**Mémoire présenté le :**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA**
**et l'admission à l'Institut des Actuaires**

Par : Guerric BRAS

Titre : Accélération et sécurisation du processus de tarification sur-mesure : comment limiter les variables tarifaires ?

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

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

*Membre présents du jury de l'Institut des Actuaires* signature

*Membres présents du jury de l'ISFA*

*Entreprise :*

Nom : ACTUARIS

*Signature :*

*Directeur de mémoire en entreprise :*

*Nom : Jennifer CHRISTIN*

*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

*Secrétariat*

Signature du candidat

*Bibliothèque :*

## RÉSUMÉ

La tarification de contrats complémentaires sur-mesure en santé collective s'avère de plus en plus complexe et chronophage. En effet, les récents bouleversements législatifs du secteur ont accru le nombre d'études soumises ainsi que leur complexité. La question du respect des dates de réponse pour les appels d'offre ainsi que celle de la qualité des tarifs est alors posée.

Comment, dans ce contexte, optimiser les temps de tarification des dossiers et obtenir rapidement une estimation du tarif ?

Dans ce mémoire nous tentons de répondre à cette question en nous intéressant à la réduction du nombre de variables à retenir dans les modèles tarifaires.

Pour mener à bien cette étude, nous utilisons les tarifications sur-mesure effectuées par le cabinet ACTUARIS depuis 2009. Dans cette base de données sont disponibles : les données démographiques des salariés couverts, le détail des garanties santé ainsi que le tarif technique associé.

La première étape de ce mémoire a été d'uniformiser l'ensemble de ces études et de les placer dans un référentiel commun. Pour pouvoir comparer facilement les garanties entre-elles, elles ont toutes été remplacées par leurs équivalents euro.

L'algorithme CART a ensuite été utilisé pour déterminer les variables les plus importantes dans les processus tarifaires actuels. Les résultats de cet algorithme, confirmés par l'utilisation de modèles linéaires généralisés, ont permis de réduire de près de 90% la dimensionnalité du modèle.

Dès lors qu'il ne reste plus que les variables apportant le plus d'information (10%), il a ensuite fallu calibrer un modèle permettant d'obtenir une estimation du tarif. Pour cela, nous avons utilisé les réseaux de neurones. L'utilisation de cette méthode nous a permis d'approcher le tarif à plus ou moins 9%.

Pour réduire encore l'erreur d'estimation, la méthode dite du gradient boosté a été appliquée à des arbres de régression. Son application a permis de réduire la fonction d'erreur par rapport aux réseaux de neurones. Grâce à cet algorithme, nous avons obtenu de meilleures estimations des tarifs à partir d'un nombre réduit de variables. De plus, la simplicité d'utilisation des arbres de régression permet de créer facilement des maquettes de calculs pour comparer ou estimer les tarifs.

On démontre ainsi qu'il est possible d'obtenir une estimation tarifaire à partir d'un nombre réduit de variables pour permettre d'optimiser et de sécuriser le processus de tarification sur-mesure.

**Mots clés** : complémentaire santé collective, tarification sur-mesure, réduction dimensionnelle, algorithme CART, modèle linéaire généralisé, gradient boosté, arbres boostés.

## SUMMARY

The pricing of a custom employer complementary health insurance is more and more difficult and time-consuming. Indeed, recent legislative changes in this sector increase the number of studies and their complexity.

We can question on the ability of respecting deadlines for proposals and on the quality of the pricing.

How, in this context, to optimize the time for pricing a product and to obtain an estimation of the price?

In this work, we will try to answer this question by reducing the number of variables used in pricing models.

In order to do this study, we use all custom employer sponsored complementary health insurance priced by ACTUARIS since 2009. In this database we have: demographic data about insured employees, all health policies and the technical price.

The first step of this work was to standardize all studies and to use a common base. In order to be able to compare all warranties, we calculated their costs in euros.

CART algorithm was used to determine the most important variables in the actual model. The result of this algorithm was confirmed by GLM. This reduces the size of the model by 90%.

Now that we have only few variables, we have to obtain an estimation of the price using only this information. We used neural networks in order to have an evaluation of the price. Using this method, we can predict the real price with more or less 9%.

After, a gradient boosting method used on trees was chosen in order to reduce the error in estimates. Its application reduces the error compared to neural networks. This algorithm permits to obtain a best estimation of all prices using a smaller number of variables. Moreover, the simplicity of regression trees allows to create a model to compare easily or to estimate prices.

As a conclusion, it is possible to obtain an estimation of the price using just a few numbers of variables in order to optimize and to secure the pricing process.

**Key words**: sponsored complementary health insurance, custom pricing, dimensional reduction, CART algorithm, GLM, gradient boosting, tree boosting.

# NOTE DE SYNTHÈSE

## INTRODUCTION

La tarification santé collective connait actuellement de nombreux bouleversements qu'ils soient d'ordre législatif, avec la mise en application de l'ANI au 1er janvier 2016, ou encore du fait de l'arrivée du *Big Data*. Cela engendre un surcroit de travail pour les services souscription avec un nombre accru de dossiers qui sont par ailleurs complexifiés. L'obtention rapide d'une estimation du tarif dans un but de validation des résultats obtenus s'avère essentielle.

L'objectif de ce mémoire est ainsi de proposer une modélisation simplifiée de la tarification sur-mesure en assurance santé collective.

Dans un premier temps la détermination des variables les plus importantes pour la tarification santé sera effectuée par le biais de l'algorithme CART. Ensuite, à partir de ces variables, des modèles de réseaux de neurones ainsi que des arbres boostés par la méthode du gradient boosté seront utilisés pour obtenir une estimation du tarif.

## ASSURANCE COMPLÉMENTAIRE SANTÉ COLLECTIVE

L'assurance complémentaire santé est actuellement en plein bouleversement. Les premières modifications importantes sont d'ordre législatif et règlementaire avec la mise en place d'une obligation de couverture santé à l'ensemble des salariés du secteur privé au 1er janvier 2016 et la fin des clauses de désignations qui ouvre à la concurrence certaines branches d'activité. Cela a entrainé une augmentation importante des demandes de tarifications sur-mesure ces derniers mois et une augmentation du temps de traitement des dossiers. Un second bouleversement d'ordre technique est apparu avec l'avènement du *Big Data.* Bien que cette masse d'informations ne soit pas, pour le moment, directement exploitable en tarification santé collective, de nouveaux champs de recherche pour connaitre encore d'avantage les caractéristiques des assurés se profilent.

## SÉLECTION DES VARIABLES PRINCIPALES

### BASE DE DONNÉES ET RETRAITEMENTS

Pour réaliser cette étude, nous avons utilisé l'historique des tarifications sur-mesure du cabinet ACTUARIS. Cette base de données contient toutes les informations concernant l'ensemble des données démographiques des sociétés, l'ensemble de leurs garanties et leur tarif uniforme. Chacune des études de cette base utilise des modèles tarifaires et des quantifications des garanties différentes. Pour pouvoir les comparer facilement, une première étape du travail a été d'uniformiser

les modèles tarifaires en n'en conservant qu'un seul, et de calculer l'équivalent euro de chacune des garanties pour toutes les études.

### DÉTERMINATION DES VARIABLES PRINCIPALES

Pour déterminer les variables les plus importantes dans la tarification santé, l'algorithme CART a été utilisé dans un premier temps. En effet, cet algorithme détermine des arbres de régression qui permettent de segmenter l'espace des variables. Les variables considérées comme apportant le plus d'information sont les variables présentes dans les nœuds de l'arbre. Pour consolider ce choix de variables, une étude des corrélations ainsi qu'une validation par l'utilisation de GLM ont été effectuées. Finalement, l'utilisation de ces méthodes permet de réduire de près de 90% le nombre de variables présentes dans la tarification.

## APPRENTISSAGE PAR RÉSEAUX DE NEURONES

Les réseaux de neurones correspondent à un algorithme d'apprentissage qui permet d'estimer le tarif uniforme à partir d'un nombre réduit de variables. Une première estimation a été obtenue en utilisant les 10% de variables conservées et en appliquant un apprentissage supervisé sur une base d'apprentissage représentant 70% des données. Ensuite, ce modèle a été amélioré par l'utilisation de plusieurs procédés. Le premier a consisté en l'utilisation de plusieurs réseaux de neurones identiques pour obtenir une estimation afin de diminuer l'impact de la pondération aléatoire des poids inter-synaptiques. Comme la base de données ne présente pas de variables aléatoires mais uniquement des données contractuelles et démographiques, le deuxième procédé utilisé pour améliorer l'estimation a été d'augmenter la taille de la base d'apprentissage. Finalement, le modèle conservé permet d'obtenir une estimation des tarifs uniformes à plus ou moins 9%.

L'utilisation des réseaux pose néanmoins problème pour l'estimation des garanties très haut de gamme qui sont présentes dans le portefeuille. Pour contourner cela et améliorer l'estimation, des méthodes dites boostées ont été utilisées.

## MÉTHODES BOOSTÉES

Les méthodes boostées consistent à réaliser plusieurs prédictions successives et à améliorer au fur et à mesure la qualité de l'estimation dans les zones où l'erreur est la plus importante. Nous nous sommes focalisés sur des classifieurs faibles de type arbres de régression, plus communément appelés arbres boostés. Cette méthode présente l'avantage de pouvoir améliorer la prédiction pour les garanties haut de gamme et plus généralement de réduire de 14% l'erreur d'estimation représentée par le RMSE par rapport au réseau de neurones.

