actes sur le dentaire permet de déterminer l'effet de la crise sanitaire sur les différents niveaux plus précisément en excluant par exemple les actes concernés par la réforme.

Sur l'optique, le redressement a été déterminé à partir de la baisse de fréquence observée sur les niveaux moyens et hautes gammes entre les deux périodes. La sélection de ces niveaux pour déterminer le coefficient s'appuie sur deux hypothèses sous-jacentes fortes. La première hypothèse consiste à considérer que le changement de comportement sur ces niveaux est uniquement lié à la crise sanitaire, la réforme ne générant pas de modification de la fréquence pour ce segment d'assurés assurés étant donné qu'ils bénéficient d'une couverture de soins importante. La seconde hypothèse est de considérer que les changements de comportement observés sur les niveaux moyen et haute gamme liés à la crise sanitaire sont identiques sur les niveaux d'entrée de gamme.

Sur le dentaire, le même principe que sur l'optique a été utilisé pour les garanties moyennes et hautes gammes. En revanche, une méthode spécifique a été utilisée sur les niveaux d'entrée de gamme potentiellement impactés par une augmentation de la fréquence sur les prothèses dentaires avec la réforme. Le coefficient de redressement a été déterminé en comparant la fréquence sur les actes non concernés par la réforme entre les deux périodes, c'est dire tous les actes excepté les prothèses dentaires. Sur ces actes, la baisse de la fréquence ne peut pas être imputée à la réforme. Cette estimation s'appuie sur l'hypothèse selon laquelle l'évolution de la fréquence sur ces actes est uniquement due à la crise sanitaire. Par ailleurs, depuis le 1^{er} janvier, les prothèses dentaires provisoires font l'objet d'une prise en charge. Cette modification génère une augmentation de la fréquence qui a été neutralisée pour déterminer les coefficients de redressement du dentaire.

4.2.2 Quelques chiffres publiés

La réforme du 100% santé a fait l'objet de différentes publications sur ses impacts depuis son lancement. Début novembre 2020, le courtier Henner (Henner, 2020) a communiqué sur une étude et indique un succès pour le dentaire et un échec sur l'optique.

Voici quelques extraits de l'article publié par newsassurancespro le 3 novembre 2020 :

« En assurance collective, 55% des couronnes dentaires sont dans les paniers encadrés, alors qu'en optique uniquement 1,5% des verres ou des montures correspondent aux équipements du 100% santé. »

« Sur les contrats collectifs, 37% des couronnes remboursées correspondent au panier 100% santé, 17,8% au panier maîtrisé et 44,6% au panier libre. Il existe de vraies différences entre la région parisienne et la province. Ainsi, le taux de recours au 100% santé sur les couronnes dentaires est faible en Île-de-France (25%) par rapport aux Pays de la Loire (53%), la Bretagne (51%), la Bourgogne (50%) ou Les Hauts-de-France (49%). »

« Par contre, il est difficile de mesurer une éventuelle hausse de la fréquence étant donné la baisse de consommation due au confinement. »

« La baisse du plafonnement de la prise en charge de la monture par l'organisme complémentaire (de 150 euros à 100 euros) a produit ses effets. Ainsi, le remboursement moyen de la monture par la complémentaire est de 103 euros, soit 29% de moins qu'en 2019, alors que le reste à charge moyen sur la monture s'élève à 29%, en hausse de 81% par rapport à 2019. »

Ces chiffres permettent d'avoir un premier niveau d'information sur les impacts de la réforme. Une analyse sur le portefeuille étudié sera réalisée dans les sections suivantes.

4.2.3 L'impact sur le poste optique

La réforme optique comporte la mise en place d'un panier 100% santé et un abaissement du plafond de remboursement de la monture de 150 euros à 100 euros sur le contrat responsable. Sur le portefeuille étudié, très peu d'assurés ont opté pour le panier 100% santé, seulement 2,6% des actes verres et montures ont été réalisés sur ce panier. Cependant ces actes sont concentrés sur les niveaux de garantie d'entrée de gamme. Sur le niveau 0, 10% des actes verres et montures sont des actes 100% santé tandis que sur les niveaux moyen et haute gamme, cela représente seulement 1% des actes.

Par ailleurs, l'abaissement du plafond de remboursement de la monture de 150 euros à 100 euros sur le contrat responsable a peut être modifier les comportements sur les garanties très élevées. La garantie étant formulée sous la forme d'un niveau de remboursement pour un équipement optique complet avec une monture et deux verres, cette modification aura un impact limité sur les garanties basses et moyennes. Sur ces garanties, la baisse du remboursement de la monture se traduira par une meilleure prise en charge des verres par la complémentaire.

4.2.3.1 L'impact sur la fréquence

Les résultats de la modélisation GLM indique principalement une déformation des coefficients liés à l'âge moyen des salariés, la zone et du niveau de garantie.

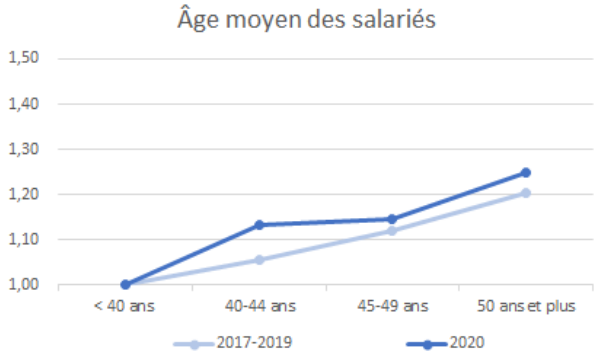


FIGURE 4.22 – Coefficients GLM de l'âge moyen des salariés

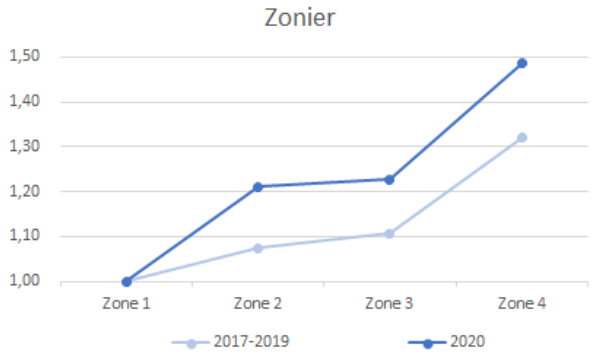


FIGURE 4.23 – Coefficients GLM du zonier

L'intensité de l'effet âge moyen des salariés est renforcé sur les plus de 40 ans et notamment sur les 40-44 ans. Sur les 40-44 ans, la fréquence à moins diminué que sur les autres segments entre les deux périodes. Il est difficile d'imputer cette évolution uniquement à la réforme. La crise COVID 19 a généré une baisse globale de la fréquence mais celle-ci n'a peut être pas été uniforme sur tous les segments. Par exemple, la fréquence est restée stable chez les salariés de 44-45 ans entre les deux périodes observées, vraisemblablement en raison d'une part plus importante de besoin d'équipement optique non reportable, la presbytie peut expliquer ce besoin. Concernant le zonier, la déformation générale traduit une augmentation de la différence de risque entre la zone 1 et les zones 2, 3 et 4. La forme de la courbe sur les zones 2, 3 et 4 est relativement similaire entre les deux périodes à un paramètre d'échelle près.

Sur la figure 4.24 ci-dessous, les coefficients des niveaux de garantie apparaissent sans surprise tous inférieurs sur 2020 du fait de la crise sanitaire.

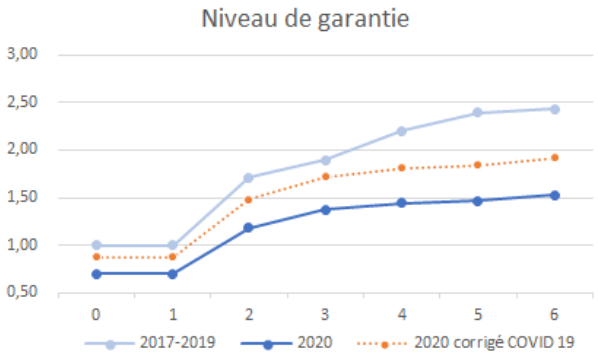


FIGURE 4.24 – Coefficients GLM du niveau de garantie sans neutralisation des autres effets

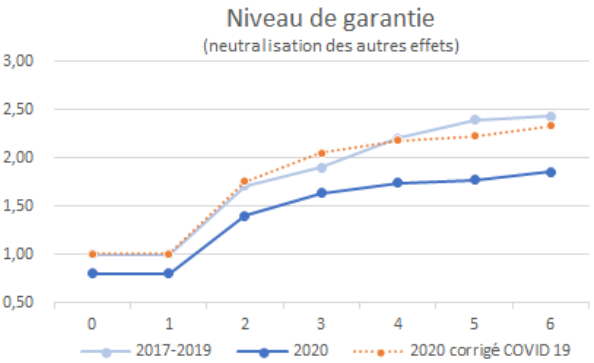


FIGURE 4.25 – Coefficients GLM du niveau avec neutralisation des autres effets

La courbe en orange est la courbe 2020 observée avec l'application du correctif de l'effet de la crise COVID 19 défini à la section 4.2.1. Avec ce correctif, il est à noté une baisse de

l'ensemble des coefficients associés au niveaux de garantie. Ce phénomène doit s'expliquer par l'effet des coefficients du zonier sur les zones 2,3 et 4 qui ont significativement augmenté.