Cette étude permet également d'obtenir un résultat plus surprenant sur les variables les plus importantes dans l'estimation du tarif. En effet, on peut noter une prépondérance des variables

représentant les garanties dentaires qui sont plus importantes pour l'estimation du tarif que certaines données démographiques.

## CONCLUSION

Dans ce mémoire, nous avons proposé plusieurs méthodes pour réduire la dimensionnalité des modèles tarifaires des complémentaires santé sur-mesure. Finalement, la meilleure estimation des tarifs est obtenue par l'utilisation des arbres boostés. Ce modèle aura pour but d'obtenir rapidement une estimation du tarif pour vérifier si des erreurs ont été commises dans une tarification complète ou encore pour connaître son positionnement marché à l'avance par rapport à la concurrence.

Ces méthodes d'estimation peuvent néanmoins être améliorées. Il serait en effet intéressant de voir l'apport de méthodes comme l'algorithme MARS ou encore l'appliquer sur un modèle tarifaire de type barème. Néanmoins il ne faut pas perdre de vue les spécificités de l'assurance complémentaire santé collective qui nécessite l'utilisation de l'ensemble des variables pour obtenir un tarif en adéquation totale avec le risque couvert.

# SYNTESIS

## INTRODUCTION

The pricing of a custom employer sponsored complementary health insurance faces numerous changes: legislative changes with the instauration of the ANI in January 1th 2016 and the apparition of the Big Data. This generates more work for subscription services with the augmentation of the number of pricing demands which are more and more complex. The possibility to obtain a quick estimation of the price in order to check the pricing process seems to be interesting.

The objective of this work is to propose a simple model for pricing a custom employer sponsored complementary health insurance.

To begin, we will determine the most important variables in actual models using CART algorithm. Then we will propose a new model using fewer variables thanks to neural networks and boosting trees in order to have an estimation of the real price.

## EMPLOYER SPONSORED COMPLEMENTARY HEALTH INSURANCE

Employer sponsored complementary health insurance is in turmoil. First important changes are legislative and regulatory changes with the obligation to insure all employees in the private sector and the end of designation clauses which introduces competition in some sectors. This causes an important augmentation of custom pricing demands the last few months and an increase of the processing time of each pricing. A second change appears, it's a technical change with the emergence of Big Data. Even if this base of information is not, for the moment, usable for custom employer sponsored complementary health insurance, new fields of research in order to know more policyholders are profiling.

## MAIN VARIABLES SELECTION

### DATA BASE AND REPROCESSING

To realize this study, we have the historic of all custom pricings made by ACTUARIS. In this database we have: demographic data about insured employees, all health policies and the technical price. Each one of those studies has a specific pricing model and a specific quantification for all warrants. In order to be able to easily compare each study, the first step of this work was to standardize all pricing models by conserving only one. After, we calculate the cost of all warranties in euros.

### MAIN VARIABLES DETERMINATION

In order to determine main variables, the CART algorithm is used in a first step. Indeed, this algorithm permits to create regression trees which allow to segment the state variables. Variables considered to bring the most information are variables in tree's nodes. To consolidate this choice, a study of correlations and GLM models are used. At the end, the utilization of those models permits to reduce the number of variables by 90%.

## NEURAL NETWORKS LEARNING

A neural network is a learning algorithm which can estimate the price using a reduce number of variables. A first estimation is obtained using supervised clustering in a data base representing 70% of the data available. After, this model is improved by the used of some processes. The first consists in using several neural networks in order to reduce the impact of the random weighting between two synapses. As the data base doesn't have random variables but only demographic and contractual ones, the second process consists in growing the size of the learning data base. Finally, the model permits to obtain an estimation of the price with more or less 9%.

However, the use of neural networks is limited for the estimation of top ranged warranties. In order to solve this problem and to improve the estimation, boosting methods are used.

## BOOSTING METHODS

Boosting methods consist in realizing several successive predictions to increase the quality of the estimation where the mistake is the most important. In this part, we focus on boosting trees. This method can improve the model by reducing the error by 14% compared with neural networks.

This study shows another result for the estimation of the price. Indeed, we can notice the predominance of variables representing dental warranties which are more important for the estimation than some demographic information.

## CONCLUSION

In this work, we propose several methods to reduce the dimension of pricing models. Finally, the best estimation is obtained using boosting trees. This model aims to have an estimation of the price rapidly to be able to see if there are mistakes in the pricing or to be able to know its own specific market positioning before the whole pricing.

Those methods can be improved. In fact, it could be interesting to measure the impact of the MARS algorithm or to use this method on a scale model. Nevertheless, we must not forget the specificity of

pricing a custom employer sponsored complementary health insurance which needs the use of all available variables.

## REMERCIEMENTS

Je tiens tout d'abord à remercier Jennifer CHRISTIN et Cécile PARADIS du cabinet ACTUARIS pour m'avoir accompagné dans la rédaction de ce mémoire, pour leurs conseils avisés et leur écoute attentive.

Je souhaite également remercier Yahia SALHI pour ses connaissances techniques, ses remarques pertinentes et pour toute l'aide qu'il m'a fournie.

Je tiens également à remercier Alexandra BARRAL, Baptiste DIELTIENS, Charles GESLIN, David HICHAM, Hélène JOURDAIN et Yufei LUO pour leur aide dans la réalisation de ce travail.

Enfin je remercie mon entourage et plus particulièrement mon frère Xavier pour leur soutien et leurs encouragements.

## SOMMAIRE

# INTRODUCTION

Le processus de tarification est le premier à être réalisé en assurance. C'est un élément clef qui ne doit pas être négligé sous peine de mauvais résultats et de pertes futures. La tarification de produits santé est un processus transverse qui associe à la fois les services actuariat, commerciaux, juridique et souscription d'une société d'assurance.

La tarification santé connait actuellement deux bouleversements. Le premier est d'ordre technique avec l'arrivée du *Big Data* qui laisse entrevoir de multiples possibilités d'amélioration des modèles actuels par l'introduction de nouvelles variables explicatives par exemple. Le deuxième est d'ordre juridique avec les profonds changements règlementaires auxquels fait face l'assurance santé collective. On peut citer notamment la fin des clauses de désignation ou encore l'accord national interprofessionnel instaurant l'obligation de la mise en place d'un régime collectif frais de santé obligatoire.

Au centre de cela se trouve le service souscription. Ce service fait face à une recrudescence du nombre de tarifications à effectuer. Il doit également suivre les modèles des actuaires et les préconisations du service juridique tout en respectant des délais de plus en plus réduits et une concurrence accrue sur certaines branches qui étaient jusqu'alors protégées par des clauses de désignation. Tous ces éléments font que la tarification d'un contrat complémentaire santé pour une société d'assurance devient de plus en plus longue et complexe. A titre d'exemple, un grand groupe paritaire a vu le nombre de demandes de tarification sur-mesure augmenter de 60% en août 2015 par rapport à août 2014. Parallèlement le temps de traitement des dossiers a augmenté de 20%.

A cela s'ajoute les spécificités de la tarification santé. La difficulté principale réside dans la multitude de variables rentrant en compte dans la tarification. Elles sont de deux types : les variables relatives à l'entreprise et aux caractéristiques de ses salariés et les variables relatives aux niveaux de garanties, acte par acte. Un modèle tarifaire peut contenir entre 50 et 150 variables en fonction des catégorisations des garanties choisies par l'organisme assureur. Pour chacune de ces variables, l'actuaire ou le souscripteur devra vérifier la conformité de la donnée et son respect des obligations conventionnelles et règlementaires.

Le but de ce mémoire est de proposer une modélisation simplifiée de la tarification santé collective en réduisant la dimensionnalité des modèles tarifaires actuels. La réduction du nombre de variables dans la tarification peut être un atout pour obtenir rapidement une évaluation du tarif. Tout d'abord, il est intéressant pour le service technique d'avoir une approximation du tarif comme outil de vérification des travaux. De plus, pour les équipes commerciales, la connaissance de cette approximation permettrait, entre autre, de savoir comment se positionne la concurrence avant même que le dossier ait été traité complètement et ainsi de prioriser les études à réaliser.

Ce mémoire se décompose en quatre parties.

Dans un premier temps le contexte actuel de la tarification santé sera présenté ainsi que le modèle tarifaire le plus couramment utilisé : le modèle fréquence * coût.

Dans un second temps nous étudierons plus en profondeur la base de données à notre disposition puis nous déterminerons les variables les plus importantes dans le processus de tarification par l'utilisation de l'algorithme CART. Puis nous confirmerons ces résultats par l'application d'une modélisation GLM.

Dans un troisième temps, un nouveau modèle ne conservant que les variables les plus importantes va être créé. Il sera basé sur la théorie de l'apprentissage supervisé des réseaux de neurones pour permettre l'obtention d'une estimation des tarifs.

Pour finir, une amélioration des estimations sera proposée par le biais de l'utilisation des méthodes boostées.

# I) L'ASSURANCE COMPLÉMENTAIRE SANTÉ COLLECTIVE

## A) LE RÉGIME DE BASE

L'assurance santé est définie tout d'abord par la présence d'un régime de base. Pour un salarié, le régime de base de la personne couverte dépend à la fois du type d'activité qu'elle exerce ainsi que de la situation géographique du siège de l'entreprise en question. Plusieurs différences existent entre les régimes. Elles peuvent être sur le montant des remboursements comme c'est le cas entre le régime général et le régime Alsace-Moselle, ou encore sur la taxation des régimes en question comme par exemple avec la Mutualité Sociale Agricole (MSA).

L'ensemble des remboursements de la Sécurité Sociale pour le régime général et pour les actes étudiés se trouve en ANNEXE 1.

## B) LA COMPLÉMENTAIRE SANTÉ

Les remboursements de la Sécurité Sociale et des différents régimes ne suffisent pas, la plupart du temps, à rembourser intégralement les soins. Cela est dû à plusieurs choses : premièrement la Sécurité Sociale doit faire des économies et limiter la surconsommation médicale. De ce fait, elle ne rembourse pas ou très peu certains actes médicaux. Deuxièmement, certains praticiens appliquent des tarifs bien supérieurs à ceux définis par la Sécurité Sociale, ainsi le reste à charge pour l'assuré est d'autant plus important et il devient nécessaire d'avoir recours à une complémentaire santé pour assurer ses soins futurs.

Le marché de la complémentaire santé est partagé entre trois types d'organismes différents. Les institutions de prévoyance pour 17% du marché en 2013, les compagnies d'assurance pour 27% et majoritairement les mutuelles santé pour 56%.

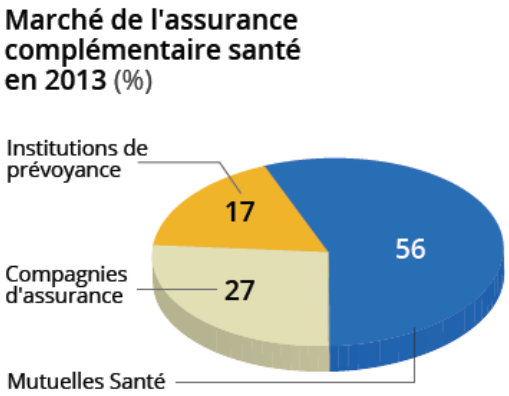



*Figure 1 : Marché de l'assurance complémentaire santé en 2013*

## 1) LES INSTITUTIONS DE PRÉVOYANCE

Les institutions de prévoyance sont des organismes d'assurance à but non lucratif et sont administrées de façon paritaire entre les partenaires sociaux. Ainsi, à la tête d'une institution de prévoyance se suivent les représentants du patronat et les représentants des syndicats. Ces organismes dépendent du code de la Sécurité Sociale.

Les institutions de prévoyance sont majoritairement recommandées dans les accords de branche.

## 2) LES COMPAGNIES D'ASSURANCE

Les compagnies d'assurance dépendent du code des assurances. La majorité du temps, ces organismes sont des sociétés de capitaux dont le but principal est l'optimisation de la rentabilité de ces capitaux.

## 3) LES MUTUELLES

Les mutuelles sont à but non lucratif et sont administrées par des représentants des assurés. Elles sont régies par le code de la mutualité et peuvent être spécialisées dans un type d'activité ou dans un secteur géographique bien précis. Ces derniers temps, on a observé un rapprochement accéléré de différentes mutuelles au vu du changement important que le secteur de l'assurance santé a connu. Ainsi leur nombre est passé de 704 en 2014 à une estimation de 206 pour 2016. Ce regroupement massif a pour but de créer des entités plus solides financièrement et plus aptes à s'adapter à une concurrence accrue dans un marché en plein développement.

## C) CONTRATS D'ASSURANCE SANTÉ

Plusieurs types de contrats complémentaires santé existent, les plus courants sont les contrats collectifs d'entreprise et les contrats individuels. La quasi-totalité de la population française est actuellement couverte par une garantie complémentaire santé individuelle ou collective.

### 1) COMPLÉMENTAIRE SANTÉ INDIVIDUELLE

N'importe quel individu peut juger qu'il a besoin d'une couverture complémentaire santé pour répondre à ses propres besoins. Le marché de l'individuel est actuellement majoritairement détenu par les mutuelles avec 66% du marché, viennent ensuite les compagnies d'assurance et enfin les institutions de prévoyance.

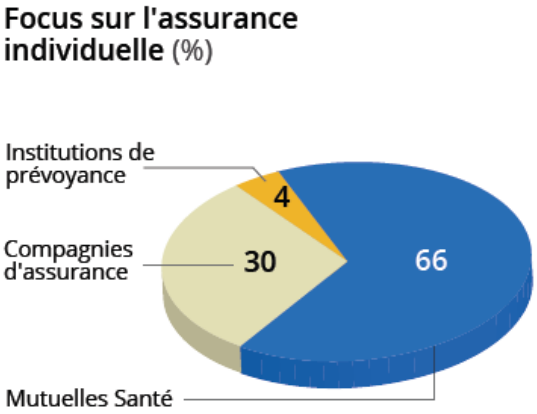



*Figure 2 : Focus de l'assurance individuelle*

Plusieurs types de contrats peuvent être proposés comme par exemple des contrats uniques où l'assuré ne choisit qu'un niveau de couverture global, ou un contrat par niveaux où l'assuré pourra décider d'être très bien couvert pour un certain type de prestations et avoir une garantie entrée de gamme dans d'autres. Au niveau de la tarification, ces deux types de contrats sont très différents car dans le deuxième cas la notion de distance entre les garanties doit être prise en considération pour mesurer l'anti-sélection.

En effet, si un assuré choisit de se couvrir à des niveaux très hauts sur certains postes et très bas sur d'autres, autrement dit si la distance entre les garanties est élevée, cela indique qu'il a un besoin spécifique sur des postes précis. Ceci se traduit par une surconsommation future sur ces postes qui doit être prise en compte par une majoration tarifaire.

## 2) COMPLÉMENTAIRE SANTÉ COLLECTIVE

Les complémentaires santé collectives correspondent à des garanties souscrites par une entreprise pour l'ensemble ou une partie de ses salariés. La mise en place d'une complémentaire santé collective peut être faite par une décision unilatérale de l'employeur, un référendum de l'ensemble des salariés ou encore être décidée en amont par un accord de branche. Le marché de la complémentaire santé est actuellement majoritairement tenu par les institutions de prévoyance puis viennent les mutuelles et enfin les compagnies d'assurance.

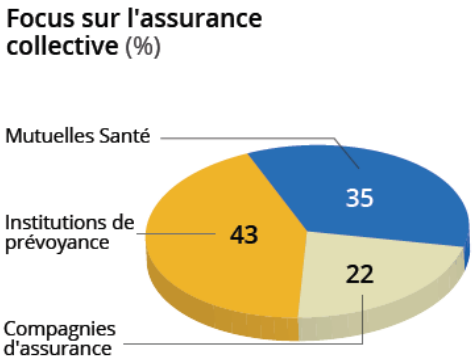



*Figure 3 : Focus de l'assurance collective*

La complémentaire santé a connu ces derniers temps des changements importants avec l'arrivée de nouvelles règlementations dont l'ANI qui laisse envisager un transfert massif du marché de l'assurance individuelle vers l'assurance collective.

Le marché de la complémentaire santé collective est celui qui a connu le plus de bouleversements en un minimum de temps. En effet trois modifications majeures ont eu lieu ces dernières années. L'ANI avec l'apparition d'une couverture complémentaire santé collective obligatoire pour tous les salariés du secteur privé, le changement de définition des contrats solidaires et responsables et l'abrogation de l'article L 912-1 du code de la Sécurité Sociale sur les clauses de désignation. Ces différents points seront vus en détail par la suite car ils impactent de façon importante la tarification des complémentaires santé.

Il y a principalement deux types de contrats en assurance santé collective. Les contrats standards et les contrats sur-mesure.

Les contrats standards correspondent à des contrats non modifiables par l'entreprise ou par le commercial en charge du dossier.

Parmi les contrats standards, on peut identifier les contrats CCN (Convention Collective Nationale), les contrats labellisés CCN, les contrats référencés et les standards propres à l'assureur.

Les contrats CCN correspondent à des garanties et des taux négociés au niveau d'une branche professionnelle. Les assureurs étaient initialement désignés. Cela signifie que les entreprises de la branche faisant partie des organisations patronales signataires devaient couvrir leurs risques chez l'assureur désigné. Actuellement, la plupart des désignations sont tombées et ont été remplacées par des recommandations.

Les contrats labellisés sont des contrats qui respectent les minimas de garanties d'une CCN mais qui n'ont pas été soumis à une négociation entre les partenaires sociaux. La labellisation permet à un assureur d'être visible au niveau de la branche en proposant les taux et les garanties qu'il souhaite sous réserve du respect des accords conventionnels. Le tarif n'est pas imposé comme dans les CCN mais il est basé sur un nombre très réduit de variables, la plupart du temps l'âge moyen et la localisation géographique suffisent à la détermination du tarif pour une société.

Les contrats référencés sont des contrats qui ont été négociés par les partenaires sociaux sans pour autant respecter l'ensemble des formalismes des appels d'offre qu'imposent les différentes conventions et le droit du travail. On retrouve ce type de contrat dans les branches où aucun accord sur les garanties frais de santé n'a été établi ou en cas d'absence de négociation au niveau de la branche. Le référencement d'offres est possible pour apporter des solutions aux entreprises de la branche. Un exemple marquant est celui de la branche de la Métallurgie. L'Union des Industries et Métiers de la Métallurgie (UIMM) a référencé les offres de quatre assureurs pour faciliter les démarches des entreprises métallurgiques dans leur choix de complémentaire santé obligatoire. Comme pour les contrats labellisés, aucun taux n'est imposé et le calcul du tarif se fait simplement à l'aide de variables clés de l'entreprise.