Sur le graphique 4.25 correspondant à la modélisation de la fréquence en neutralisant tous les effets hors effet garantie et en appliquant le correctif global de la crise COVID 19, les courbes des deux périodes apparaissent relativement proches. Une légère tendance à la baisse est observée sur les niveau 5 et 6. Les effets de déformations observés précédemment sur l'âge moyen des salariés, le zonier et le niveau de garantie semblent se compenser.

L'impact de la réforme du 100% santé semble relativement limité sur la fréquence en optique à ce stade, il n'y a pas eu de changement notable de comportement sur les formules d'entrée de gamme. Les actes 100% santé réalisés auraient vraisemblablement été réalisés sans la réforme.

4.2.3.2 L'impact sur le coût moyen

Les résultats de la modélisation du coût moyen montrent de légères déformations des coefficients tarifaires liés à l'âge moyen des salariés et la zone entre les deux périodes 2017-2019 et 2020.

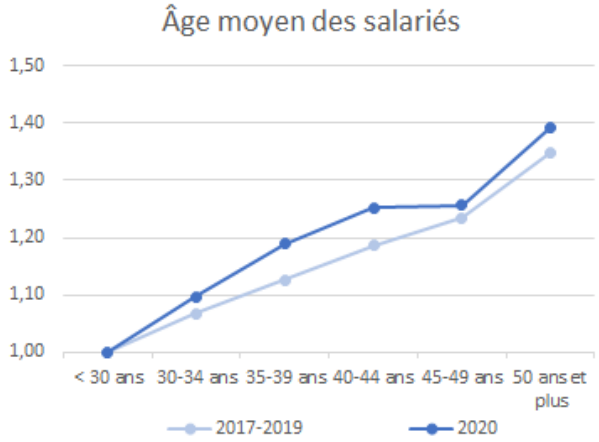


FIGURE 4.26 – Coefficients de l'âge moyen des salariés du modèle GLM de la fréquence optique

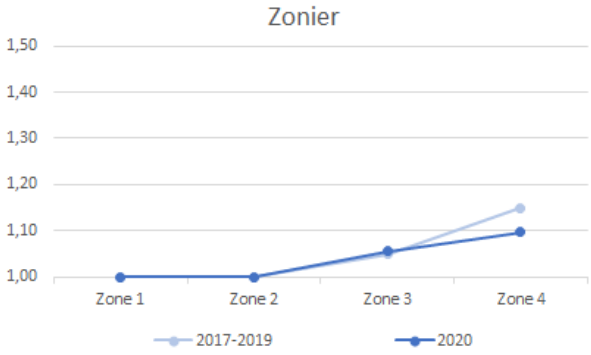


FIGURE 4.27 – Coefficients du zonier du modèle GLM de la fréquence optique

En ce qui concerne l'âge moyen des salariés, la déformation est une déformation à la hausse avec le segment des 40-44 ans qui est le plus impacté. La déformation des coefficients de zone reste modérée, seul le coefficient de la zone 4 est concerné avec une légère baisse.

Sur la figure 4.28, les coefficients associés au niveau de garantie apparaissent globalement stables entre les deux périodes. La courbe en pointillés des coefficients du niveau de garantie avec neutralisation des autres effets est superposée à celle sans neutralisation jusqu'au niveau 4.

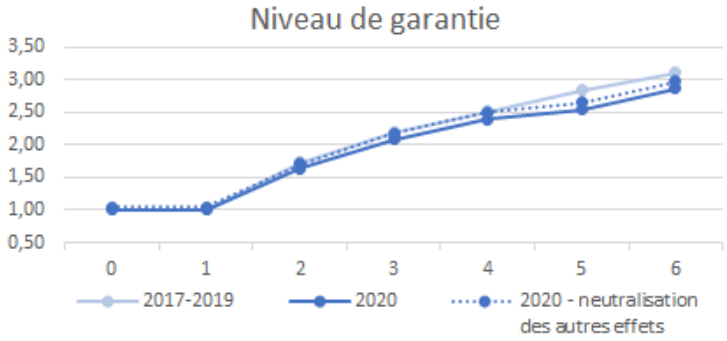


FIGURE 4.28 – Coefficients GLM du niveau de garantie

Le coût moyen sur les niveaux d'entrée de gamme est resté stable malgré une part non négligeable des actes réalisés sur le panier 100% santé. Ceci s'explique par un différentiel de coût faible entre un équipement 100% santé et un équipement panier libre sur les garanties d'entrée gamme. En santé collective, le niveau de garantie minimum défini par le panier ANI est de 100 euros pour un équipement simple et 200 euros pour un équipement complexe, or les coûts moyens pour la complémentaire santé des équipements 100% santé simples et complexes sont légèrement inférieurs à ces minimas de garantie.

Sur les niveaux 5 et 6, la légère baisse du coût moyen s'explique par la baisse du remboursement de la monture. Une partie des prestations qui servaient à rembourser la monture s'est reportée sur le remboursement des verres mais pas l'intégralité.

4.2.3.3 L'impact en dentaire

La réforme du dentaire concerne la revalorisation des soins conservateurs et une offre de prothèses dentaires sans reste à charge. Les mesures de la réforme sont échelonnées sur 2020 et 2021. La première étape sur 2020 concerne la mise en place du panier 100% santé sur les prothèses fixes, couronnes et bridges, et une deuxième étape en 2021 concernera les prothèses amovibles ou dentiers. Compte tenu de son étalement, l'impact de la réforme mesuré à partir des données 2020 ne permet pas d'avoir une vision exhaustive des impacts de la réforme mais permet permet d'établir les grandes tendances.

Sur le portefeuille étudié, la part des couronnes dentaires est de 39,5% sur le panier 100% santé, 24,5% sur le panier maîtrisé et 36% sur le panier libre avec une forte hétérogénéité selon le niveau de garantie. Sur les garanties d'entrée de gamme, la part sur le panier 100% santé est comprise entre 50% et 60%. Ce succès de la réforme sur le dentaire va donc modifier très fortement la prime pure sur le poste dentaire.

Les prothèses provisoires sont également prise en charge par la sécurité sociale depuis le 1^{er} janvier 2020, cela impacte à la hausse la fréquence du poste dentaire.

4.2.3.4 L'impact sur la fréquence

Hormis pour le niveau de garantie, les coefficients sont relativement stables entre les deux périodes. Les deux courbes des coefficients associés à l'âge moyen des salariés sur la figure 4.29 se confondent. Concernant le zonier, les effets de zones se sont légèrement accentués.

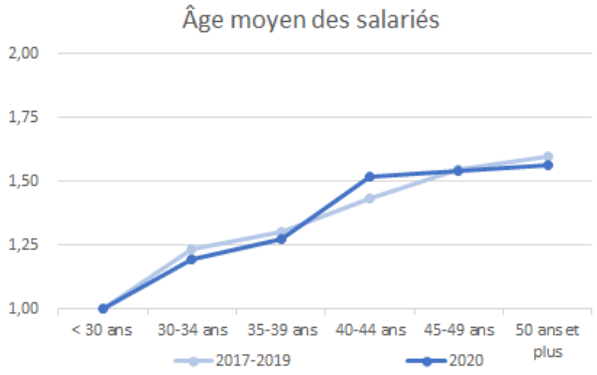


FIGURE 4.29 – Coefficients GLM de l'âge moyen des salariés

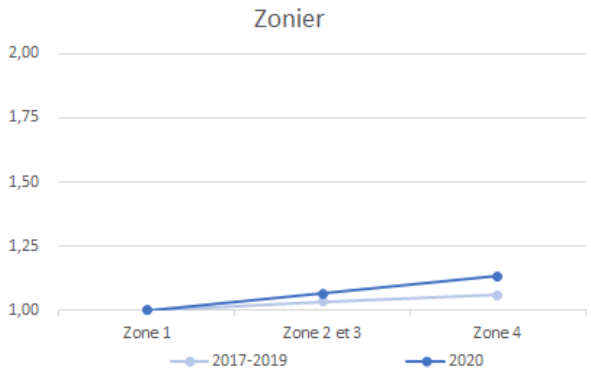


FIGURE 4.30 – Coefficients GLM du zonier

En ce qui concerne le niveau de garantie, la forme générale des courbes des coefficients est assez proche entre les deux périodes mais une augmentation de la fréquence se dégage tout de même sur les niveaux de garantie inférieurs au niveau 3.

La courbe en orange sur le graphique 4.32 correspond à l'exercice de survenance 2020 avec l'application d'un correctif de l'effet de la crise COVID 19 et la neutralisation des autre effets. Cette courbe indique une hausse de la fréquence de l'ordre de 7 à 8% sur les niveaux d'entrée de gamme. Cette hausse s'explique par la prise en charge des couronnes provisoires, environ 45% de la hausse et le reste est lié à l'augmentation de la fréquence sur les couronnes dentaires, les inlays-cores et les Inlays-onlays. Sur les niveaux supérieurs au niveau 2, la hausse est liée aux couronnes provisoires.