Le dernier type de contrat, que l'on peut appeler standard, est le standard propre à la société d'assurance. Ces contrats sont adaptés aux petites structures, ils ne prennent en considération dans la tarification que certaines informations sur l'entreprise comme une approximation de l'âge moyen ou encore la localisation géographique. La création de modèle tarifaire pour les contrats standards est basée sur la consommation de l'ensemble des entreprises adhérentes à ce standard. L'ensemble des résultats sont mutualisés sur ce standard ce qui ne permet pas à une entreprise adhérente de négocier son régime en fonction de ses résultats propres.

Le deuxième grand type de contrat est le contrat sur-mesure. Ces contrats sont entièrement personnalisables. Les entreprises peuvent demander les garanties qu'elles souhaitent sous réserve qu'elles respectent les minimas conventionnels et qu'elles ne soient pas en dehors des limites de souscription de la société d'assurance. Une contrainte sur les effectifs couverts peut cependant être mise en place pour obtenir une mutualisation suffisante du risque. La tarification de ce type de contrat est assez longue car elle nécessite d'analyser l'ensemble des propositions de garanties demandées et elle impose une connaissance parfaite des caractéristiques démographiques de l'entreprise. Ainsi la connaissance de variables aussi différentes que la proportion de conjoints à charge, le nombre maximal de séances d'acupuncture ou encore les plafonds des garanties dentaires sont nécessaires pour chacune des études de contrat sur-mesure. Ce type de contrat à l'avantage d'être entièrement pilotable par l'entreprise en fonction de ses résultats. Ainsi une entreprise ayant

un ratio de sinistres sur primes inférieur à 100% pourra choisir d'augmenter ses garanties pour un coût équivalent ou alors de réduire le prix de sa complémentaire santé. Ce type de contrat coûte relativement cher aux sociétés d'assurance car il impose d'importants frais de gestion, liés à l'absence d'automatisation des remboursements des garanties qui sont toutes différentes d'une société à l'autre, et au processus de tarification plus lourd. C'est pour cela que ces contrats sont réservés aux entreprises avec des effectifs suffisamment importants.

# 3) BOULEVERSEMENTS LÉGISLATIFS

## 3.1) L'ACCORD NATIONAL INTERPROFESSIONNEL

Le 11 janvier 2013, un accord de sécurisation et de flexibilité des parcours professionnels des salariés a été signé : l'ANI.

Le premier article de l'ANI prévoit la généralisation de la couverture santé obligatoire à l'ensemble des salariés des entreprises avant le 1er janvier 2016. Cette obligation de couverture est accompagnée de nouvelles contraintes comme le respect de certaines garanties minimales ou encore une participation minimale de l'employeur à hauteur de 50% des cotisations. Des organismes peuvent être recommandés au sein d'une branche et donc mutualiser au-delà du niveau de l'entreprise.

En 2015, environ 86% des salariés sont couverts par leur entreprise pour leurs frais de santé. Cela représente plus de 2,5 millions de plus par rapport à l'année 2014. L'arrivée de l'ANI va bousculer le marché de la complémentaire santé puisque cette proportion va atteindre les 100%.

L'ANI engendrera le plus de changements dans la répartition entre les différents types d'assurance. Une estimation fait référence à un transfert de 7 millions de salariés de l'assurance individuelle vers l'assurance collective. Ce transfert aura des répercussions majeures dans le secteur de l'assurance santé. Certains acteurs de l'assurance individuelle verront leur part de marché réduite et une concurrence accrue est née sur le marché de la complémentaire santé collective après l'abrogation de l'article L 912-1 du code de la Sécurité Sociale. Cet article faisait référence aux closes de désignation. Ces closes permettaient de contraindre les entreprises à adhérer à une société d'assurance désignée dans la convention de la branche à laquelle elle appartient. La fin de cette clause permet d'augmenter la concurrence et d'ouvrir de nouveaux marchés pour certains acteurs de l'assurance.

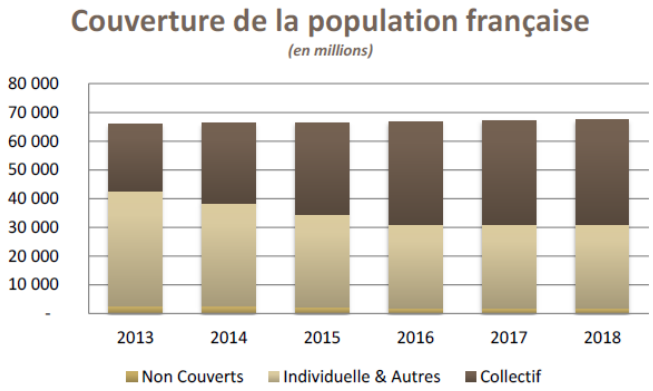



*Figure 4 : Couverture de la population française*

Ces changements majeurs dans la répartition du poids entre l'individuel et le collectif aura un impact important sur les services de souscription des différents acteurs. En effet, l'obligation de la complémentaire santé va entraîner une augmentation du nombre de dossiers à traiter dans ces services ce qui peut augmenter les délais de traitements et les erreurs dans la tarification.

## 3.2) LES CONTRATS SOLIDAIRES ET RESPONSABLES

Un contrat est dit solidaire et responsable s'il respecte un certain nombre de conditions. Il a pour but d'assurer une bonne couverture des dépenses de santé de l'assuré sans pour autant encourager des dépenses inutiles de sa part. Cela conduit à une modulation du montant des remboursements en fonction du respect du parcours de soin coordonné. Le contrat responsable reprend les mêmes mécanismes incitatifs et limitant que le régime de base de la Sécurité Sociale pour obtenir une politique de santé plus harmonieuse.

Ils bénéficient d'incitations fiscales plus favorables que les contrats non responsables. L'élément le plus marquant est la diminution de la taxe sur les conventions d'assurance de 14% à 7% pour les contrats responsables.

Pour être qualifié de responsable, un contrat d'assurance collective doit respecter des minima et des maxima sur les garanties.

| Actes | Plancher | Plafond |
|---|---|---|
| Honoraires des médecins (ville et hospitalisation) | 100% TM | 225% BR-SS puis 200% BR-SS à partir de 2017 et -20% de la couverture des médecins adhérents au Contrat d'Accès aux Soins (CAS) |
| Prothèses dentaires et orthodontie acceptées par la Sécurité Sociale | 125% BR-SS | |
| Optique | Limitation bisannuelle pour les adultes et annuelle pour les enfants | |
| A - Verres simple foyer dont la sphère est comprise entre -6.00 et +6.00 dont le cylindre est inférieur ou égal à +4.00 | 50 € | 470 € dont 150 € maximum pour les montures |
| C - Verres simple foyer dont la sphère est supérieure à -6.00 et +6.00 ou dont le cylindre est supérieur à +4.00 et à des verres multifocaux progressifs | 200 € | 750 € dont 150 € maximum pour les montures |
| F - Verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est hors zone -8.00 à +8.00 ou à verres multifocaux ou progressifs sphériques dont le cylindre est hors zone -4.00 à +4.00 | 200 € | 850 € dont 150 € maximum pour les montures |
| B - Verres définis en A et en C | 125 € | 610 € dont 150 € maximum pour les montures |
| D - Verres définis en A et en F | 125 € | 660 € dont 150 € maximum pour les montures |
| E - Verres définis en C et en F | 200 € | 800 € dont 150 € maximum pour les montures |

*Tableau 1 : Planchers et plafonds du contrat responsable*

L'apparition du contrat responsable a également participé à un certain bouleversement de la tarification des régimes complémentaires santé collectifs. En effet, elle a entrainé la création d'une distinction entre les médecins CAS qui ont signé le contrat d'accès aux soins qui limite entre autres les dépassements d'honoraires et les médecins non CAS. Cette nouvelle contrainte augmente encore le temps de traitement des services de souscription qui doivent vérifier que chacune des garanties non CAS est inférieure d'au moins 20% de la base de remboursement que la même garantie pour les CAS tout en étant jamais inférieure au ticket modérateur et supérieure à 200% de la base de remboursement y compris le remboursement de la Sécurité Sociale.

# D) TARIFICATION DE PRODUITS COMPLÉMENTAIRES SANTÉ COLLECTIFS

## 1) ORGANISATION

### 1.1) L'ACTUARIAT

Le service actuariat est présent dans le processus de tarification dès le départ. Il analyse les spécificités de la consommation santé des bénéficiaires et propose des modèles adaptés.

Depuis l'apparition du *Big Data*, les actuaires ont tenté d'incorporer ces innovations technologiques et techniques dans le processus de tarification des complémentaires santé. De nouveaux modèles sont apparus, demandant d'autant plus de données et permettant une analyse encore plus fine du portefeuille.

Par exemple, le critère géographique ne s'analyse plus département par département mais en fonction du code postal ou du code INSEE de l'assuré en question. En effet, celui-ci va entre-autre permettre de connaitre la densité de dentistes ou d'ophtalmologistes dans la commune de résidence de l'assuré. Avec cette information, l'actuaire va pouvoir déterminer à quel point la présence de ces deux spécialistes dans la commune favorise la consommation médicale.

### 1.2) LE SERVICE DE SOUSCRIPTION

Le service de souscription est un service qui va tarifer « à la chaîne » des contrats complémentaires santé. C'est dans ce service que sont reçues les demandes de tarification pour le renouvellement d'un contrat ou encore pour la prospection.

Ces services utilisent les modèles actuariels définis par les actuaires pour tarifer chacune des entreprises qui sont ou qui peuvent entrer dans le portefeuille de l'organisme assureur. Ils doivent également vérifier les conditions d'acceptation de la tarification d'une complémentaire santé pour une entreprise à partir de la politique de souscription. Le caractère responsable des garanties santé proposées et les différentes recommandations et garanties minimales des conventions collectives nationales doivent être vérifiées. Pour parvenir à cela, ils ont besoin des informations sur les garanties à tarifier et aussi sur les salariés des entreprises en question ou encore sur le tenant des contrats actuels.

### 1.3) LE SERVICE COMMERCIAL

Le service commercial est là pour obtenir de nouvelles affaires, de nouveaux contrats que ce soit en affaire directe ou par le biais de courtiers en assurance. Ce sont eux qui fournissent aux souscripteurs

les informations nécessaires à la tarification. Ces informations capitales pour le tarificateur ne sont pas toujours des plus fiables et des plus complètes. Ainsi il n'est pas rare que le service commercial ne connaisse pas le nombre exact de salariés d'une entreprise, le nombre de bénéficiaires potentiels ou encore parfois les spécificités des garanties vendues ou attendues par l'entreprise. Ce sont donc des données tronquées, manquantes et pas toujours totalement fiables qui sont fournies au service souscription pour la tarification des contrats.

### 2) CONSÉQUENCE SUR LA TARIFICATION

L'explosion des demandes de tarification pour les services de souscription due à l'obligation de la mise en place d'une complémentaire santé et le manque de fiabilité de certaines données font qu'une erreur dans la tarification est vite arrivée.

Les modèles de plus en plus performants nécessitant des informations complètes sur les assurés en question semblent difficilement utilisables lorsqu'on ne possède que quelques chiffres clés sur des données agrégées de l'entreprise. La complexification des modèles tarifaires en santé collective ne va pouvoir s'appliquer que si les services commerciaux arrivent à obtenir les données correspondantes. Ainsi il semble trop ambitieux de vouloir acquérir le code INSEE de chacun des bénéficiaires d'un régime collectif pour la tarification d'une complémentaire santé.

Finalement, on peut remarquer un écart important entre les éléments attendus de l'actuariat pour créer des modèles permettant de connaitre avec précision le risque santé associé à une population et ce qui est techniquement accessible par les souscripteurs. D'autre part, des hypothèses doivent souvent être prises dans le cadre de données manquantes comme le pourcentage d'hommes ou le nombre moyen d'enfants. Lorsqu'on doit prendre des hypothèses dans une tarification, il serait souhaitable de savoir s'il n'y a pas de contradiction entre les tarifications passées et celles que l'on doit faire avec les données manquantes.

## E) MODÈLE TARIFAIRE SANTÉ

### 1) MODÈLES FRÉQUENCE * COÛT

Le modèle fréquence * coût est un des modèles les plus répandus en tarification santé. La mise en place de ce modèle consiste à étudier d'une part la fréquence des sinistres pour chacune des garanties et d'autre part la dépense engagée pour ce sinistre.

Mathématiquement, cela correspond à la formule suivante :

$$P_{a,i} = \mathbb{E}\left[\sum_{n=1}^{f_{a,i}} C_{a,i}^{n}\right]$$

Avec $P_{a,i}$ la prime pure associée à l'acte $a$ pour l'individu $i$, $f_{a,i}$ une variable aléatoire dans $\mathbb{N}$ représentant le nombre de consommations d'acte $a$ par l'assuré $i$ et enfin $C_{a,i}^{n}$ dans $\mathbb{R}^{+}$ représente le coût total de la n$^{ième}$ consommation de l'acte $a$ par l'assuré $i$.

Une fois la fréquence de consommation et les coûts déterminés, des correctifs sont appliqués pour permettre de mieux spécifier la consommation médicale d'une population.

Ainsi le modèle devient :

$$P_{a,i} = \mathbb{E}\left[\sum_{n=1}^{f_{a,i}} k_{a,i} C_{a,i}^{n}\right]$$

Avec $k_{a,i}$ le correctif de l'acte $a$ prenant en compte les caractéristiques de l'assuré $i$.

Les correctifs les plus courants sont des correctifs par rapport à l'âge, au sexe, à la catégorie socioprofessionnelle et au département de l'assuré.

Ainsi pour chacun des postes que l'assureur souhaite garantir, il devra déterminer un jeu de correctifs, un jeu de fréquences de consommation et un jeu de dépenses engagées pour pouvoir tarifer.

Ce modèle nous permet d'obtenir la prime pure et plus spécifiquement un coût adulte et un coût enfant pour une certaine population couverte. Ces deux coûts déterminés, nous pouvons ensuite calculer le budget technique :

$$Budget\ technique = Coût\ adulte \times Nombre\ d'adultes + Coût\ enfant\ \times Nombre\ d'enfants$$

Si l'on souhaite avoir un budget commercial, il faut rajouter la taxe sur les conventions d'assurance (TCA), la couverture maladie universelle (CMU) et les chargements.

$$Budget\ commercial = \frac{Budget\ technique \times (1 + TCA + CMU)}{1 - Chargements}$$

Un autre modèle est également couramment utilisé en tarification, il s'agit de la tarification par l'utilisation de barèmes. Ce cas ne sera pas développé dans ce mémoire.

## 1.1) OBTENTION D'UN TARIF UNIFORME

Le tarif uniforme se calcule comme explicité ci-dessous à partir des budgets commerciaux.

$$Tarif\ uniforme = \frac{Budget\ commercial}{Nombre\ de\ têtes\ payantes}$$

Le nombre de têtes payantes correspond à l'ensemble des salariés cotisants au régime. Ainsi, ne sont pas comptés dans le nombre de têtes payantes les personnes bénéficiant de ce régime à titre gratuit comme les personnes en CDD qui bénéficient du maintien de leur couverture pendant 12 mois (ou pendant la durée du contrat temporaire si elle est inférieure à un an) ou encore les licenciés de l'entreprise. Pour prendre en compte ces nouvelles spécificités qu'a apportées l'ANI, on applique un coefficient au tarif. Ce coefficient dépend entre autre de la durée des contrats précaires, de leur proportion par rapport à l'effectif total ou encore de la mise en place d'un plan de restructuration récent.

Le tarif uniforme est le plus favorable pour les familles car chacun des salariés va payer la même somme quelle que soit sa situation familiale ou le nombre de ses bénéficiaires.

D'autres formes de tarif existent comme par exemple de l'Adulte/Enfant, ou de l'Isolé/Famille.

Dans la suite nous nous intéresserons uniquement au tarif uniforme non taxé et hors chargements et nous n'intègrerons pas de majoration particulière par rapport à l'ANI.

# II) SÉLECTION DES VARIABLES PRINCIPALES

Pour pouvoir réduire la dimensionnalité des modèles tarifaires actuels, il faudra dans un premier temps déterminer quelles sont les variables qui ont le plus d'impact sur le tarif final. Pour cela nous allons utiliser l'algorithme CART qui va nous permettre d'obtenir des arbres de régression. Les différentes variables utilisées dans chaque nœud pour séparer les branches vont être considérées comme les variables principales pour la tarification. Une fois les résultats obtenus nous analyserons les différentes corrélations entre les variables conservées par l'algorithme. Pour finir, une étude par le biais de l'utilisation des GLM nous permettra de confirmer ou non l'importance des variables.

## A) DONNÉES

### 1) PRÉSENTATION DE LA BASE DE DONNÉES

La base de données utilisée correspond à l'historique des études santé tarifées par le cabinet ACTUARIS depuis 2009. Par études, il est fait référence à la démographie des entreprises ainsi que l'ensemble des garanties sur-mesure de celles-ci. Chaque tarification santé a été réalisée par un consultant puis validée par la suite par l'un de ses supérieurs ce qui garantit la fiabilité des données. D'autre part, les études de benchmarks et de tarification de contrats standards ont été retirées de la base de données.

Ainsi une ligne de cette base de données nous donne les informations sur l'entreprise que ce soit le pourcentage de mariés, le pourcentage d'hommes, le nombre moyen d'enfants, le pourcentage de cadres ou encore l'effectif et également les informations sur les garanties tarifées.

Les garanties sont séparées en plusieurs postes, Honoraires-Pharmacie, Hospitalisation, Dentaire, Optique, Autres actes remboursés par la Sécurité Sociale, Actes hors nomenclature et Spécifiques. Chacun de ces postes est composé de plusieurs garanties principales elles-mêmes divisées en plusieurs sous garanties (la liste complète des variables entrant dans le champ de la tarification se trouve en ANNEXE 2).

L'une des spécificités de cette base est qu'elle ne s'intéresse pas à une liste d'individus tête par tête mais aux caractéristiques générales de l'entreprise et aux spécificités de son contrat santé collectif. Nous nous sommes également intéressés aux tarifs. Par tarif, nous entendons tarif uniforme annuel en euros hors taxe et hors chargements.

Les différents tarifs ont été obtenus par le biais du logiciel ADDACTIS Prévoyance Office®.