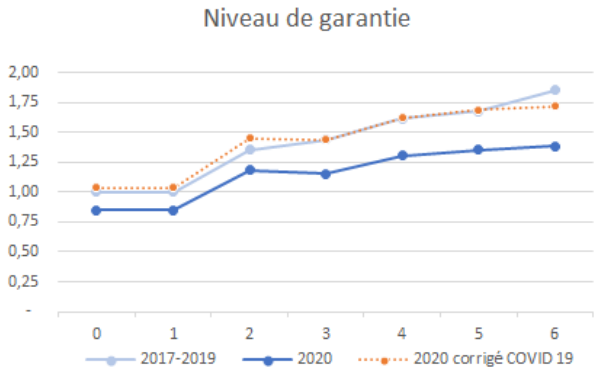


FIGURE 4.31 – Coefficients GLM du niveau de garantie

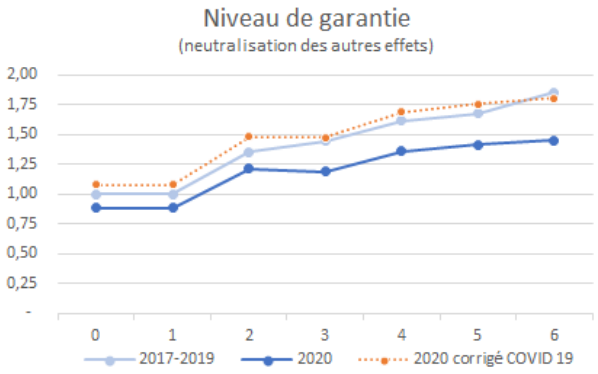


FIGURE 4.32 – Coefficients GLM du niveau de garantie avec neutralisation des autres effets

4.2.3.5 L'impact sur le coût moyen

Les coefficients tarifaires de l'âge moyen des salariés ont légèrement évolué sur 2020. Une baisse du coût moyen sur le segment des 50 ans et plus est observée. Sur le zonier la déformation est plus marquée, l'écart relatif entre les zones s'est réduit, il n'y a plus d'écart entre les zones 3 et 4.

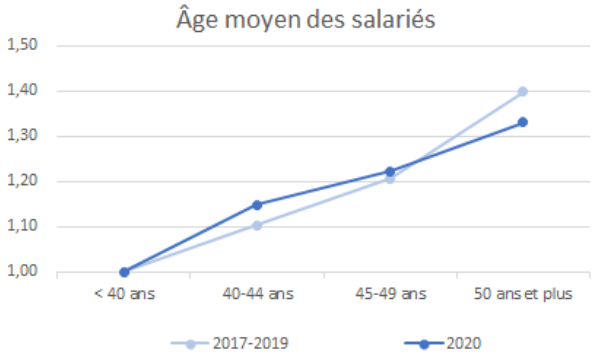


FIGURE 4.33 – Coefficients GLM de l'âge moyen des salariés

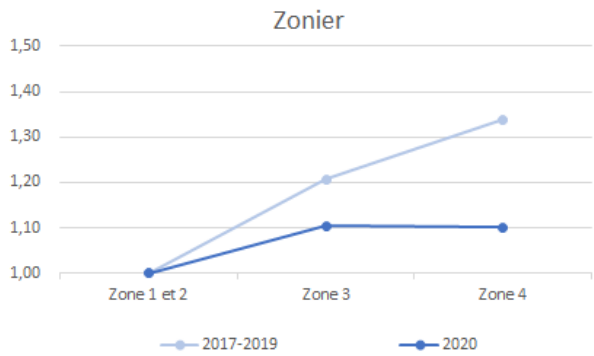


FIGURE 4.34 – Coefficients GLM du zonier dentaire

La déformation des coefficients du coût moyen sur le niveau de garantie est sans surprise très marquée. La courbe en pointillés avec la neutralisation des autres effets indique une hausse du coût moyen sur les niveaux 0 à 3 et une baisse sur les niveaux 5 et 6.

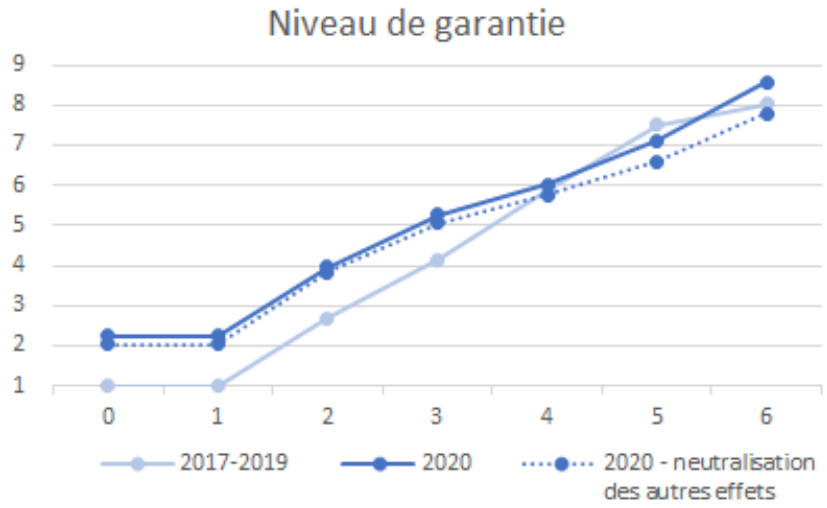


FIGURE 4.35 – S/P net sur l'échantillon d'apprentissage

La hausse est particulièrement forte sur les deux premiers niveaux avec un doublement du coût moyen. Cette évolution s'explique par le mix des types d'actes (soins dentaires, couronnes dentaires,...) qui s'est déformé et la hausse du coût de certains actes, notamment les actes prothétiques réalisés sur le panier 100% santé. Le premier effet génère une hausse de 15% du coût moyen et le second une hausse de 80% du coût moyen, le produit des deux effets générant un doublement du coût moyen. L'effet mix s'explique par une part plus importante

d'actes plus chers notamment les couronnes dentaires. Cette effet doit être considérer avec prudence car la baisse de fréquence sur ces actes a surement été hétérogène selon le type d'acte. Si l'on considère que la fréquence des actes de prothèses n'a pas été impactée par le confinement, l'effet mix diminue à 5%. La hausse du coût moyen des actes est à attribuer pleinement à la réforme et principalement au couronnes dentaires. Le coût moyen d'une couronne passe de 60 euros à plus de 300 euros dans le cadre du panier 100% santé.

Sur les niveaux 5 et 6, la baisse du coût moyen est de l'ordre de 10%, elle s'explique essentiellement par la baisse du coût des inlays-cores. Sur ces niveaux, une partie des inlays-core ont été réalisé sur le panier 100% santé, or le coût moyen des inlays-cores est deux fois inférieur à ceux réalisés sur le panier libre.

4.2.4 Synthèse des impacts de la réforme

Les impacts de la réforme sont donc relativement hétérogènes entre les différents postes.

Sur le poste optique, la réforme ne semble pas produire d'effets notables en terme d'évolution de la prime pure. Les formules d'entrée de gamme concentrent une large partie des actes du panier 100% santé mais cela est sans effet sur la fréquence et le coût moyen. Ceci s'explique par le coût de l'équipement 100% santé proche des garanties minimales imposées par le panier ANI en optique. Sur les garanties haute gamme, une légère baisse du coût moyen est à noter en raison de la baisse du remboursement de la monture.

Sur le dentaire, le succès de la réforme ne se dément pas. Il se traduit par une hausse de la fréquence sur les garanties d'entrée de gamme de 7 à 8% en raison de la prise en charge des couronnes provisoires et la hausse de la fréquence sur les couronnes dentaires, les inlays-cores et les Inlays-onlays. Une large proportion de ces actes prothétiques ont été réalisés sur le panier 100% santé, ce qui se traduit par un doublement du coût moyen global sur les niveaux 0 et 1. Sur les niveaux supérieurs au niveau 2, une hausse de la fréquence est constatée en raison de la prise en charge des couronnes provisoires. Concernant le coût moyen, une baisse de l'ordre de 10% est observée sur les niveaux hautes gammes. Cette baisse est liée aux actes inlays-cores qui ont été réalisés pour une partie d'entre-eux sur le panier 100% santé pour lequel le coût moyen est deux fois plus faible que sur le panier libre.

Cette analyse a pour objectif de donner une vision des premières tendances observées sur les impacts de la réforme. Compte tenu de la crise sanitaire, de l'échelonnement de la réforme et de la période d'observation des effets de la réforme relativement courte, ces premiers impacts sont susceptibles d'évoluer.

Conclusion

La science actuarielle évolue et il est nécessaire de remettre régulièrement en cause les approches utilisées afin de s'assurer qu'elles restent performantes. La forte concurrence sur le marché de la santé collective, les évolutions réglementaires assez fréquentes et les changements de comportement des assurés qui peuvent en découler l'imposent également.

Les modèles linéaires généralisés sont fréquemment utilisés pour les besoins de tarification mais cette étude a mis en évidence certaines de leurs limites bien connues. Tout d'abord l'adéquation des données aux lois théoriques est apparue difficile à satisfaire notamment sur la fréquence avec la surreprésentation des non consommants. Ensuite la présence d'interaction entre les variables explicatives, notamment entre le niveau de garantie et le zonier conduit à créer certains biais d'estimation de la prime pure avec cette méthode. Les interactions peuvent être spécifiées dans la modélisation GLM mais cela suppose un travail d'identification important en amont. Par ailleurs, l'intégration de multiples interactions dans une modélisation GLM complexifie grandement le tarif et altère d'autant plus sa lisibilité, atout majeur de cette méthode.

Les méthodes d'apprentissage CART et Random forest au travers de cette étude ont montré qu'elles pouvaient être une alternative intéressante à la méthode des modèles linéaires généralisés. Les gains de performance ont été relativement modestes mais les biais identifiés sur certains croisements de variables avec la méthode GLM n'étaient pas présents. Elle permettent de prendre en compte naturellement les interactions entre les variables.

L'algorithme CART est une méthode relativement facile à mettre en œuvre. Elle nécessite tout de même d'appliquer certaines précautions d'usage notamment avec l'utilisation des variables explicatives numériques. Le nombre de valeurs distincts qu'elles comportent sont autant de possibilités de découpage lors du processus de division, ce qui peut générer du surapprentissage. La solution consiste à les transformer en classe tout en veillant à préserver une certaine cohérence quand cela est nécessaire. Un critère de division comportant une sélection de classes d'âge avec des discontinuités n'apparaîtrait pas très pertinent en tarification santé étant donné que la consommation médicale est monotone croissante avec l'âge. L'élagage de l'arbre est une des étapes clés de la méthode pour laquelle il n'existe pas de procédé universel. Deux règles sont fréquemment utilisées, la règle de l'écart-type mais elle a rapidement été exclue car elle ne permettait pas d'obtenir une finesse de tarif suffisante et la règle basée sur la minimisation de l'erreur. Cette dernière a été retenue mais elle a tendance à sélectionner des arbres relativement complexes. Le gain apporté par les dernières branches de l'arbre s'avère relativement faible. Un compromis a été recherché, il a été défini en analysant la courbe d'évolution de l'erreur en fonction de la complexité. Ce seuil a été fixé au seuil qui minimise l'erreur plus 2,5% de son écart type. Ce seuil permet de simplifier fortement

l'arbre. Sur l'arbre de la fréquence du poste soins courants, il a permis de réduire le nombre de feuilles de 35% pour une perte de performance inférieure à 1%. L'interprétabilité de la méthode constitue un atout, il est relativement aisé de s'assurer de la cohérence de la prime pure modélisée sur les différents feuilles.

Concernant l'application de la méthode Random Forest, les résultats ont été contrastés. La performance s'est avérée meilleure que celle atteinte avec la méthode CART mais la réalisation de test de sensibilité a mis en évidence que c'était la stratégie d'agrégation qui expliquait cette amélioration. En intégrant l'ensemble des variables, ce qui revient à un modèle de type BAGGING, la performance était tout aussi bonne voir meilleure que celle obtenue sur le modèle Random Forest. L'ajout de la dimension aléatoire sur la sélection du nombre de prédicateurs n'a pas permis d'améliorer significativement les résultats. Ceci s'explique vraisemblablement par le nombre de prédicateurs disponibles trop faible. Le fait qu'il n'y ai pas suffisamment de variables à fort pouvoir prédictif peut être aussi un autre élément d'explication.

Les méthodes d'apprentissage permettent une amélioration de la performance mais elles n'ont pas pleinement exprimé leur potentiel dans cette étude. La recherche de nouvelles informations tarifaires constitue un axe d'amélioration possible. Il existe un grand nombre de sources externes de données notamment les données publiques de l'Insee, les données du bilan de l'entreprise qui permettraient peut-être d'affiner encore davantage le tarif. Cependant, il faudra être vigilant à l'interprétabilité des effets liés à ces variables.

Par ailleurs l'analyse sur l'impact de la réforme 100% santé a montré que cette réforme avait des impacts forts en dentaire. Les formules d'entrée de gamme sont les plus touchées avec sur certains niveaux de garantie un doublement de la prime pure sur le poste dentaire. Sur l'optique, la réforme n'a pas eu d'effet en terme de fréquence et de coût moyen. Les actes réalisés sur le panier 100% santé auraient vraisemblablement été effectués sans la réforme. Ces actes ne génèrent pas non plus de dégradation de la sinistralité car leur coût est proche des garanties minimales imposées par le panier ANI. Sur les garanties hautes gammes, la réforme entraîne plutôt une baisse des prestations en raison de l'abaissement du plafond de remboursement de la monture. Ces constats illustrent le besoin d'ajuster régulièrement les tarifs aux évolutions réglementaires. Ces analyses devront être poursuivies afin d'affiner la mesure des effets de la réforme.

Bibliographie

DREES Santé - La complémentaire santé - Acteurs, bénéficiaires, garanties - Edition 2019
https://drees.solidarites-sante.gouv.fr/IMG/pdf/cs2019.pdf

DREES, Rapport 2020 - Sur la situation financière des organismes complémentaires assurant une couverture santé, 2020
https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-12/rapport-oc-2020.pdf

Ministères des solidarités et de la santé Dossier de presse « 100% santé - Des soins pour tous, 100% pris en charge », 2019
https://solidarites-sante.gouv.fr/IMG/pdf/dicom_dp_100_sante_2019__301219.pdf

Assurance maladie - Observatoire des pratiques tarifaires - Dépassements d'honoraires des médecins une tendance à la baisse qui se confirme, 2017
https://www.ameli.fr/fileadmin/user_upload/documents/Observatoire_des_pratiques_tarifaires.pdf

M. ARSALANE - Modélisation de la consommation médicale en assurance collective, Mémoire 2016

C. SEPULVEDA - Modélisation du risque géographique en santé, pour la création d'un nouveau zonier. Comparaison de deux méthodes de lissage spatial, Mémoire 2016

Ricco RAKOTOMALALA - Classification ascendante hiérarchique, Université Lumière Lyon-2

S. TUFFERY - Data mining et statistique décisionnelle - La science des données, 2017

F. PLANCHET, A. MISERAY - Tarification IARD Introduction aux techniques avancées, ISFA, 2017

A. CHARPENTIER - Actuariat IARD - ACT2040- Partie 4 - Modèles linéaires généralisés, UQAM, 2013

G. BOUCHTA - Mémoire Mise en œuvre de méthodes innovantes de tarification, Mémoire 2017

L.BREIMAN, J.FRIEDMAN, CJ.STONE, R.A.OLSHEN - Classification and Regression

Trees, 1984

C. MALOT-TULEAU - Méthodes CART Introduction à la sélection de variables, Cours de MASTER MASS Université Nice Sophia-Antipolis, 2006-2007

R. GENUER, J.M. POGGI - Arbres CART et Forêts aléatoires, Importance et sélection de variables
https://hal.archives-ouvertes.fr/hal-01387654v2/document

L.BREIMAN - Random forests are a combination of tree predictors, 2001

Newsassurancespro - 100% santé : Un succès en dentaire, un échec en optique - 03/11/2020https://www.newsassurancespro.com/100-sante-un-succes-en-dentaire-un-echec-en-optique/01691430320

Annexes

Annexe A - Compléments sur l'analyse bivarée

Le collège de salariés

La consommation médicale annuelle moyenne est plus importante pour la population des cadres que sur les populations non cadres et ensemble du personnel. Cette observation est à confirmer avec des analyses croisées car ce constat peut être le reflet d'un effet caché d'une autre variable telle que le niveau de garantie en moyenne plus élevé sur le collège cadre.

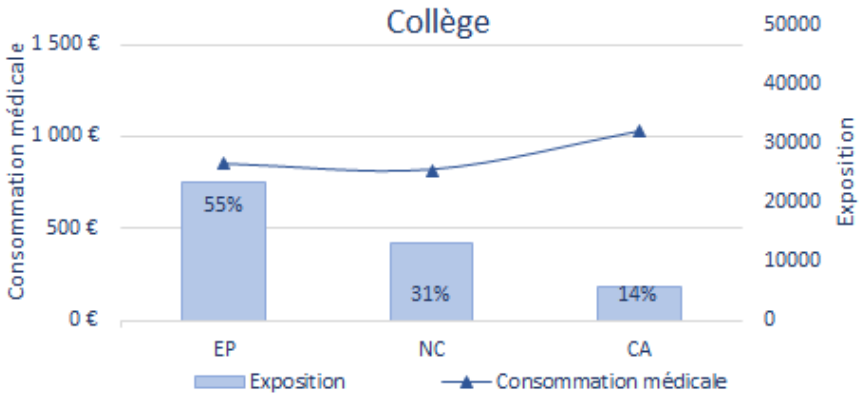


FIGURE 4.36 – Consommation médicale en fonction du collège de salariés

Zone géographique

La zone géographique est également un élément à prendre en compte dans la tarification. Le zonier actuel montre des écarts de consommation moyenne selon la zone. Une étude approfondie est réalisée à la section 5.2.