Pour pouvoir utiliser chacune de ces études, une uniformisation des modèles de tarification a dû être effectuée. En effet, l'utilisation d'études sur plusieurs années est problématique notamment lorsque

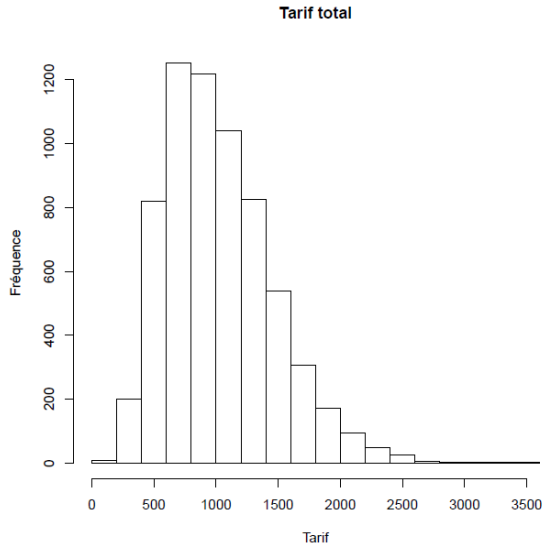
les garanties sont exprimées en pourcentage du plafond mensuel de la Sécurité Sociale qui augmente chaque année ou encore lorsque différents jeux de paramètres sont utilisés. Pour remédier à cela, une modification de l'historique des études a été réalisée par le biais du logiciel ADDACTIS Workflow® pour imposer l'année 2015 comme année de cotation ainsi que le jeu de paramètres actuel pour le modèle fréquence * coût.

Une chose importante concernant cette base de données est qu'elle est basée sur un historique de tarification. Ainsi des études ont été créées alors que le contrat responsable n'existait pas. Ainsi aucune distinction ne sera faite quant au contrat d'accès aux soins et par rapport à des minima et maxima de garanties.

## 2) QUELQUES STATISTIQUES SUR LA BASE DE DONNÉES

La base de données est composée de 6 561 études et de 149 variables.

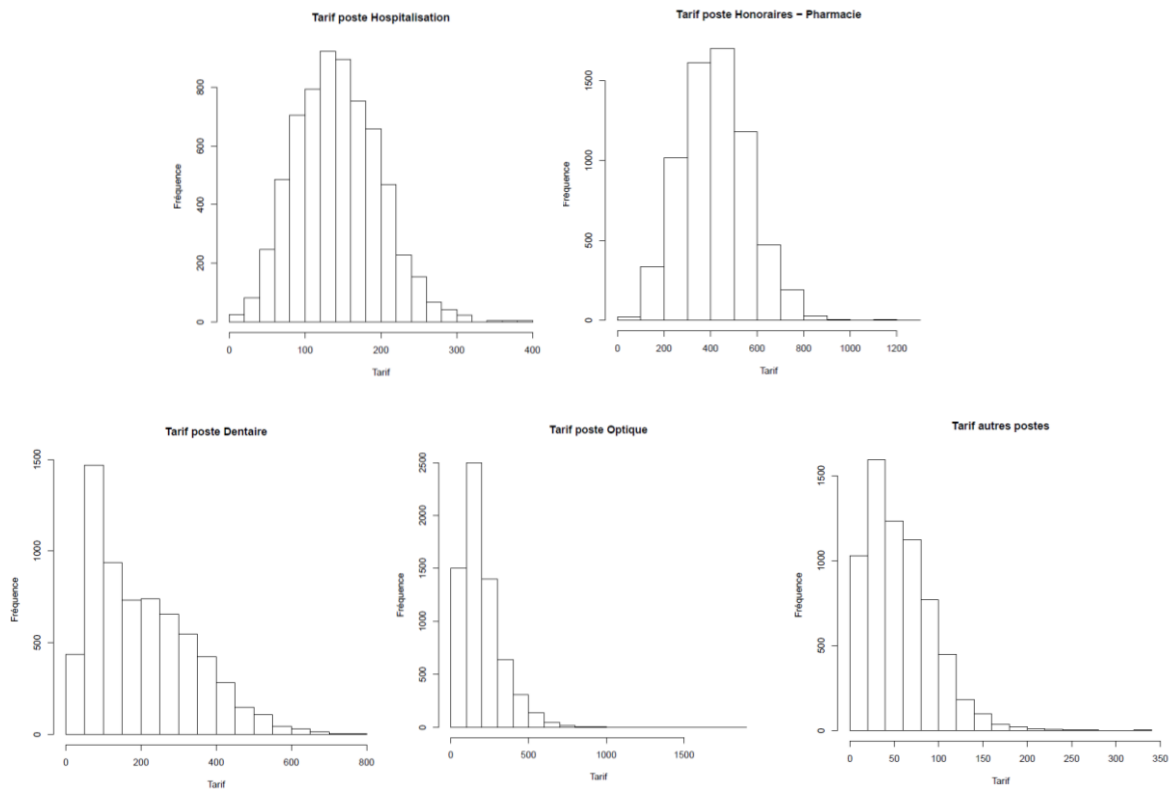
Nous obtenons également un tarif uniforme annuel global ainsi que sa décomposition en plusieurs postes : Honoraires-Pharmacie, Hospitalisation, Optique, Dentaire et Autres postes remboursés ou non par la Sécurité Sociale.




*Graphique 1 : Histogramme des tarifs totaux*

On peut observer quelques tarifs très faibles (inférieur à 300 € annuel) et très élevés (supérieur à 3 000 € annuel). D'autre part, 57% des tarifs sont compris entre 2 et 4% PMSS.

Dans la suite, on ne conservera pas les tarifs inférieurs à 300 € annuel pour ne pas prendre en compte des données aberrantes dans le modèle. Cette nouvelle contrainte retire 42 études de la base de données.




*Graphique 2 : Histogrammes des tarifs pour les différents postes*

L'étude des histogrammes des tarifs poste par poste montre que l'on peut les classer en deux catégories. La première regroupe le poste Honoraires – Pharmacie avec le poste Hospitalisation. En effet, on peut observer une répartition très proche avec peu de valeurs extrêmes.

Les autres garanties n'ont pas la même distribution. On peut observer des tarifs très bas et d'autres très élevés. Cela montre bien le caractère plus ou moins optionnel de ces garanties qui dans un contrat « entrée de gamme » sont au minimum et pour les contrats « très haut de gamme » peuvent rembourser les frais réels.

Pour mieux comprendre le type de garanties santé présentes dans la base de données, nous allons étudier les niveaux de garanties.

Il est possible de définir des niveaux de garanties pour les postes Honoraires-Pharmacie, Hospitalisation, Dentaire et Optique en s'intéressant au remboursement d'une seule garantie pour chacun de ces postes. Les garanties en question sont le remboursement des consultations de

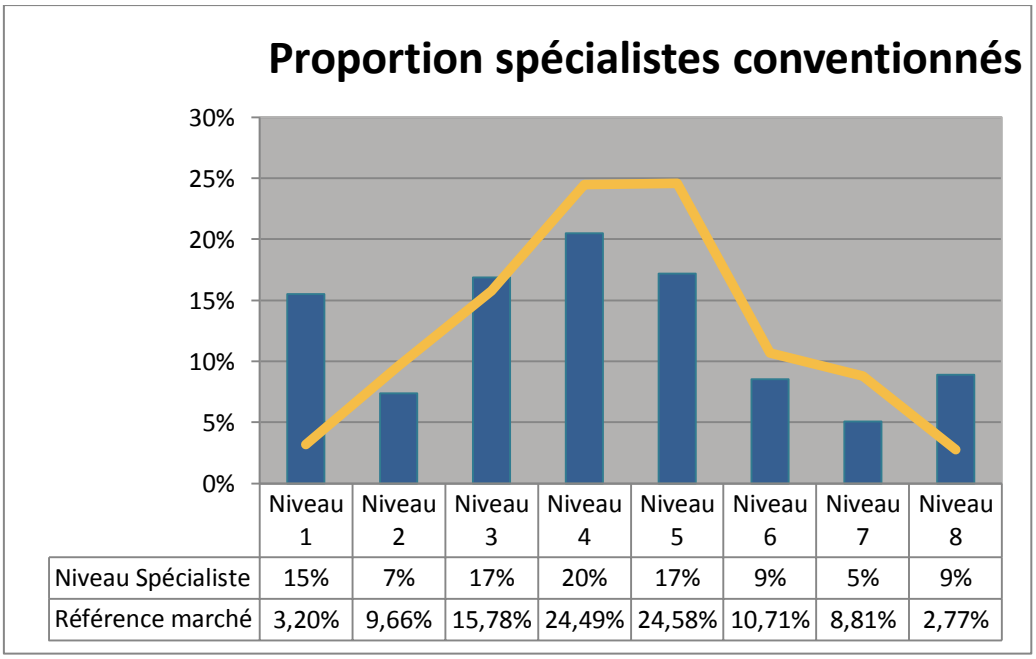
spécialistes conventionnés, le remboursement des honoraires d'hospitalisation en secteur conventionné, ceux des prothèses dentaires acceptées par la Sécurité Sociale et pour finir ceux des montures.