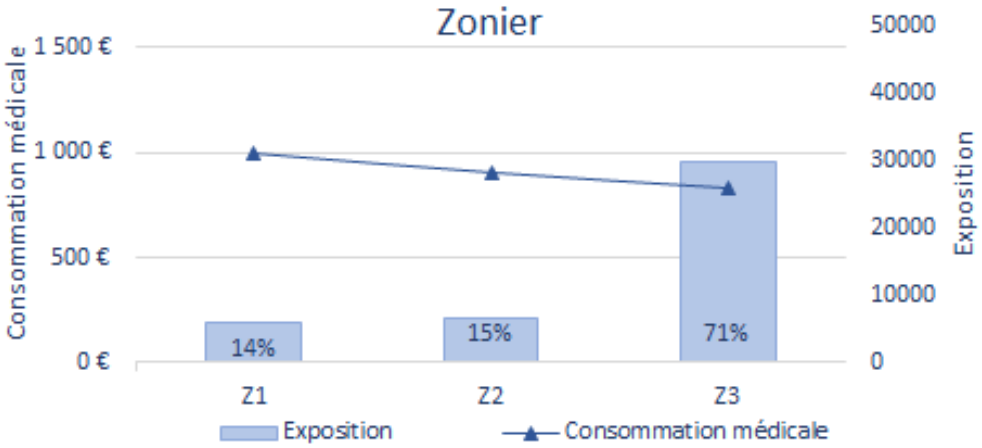


FIGURE 4.37 – Consommation médicale en fonction de la zone géographique

Le régime

Le nombre de bénéficiaire du régime Alsace-Moselle est faible dans le portefeuille avec 2% des bénéficiaires. Sur ce régime, la consommation médicale ne semble pas différente de celle observée sur le régime général.

Cependant le régime Alsace-Moselle assure un complément à la prise en charge par le régime général des prestations en nature (soins de ville, hospitalisation, médicaments), le niveau de prise en charge par le régime obligatoire est donc plus important pour ses bénéficiaires. A titre d'exemple, le niveau de remboursement des honoraires des médecins et dentistes est de 90% sur le régime Alsace-Moselle au lieu de 70% sur le régime général. A consommation médicale équivalente, le reste à charge est plus faible sur le régime Alsace-Moselle et donc les prestations remboursées par la complémentaire santé seront plus faibles également.

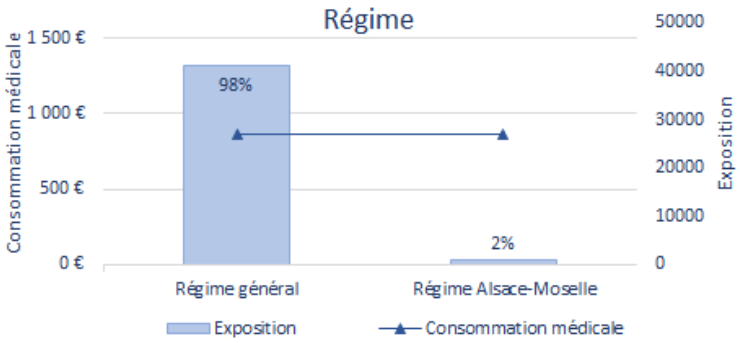


FIGURE 4.38 – Consommation médicale en fonction du régime

La détention d'option

Les bénéficiaires ayant souscrit des renforts de garantie semblent également consommer davantage que les non souscripteurs. Le graphique 5.12 montre une surconsommation de l'ordre de 10% sur les garanties de la base. Cet écart est vraisemblablement minoré. Les détenteurs d'option ont généralement des garanties de base moins couvrantes que les autres assurés.

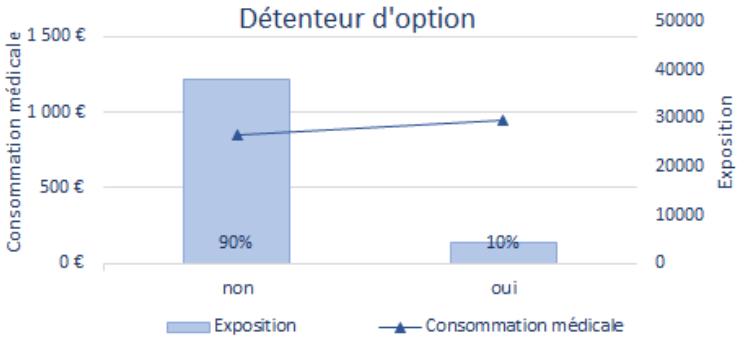


FIGURE 4.39 – Consommation médicale en fonction de la détention d'option

Annexe B - Cartes des différents zoniers

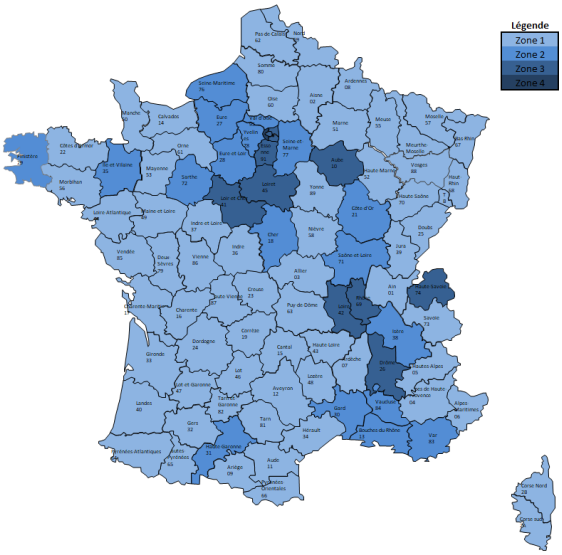


FIGURE 4.40 – Zonier hospitalisation (chambre particulière)

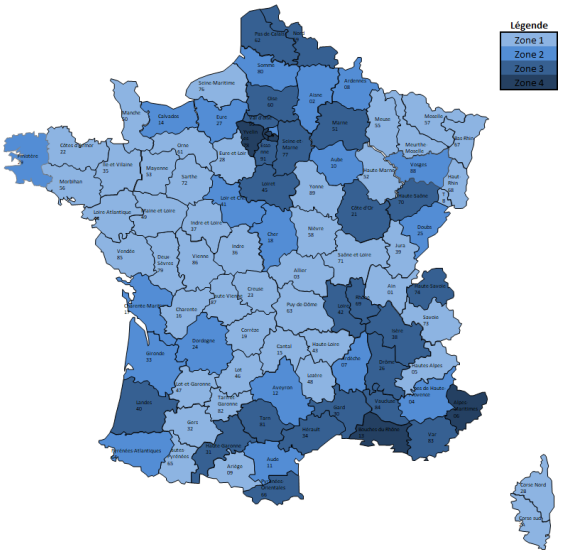


FIGURE 4.41 – Zonier soins courant

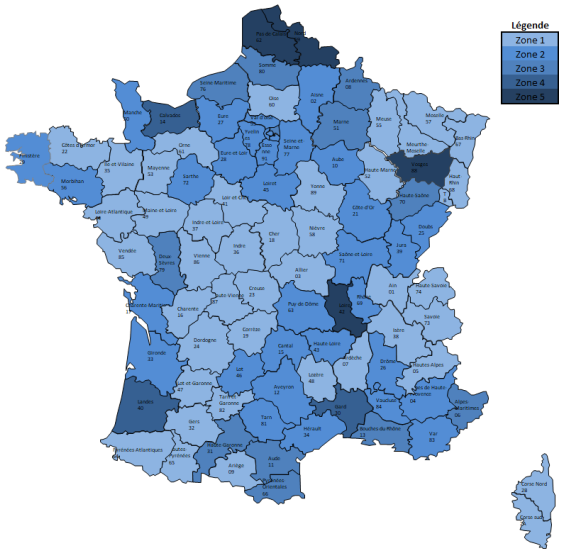


FIGURE 4.42 – Zonier pharmacie

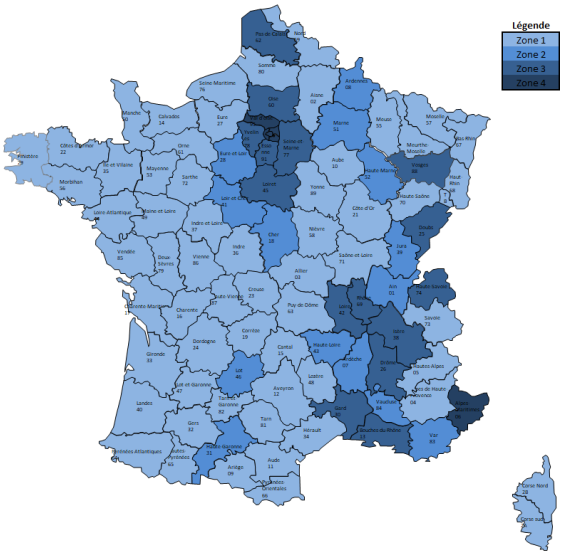


FIGURE 4.43 – Zonier dentaire

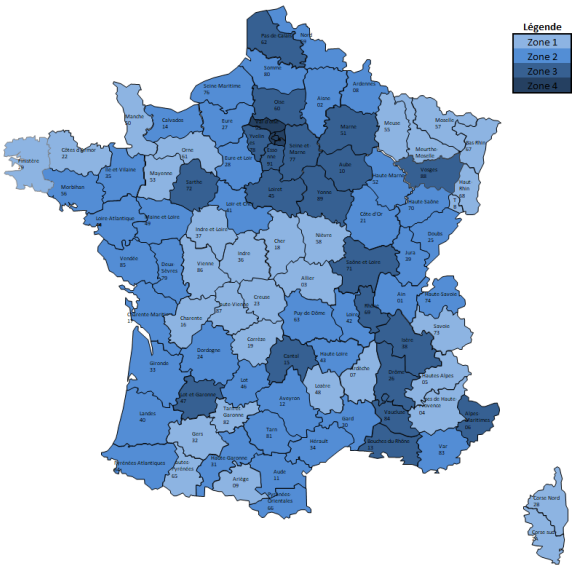


FIGURE 4.44 – Zonier optique

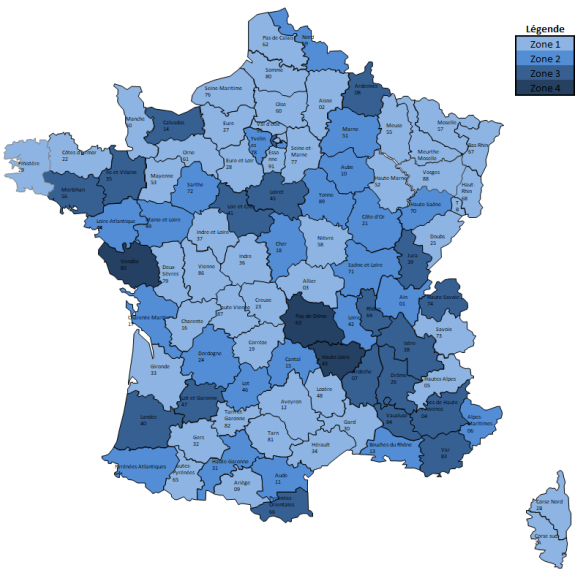


FIGURE 4.45 – Zonier bien-être

Annexe C - Graphiques des ajustements de la fréquence par les lois de Poisson et Binomiale Négative sur les autres postes

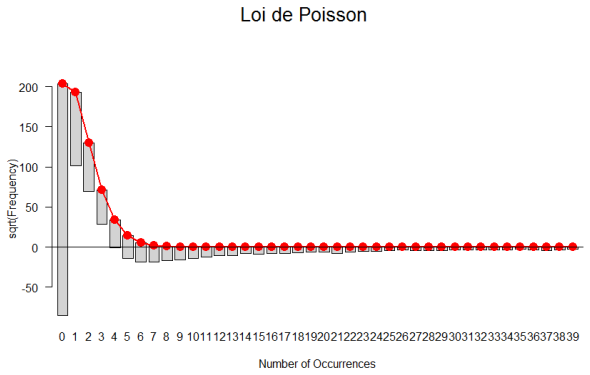


FIGURE 4.46 – Hospitalisation - Ajustement de la fréquence par une loi de Poisson

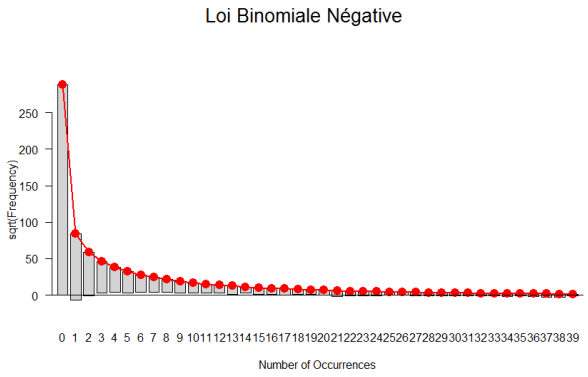


FIGURE 4.47 – Hospitalisation - Ajustement de la fréquence par une loi Binomiale Négative

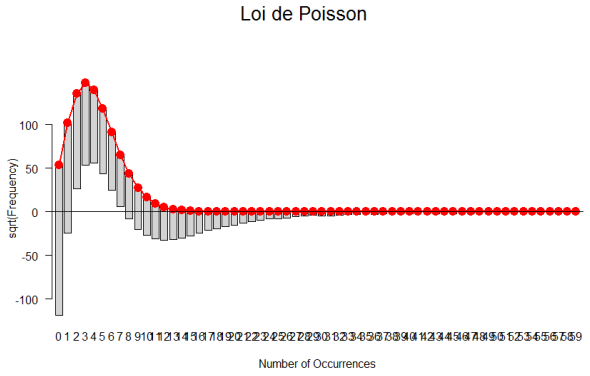


FIGURE 4.48 – Pharmacie - Ajustement de la fréquence par une loi de Poisson

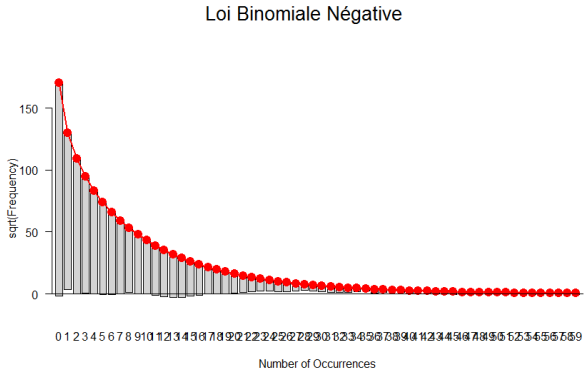


FIGURE 4.49 – Pharmacie - Ajustement de la fréquence par une loi Binomiale Négative

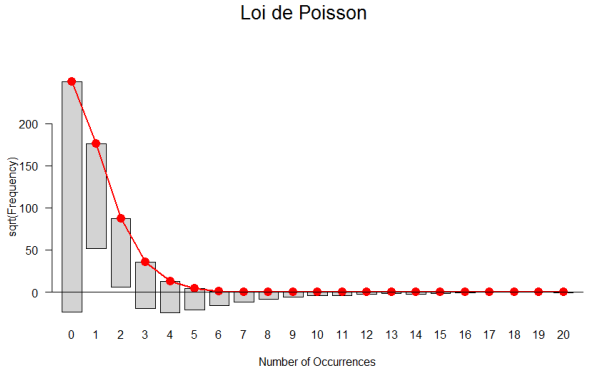


FIGURE 4.50 – Dentaire - Ajustement de la fréquence par une loi de Poisson

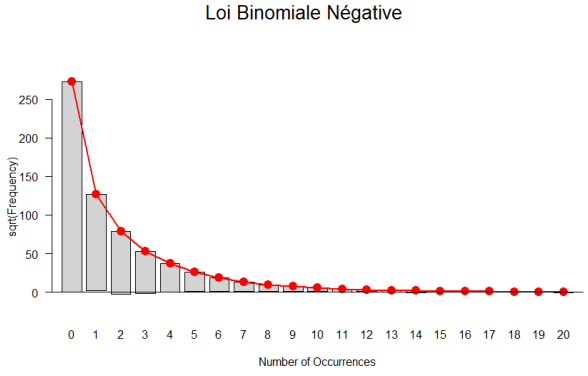


FIGURE 4.51 – Dentaire - Ajustement de la fréquence par une loi Binomiale Négative

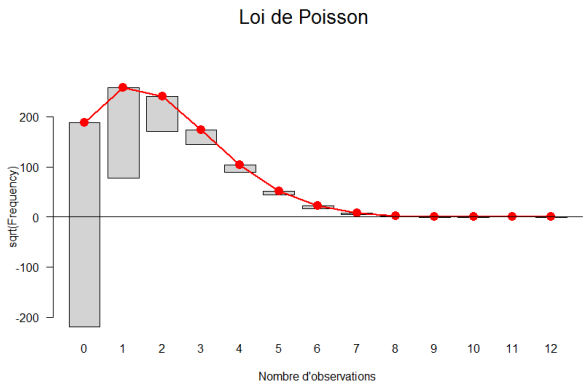


FIGURE 4.52 – Optique - Ajustement de la fréquence par une loi de Poisson

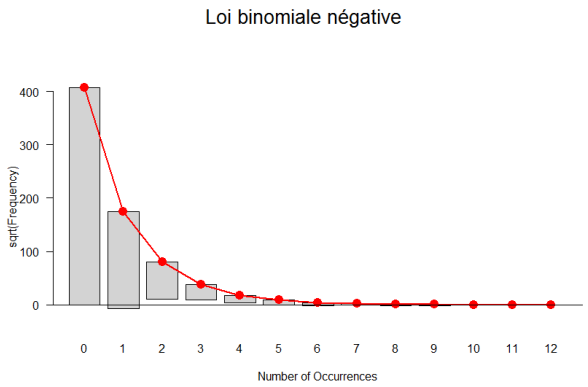


FIGURE 4.53 – Optique - Ajustement de la fréquence par une loi Binomiale Négative