Les différents seuils des niveaux sont déterminés de la façon suivante :

| Actes | Moyen | Niveau 1 | Niveau 2 | Niveau 3 | Niveau 4 | Niveau 5 | Niveau 6 | Niveau 7 | Niveau 8 |
|---|---|---|---|---|---|---|---|---|---|
| Poids | | 3,20% | 9,66% | 15,78% | 24,49% | 24,58% | 10,71% | 8,81% | 2,77% |
| C&V de spécialistes | 245% BR-SS | 100% BR-SS | 125% BR-SS | 160% BR-SS | 200% BR-SS | 270% BR-SS | 360% BR-SS | 420% BR-SS | 485% BR-SS |
| Honoraires méd. Et Chir. | 295% BR-SS | 120% BR-SS | 170% BR-SS | 215% BR-SS | 285% BR-SS | 340% BR-SS | 360% BR-SS | 425% BR-SS | 490% BR-SS |
| Prothèses Acceptées | 360% BR-SS | 160%BR-SS | 245%BR-SS | 290%BR-SS | 340%BR-SS | 400%BR-SS | 440%BR-SS | 470%BR-SS | 490%BR-SS |
| Monture | 170 € | 60 € | 60 € | 125 € | 160 € | 195 € | 225 € | 245 € | 275 € |

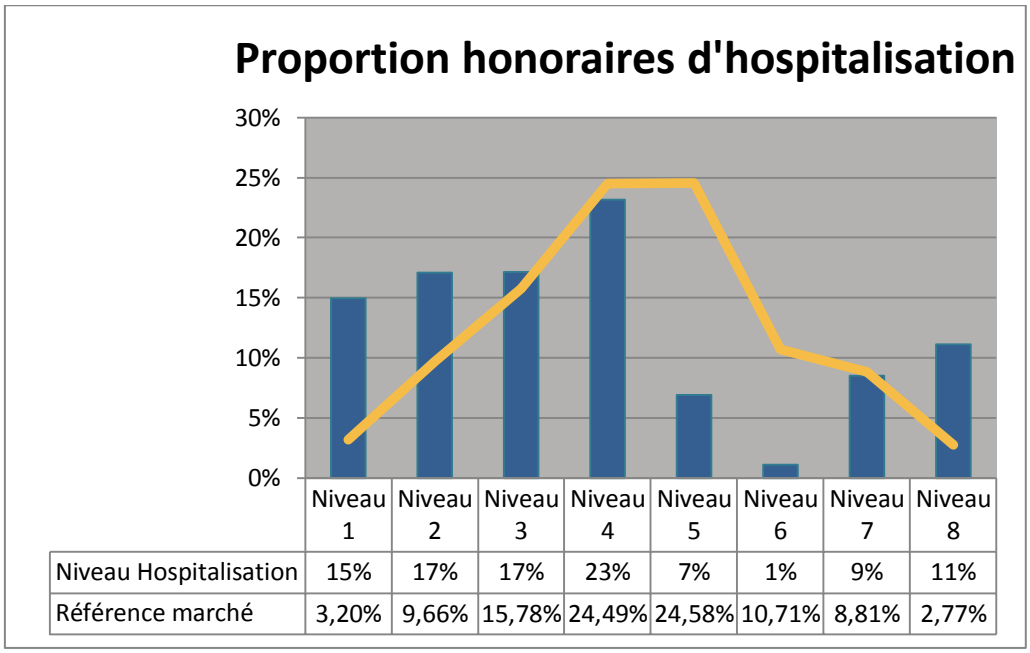
*Tableau 2 : Niveaux des garanties par poste*

Cette répartition va totalement être bouleversée par la mise en conformité des garanties au contrat responsable. En effet, actuellement le niveau moyen de remboursement pour les montures est de 170 € alors que le remboursement moyen de ce poste sera inférieur à 150 € lorsque l'ANI sera mis en place.




*Graphique 3 : Comparaison entre les niveaux de couverture des spécialistes et le marché de l'assurance collective*

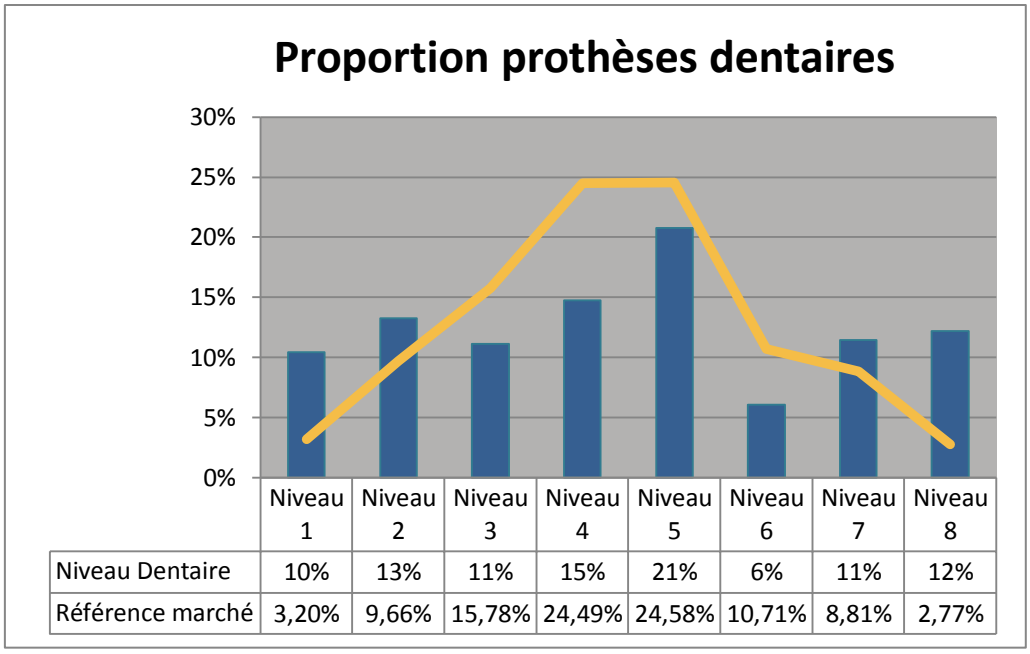
Les garanties spécialistes étudiées sont relativement proches de ce qui est observé sur le marché. On notera cependant une plus forte présence de garanties « entrée de gamme » et « très haut de gamme ».




| | Niveau 1 | Niveau 2 | Niveau 3 | Niveau 4 | Niveau 5 | Niveau 6 | Niveau 7 | Niveau 8 |
|---|---|---|---|---|---|---|---|---|
| Niveau Hospitalisation | 15% | 17% | 17% | 23% | 7% | 1% | 9% | 11% |
| Référence marché | 3,20% | 9,66% | 15,78% | 24,49% | 24,58% | 10,71% | 8,81% | 2,77% |

*Graphique 4 : Comparaison entre les niveaux de couverture des honoraires d'hospitalisation et le marché de l'assurance collective*

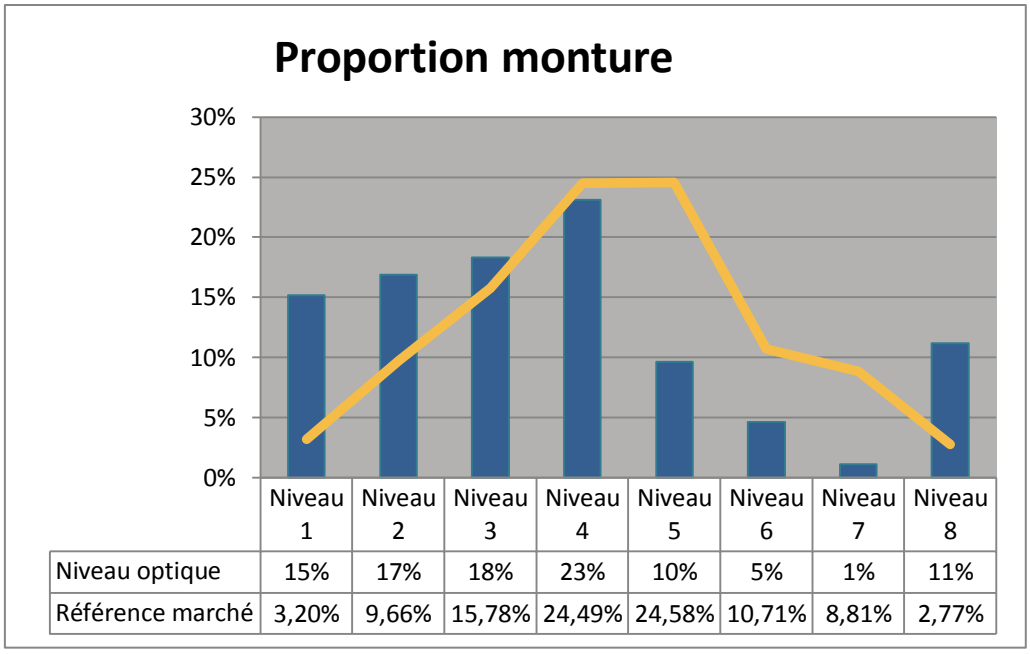
Comme pour la garantie spécialiste, la garantie honoraires d'hospitalisation est majoritairement « entrée de gamme ». Les tarifications de contrats « très haut de gamme » sont supérieures à la moyenne du marché. Un autre point à prendre en compte est la représentation dix fois plus faible du niveau 6 par rapport au marché.