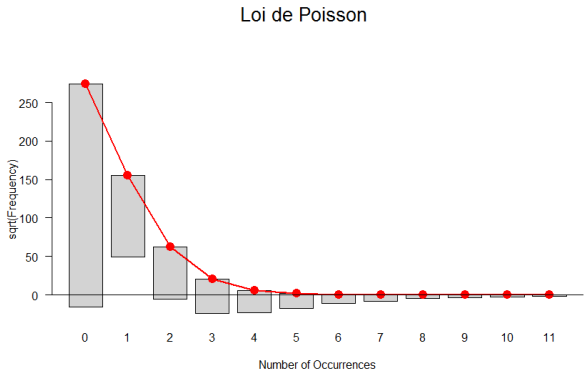


FIGURE 4.54 – Bien-être - Ajustement de la fréquence par une loi de Poisson

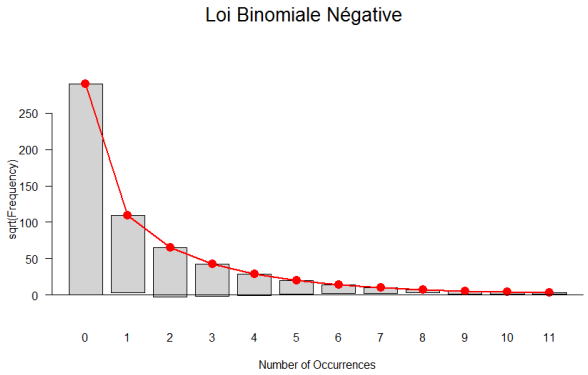


FIGURE 4.55 – Bien-être - Ajustement de la fréquence par une loi Binomiale Négative

Annexe D - Résultats de la modélisation CART sur les autres postes

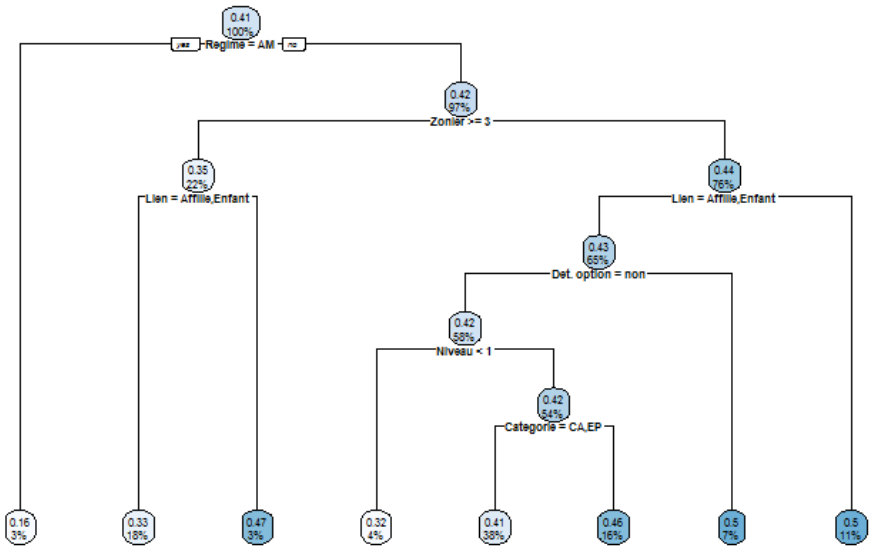


FIGURE 4.56 – Hospitalisation - Arbre CART sur la fréquence

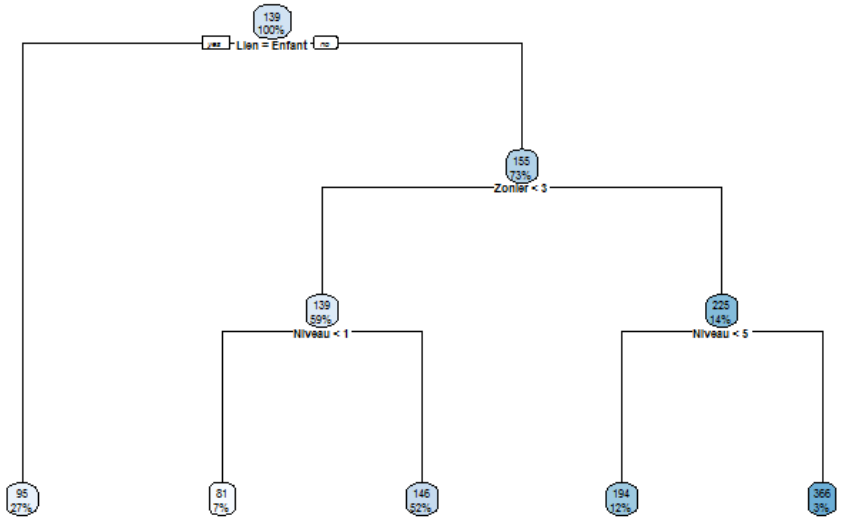


FIGURE 4.57 – Hospitalisation - Arbre CART sur le coût moyen

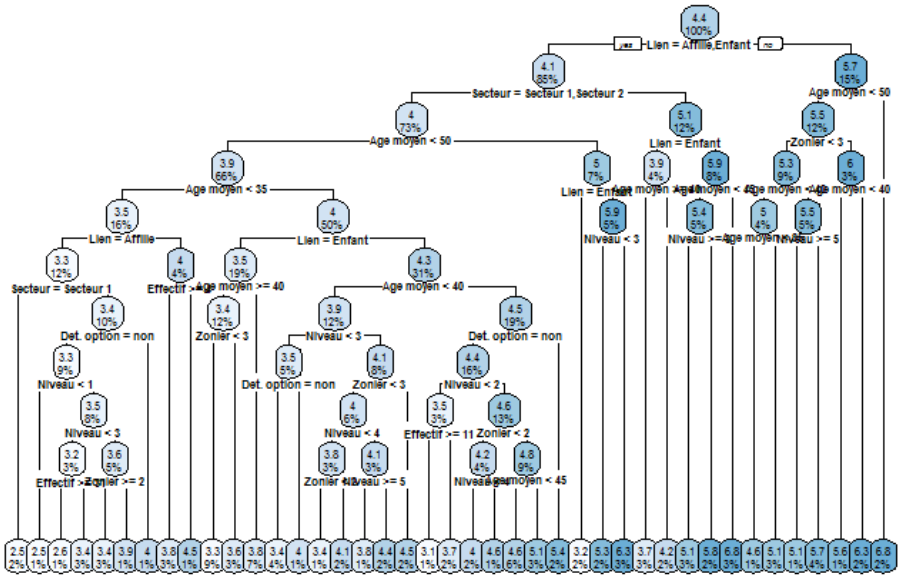


FIGURE 4.58 – Pharmacie - Arbre CART sur la fréquence

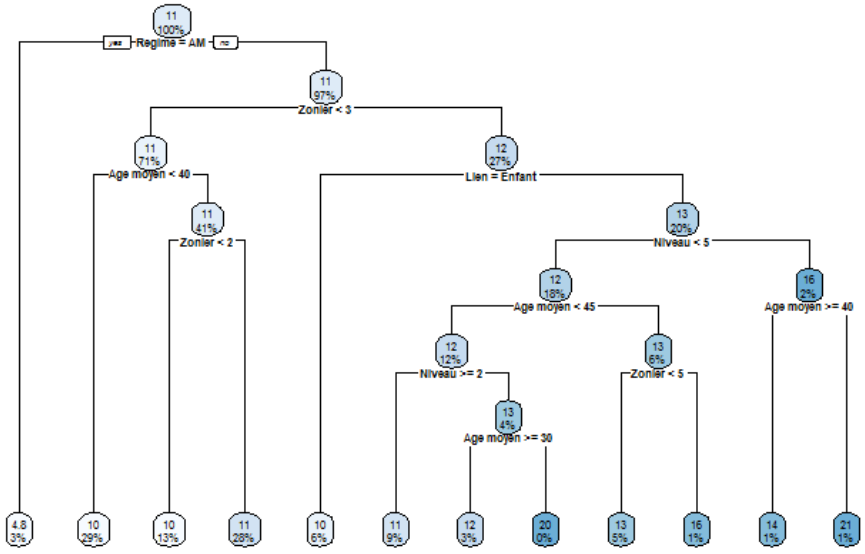


FIGURE 4.59 – Pharmacie - Arbre CART sur le coût moyen

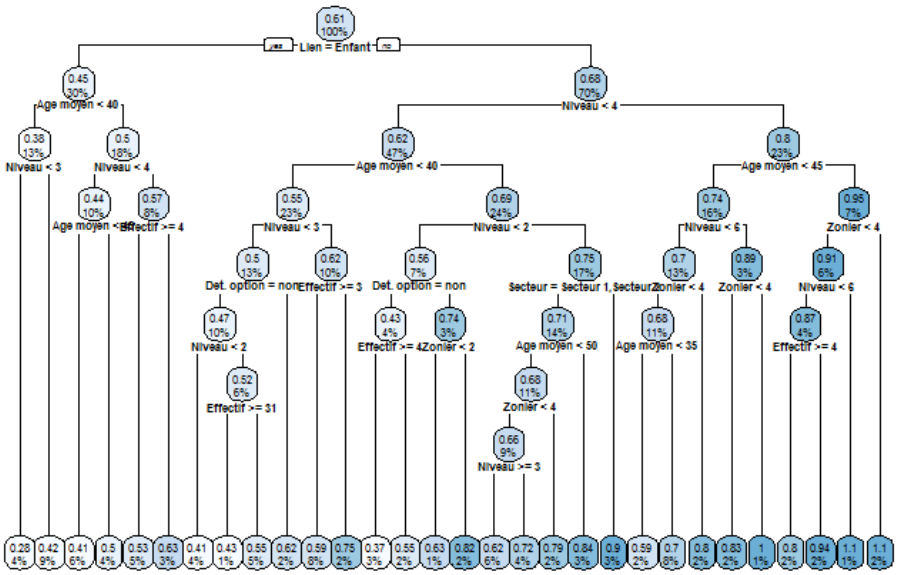


FIGURE 4.60 – Dentaire - Arbre CART sur la fréquence

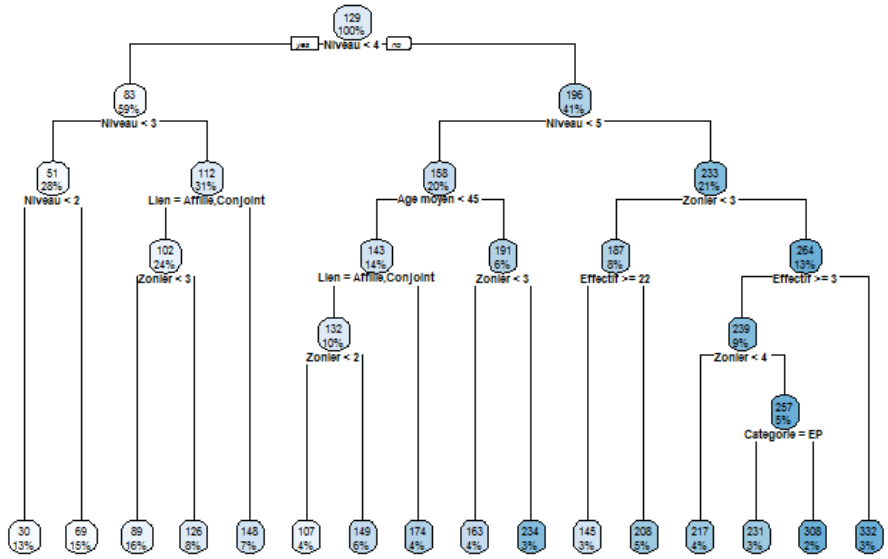


FIGURE 4.61 – Dentaire - Arbre CART sur le coût moyen

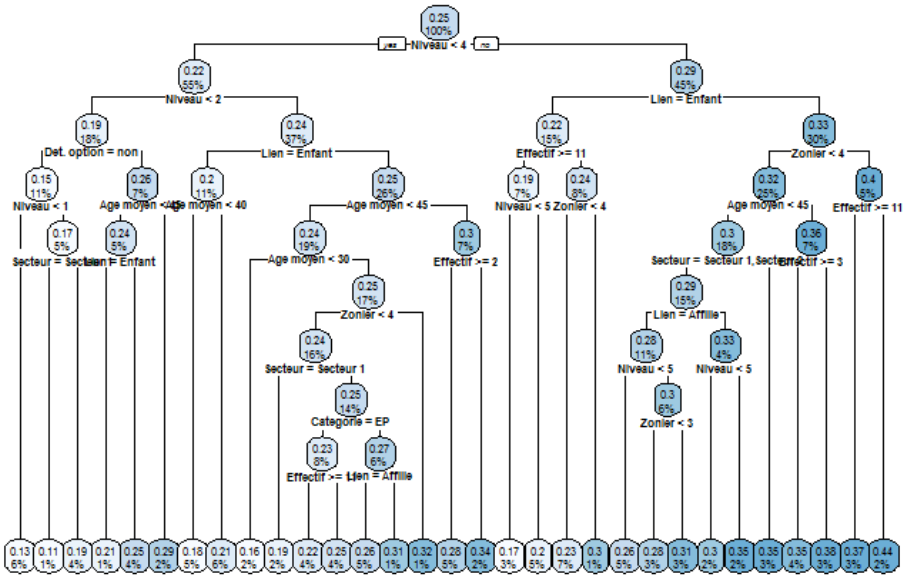


FIGURE 4.62 – Optique - Arbre CART sur la fréquence

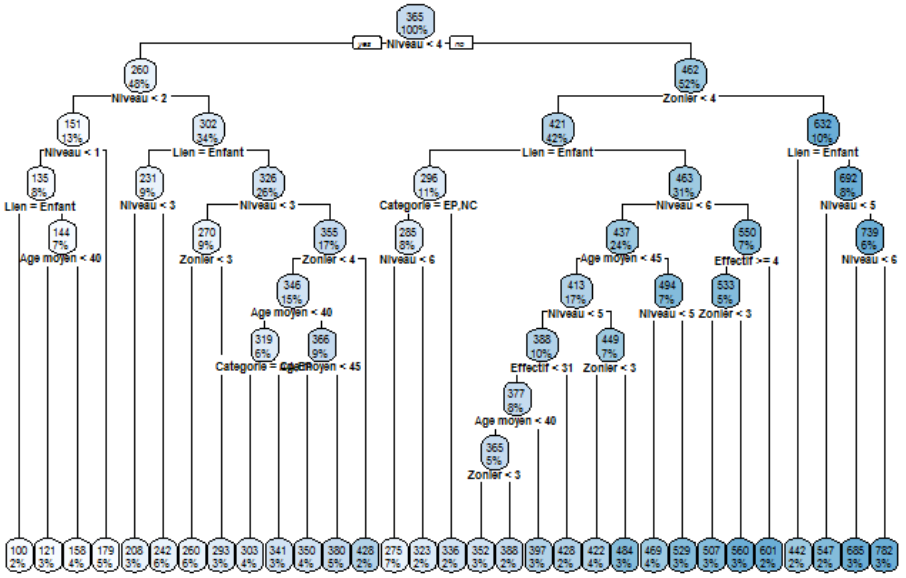


FIGURE 4.63 – Optique - Arbre CART sur le coût moyen

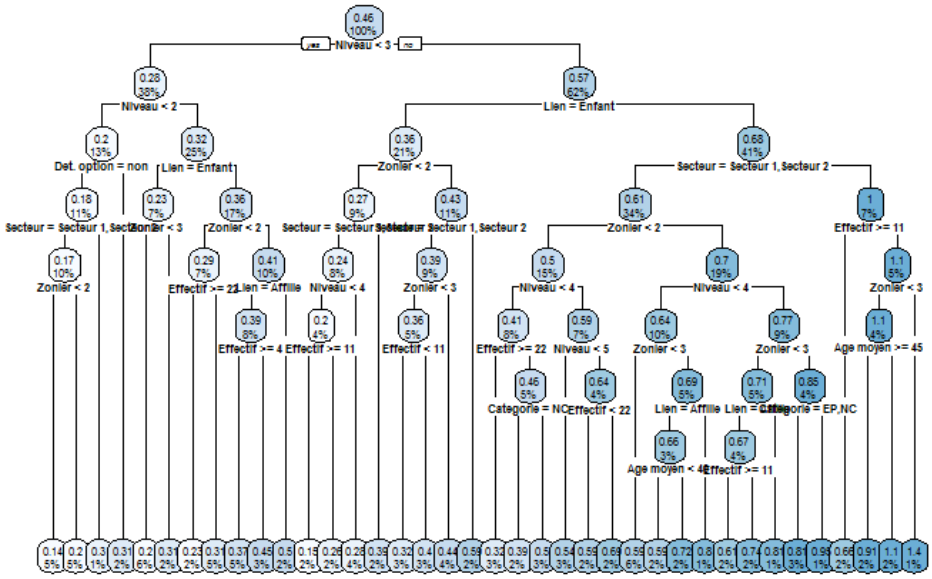


FIGURE 4.64 – Bien-être - Arbre CART sur la fréquence

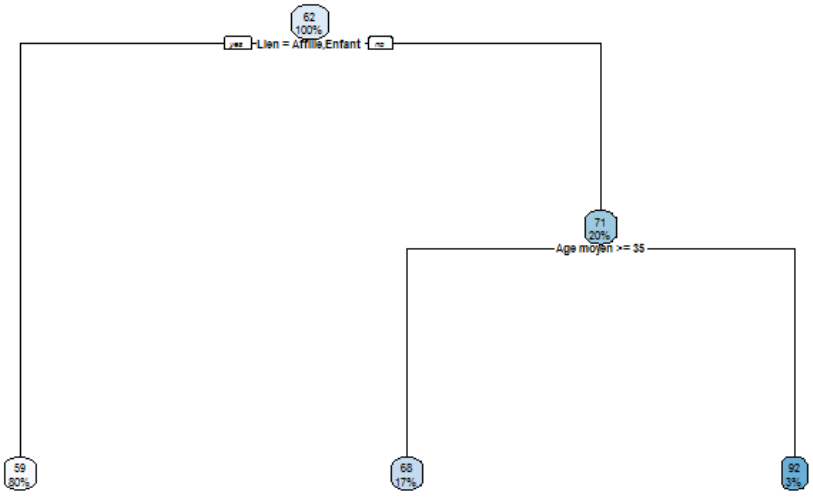


FIGURE 4.65 – Bien-être - Arbre CART sur le coût moyen