| | Niveau 1 | Niveau 2 | Niveau 3 | Niveau 4 | Niveau 5 | Niveau 6 | Niveau 7 | Niveau 8 |
|---|---|---|---|---|---|---|---|---|
| Niveau Dentaire | 10% | 13% | 11% | 15% | 21% | 6% | 11% | 12% |
| Référence marché | 3,20% | 9,66% | 15,78% | 24,49% | 24,58% | 10,71% | 8,81% | 2,77% |

*Graphique 5 : Comparaison entre les niveaux de couverture des prothèses dentaires et le marché de l'assurance collective*

La répartition des prothèses dentaires est relativement stable d'un niveau à l'autre et ne respecte pas la répartition du marché. Ainsi, les résultats obtenus pour les garanties dentaires dans la suite de l'étude pourront être très différents sur un autre portefeuille qui sera lui plus proche de ce que l'on observe sur le marché.




| | Niveau 1 | Niveau 2 | Niveau 3 | Niveau 4 | Niveau 5 | Niveau 6 | Niveau 7 | Niveau 8 |
|---|---|---|---|---|---|---|---|---|
| Niveau optique | 15% | 17% | 18% | 23% | 10% | 5% | 1% | 11% |
| Référence marché | 3,20% | 9,66% | 15,78% | 24,49% | 24,58% | 10,71% | 8,81% | 2,77% |

*Graphique 6 : Comparaison entre les niveaux de couverture des montures et le marché de l'assurance collective*

La répartition du remboursement des montures ressemble beaucoup à ce que l'on observe pour la garantie honoraires d'hospitalisation.

D'autre part, on peut remarquer pour l'optique que le niveau 1 est cinq fois plus représenté dans la base de données comparativement aux niveaux de couverture observés sur tout le marché de l'assurance et que les garanties « haut de gamme » (niveaux 5, 6 et 7) sont très peu représentées. En revanche le niveau « très haut de gamme » avec un remboursement en frais réels est quatre fois plus représenté que sur le marché.

Ces différents graphiques montrent bien que les données auxquelles nous nous intéressons ne représentent pas le marché de l'assurance collective dans sa globalité. En effet, il y a beaucoup plus de garantie « entrée de gamme » et « très haut de gamme » dans cette base de données que sur le marché actuel. Un point important est à prendre en compte, ces données ne prennent pas en considération les bouleversements dans les garanties santé qu'ont apportés la mise en place de l'ANI et la généralisation des couvertures santé collectives à tous les salariés.

## 3) PARTICULARITÉS DES GARANTIES ETUDIÉES

Les garanties étudiées possèdent une granularité dépendant du suivi ou non du parcours de soins ainsi que de la lettre clé de l'acte médical.

### 3.1) PARCOURS DE SOINS

Le parcours de soins a été mis en place pour permettre à chaque personne de bénéficier de soins coordonnés et pour respecter un protocole de soins. Il se traduit par la déclaration d'un médecin traitant auprès de l'assurance maladie et de la nécessité d'obtenir de sa part des ordonnances avant de consulter un spécialiste.

Cela a également pour but de diminuer les abus et le surplus de consommation médicale.

Au niveau des garanties des complémentaires santé, cela se traduit par la différenciation des remboursements de soins en fonction du respect ou non du parcours de soins.

La base de données différencie les garanties en fonction du respect de ce parcours de soins pour les garanties consultations et visites de généralistes et de spécialistes.

### 3.2) LETTRES CLÉS

Au vu de la multitude d'actes en santé, des lettres clés ont été introduites pour codifier chacun de ces actes. La base de données segmente les garanties hospitalisation-chirurgie et petite chirurgie en fonction de cinq lettres clés.

Ces lettres sont :

- ADA : Actes d'anesthésie
- ADC : Actes de chirurgie
- ATM : Actes techniques médicaux
- ADE : Actes d'échographie
- ACO : Actes d'obstétrique

Cette distinction a été effectuée pour les actes conventionnés et non conventionnés.

## 4) DÉFINITION D'UN ÉQUIVALENT EURO DES GARANTIES

L'une des difficultés de la tarification santé est le nombre important de garanties possibles et imaginables. Ceci est dû au fait que l'on a des garanties qui peuvent s'exprimer y compris les remboursements de la Sécurité Sociale ou en complément du régime obligatoire et d'autre part au nombre important d'assiettes disponibles.

Les différentes assiettes en tarification santé sont :

- €
- % Plafond annuel/mensuel de la Sécurité Sociale (%PASS/%PMSS)

    - La valeur de ce plafond est revue annuellement à la hausse
- % Montant remboursé (%MR)
    - Montant que le régime de base a remboursé à l'assuré
- % Remboursement de la Sécurité Sociale (%RSS)
    - Montant remboursé par la Sécurité Sociale
- % Remboursement de la Sécurité Sociale Reconstitué (%RSSR)
    - Montant qu'aurait remboursé la Sécurité Sociale si l'acte en question avait été effectué par un organisme conventionné
- % Base de Remboursement (%BR)
    - Tarif appliqué par l'Assurance Maladie pour le remboursement de chaque acte médical, la base de remboursement est égale à 0 € pour les actes non conventionnés
- % Tarif de Convention (%TC)
    - Pour les actes en secteur conventionné on a %TC=%BR, pour les actes en secteur non conventionné, utiliser cette assiette revient à considérer le même acte en secteur conventionné avec la base de remboursement
- % Ticket modérateur (%TM)
    - Correspond à la base de remboursement moins le remboursement de la Sécurité Sociale

$$TM = BR - RSS$$

- % Tarif de responsabilité
    - Base tarifaire retenue par l'Assurance Maladie
- % Dépassement
    - Montant dépassant le tarif de convention
- % Frais Réels
    - Coût global de la prestation

Une garantie peut également être définie par plusieurs de ces assiettes, par exemple 90% des frais réels dans la limite de 400% BR.

On peut schématiser un remboursement de soin de la façon suivante :

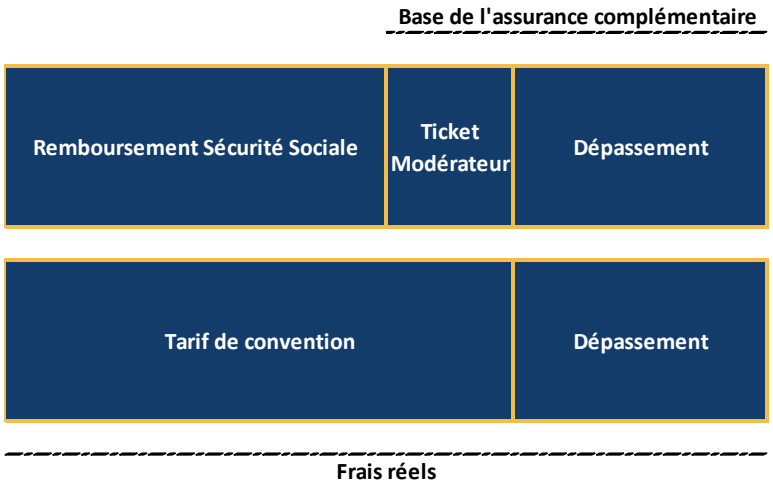



*Figure 5 : Schématisation des remboursements de soins*

L'ensemble de ces éléments font référence à un certain montant en euros. La façon la plus rapide pour comparer des garanties santé est donc de toutes les transposer en euros et en complément des remboursements de la Sécurité Sociale. Sans cela, une étude comparative des niveaux de couverture est complexe.

Ainsi, une première étape et non des moindres de ce mémoire a été le calcul des équivalents euro pour toutes les garanties existantes. Par la suite seul un exemple basé sur une garantie simple sera exposé ainsi que les postes où un retraitement particulier a été nécessaire.

Exemple : la consultation d'un généraliste conventionné

Pour comprendre comment ont été calculés les équivalents euro des garanties, on peut étudier le cas des généralistes conventionnés.
La base de remboursement est de 23 €.

| Garantie | Equivalent € |
|---|---|
| 5% PMSS complément SS | 5% * 3170 = 158.50 € |
| 5% PMSS y compris SS | 5% * 3170 - 70% * 23 = 142,40 € |
| 200% MR complément SS | 2 * 70% * 23 = 32.20 € |
| 200% MR y compris SS | 2 * 70% * 23 - 70% * 23 = 16.10 € |
| 200% BR complément SS | 200% * 23 = 46 € |
| 200% BR y compris SS | 200% * 23 - 70% * 23 = 29.90 € |
| 100% TM en complément SS | 100% * (100% - 70%) * 23 € = 6.90 € |

*Tableau 3 : Garanties et remboursements en euros associés*

Les participations forfaitaires n'ont pas été prises en compte dans le calcul de l'équivalent euro des garanties. En effet, comme dans la suite on ne va s'intéresser qu'à comparer une garantie avec une autre, si leur équivalent euro a été augmenté de la même valeur cela ne change rien au résultat.

## 4.1) CAS PARTICULIERS : LES FRAIS RÉELS ET LES DÉPASSEMENTS

Deux types de garantie ne peuvent pas être reliées aux remboursements de la Sécurité Sociale. Il s'agit des garanties en pourcentage des frais réels et en pourcentage des dépassements. Pour résoudre ce problème lorsque l'on se trouve dans le cadre d'un modèle de type fréquence * coût, nous utiliserons la dépense engagée maximale comme le montant en euros équivalent à 100% des frais réels. Ainsi une garantie à 80% des frais réels correspondra au montant en euros suivant :

$$80\%\ Frais\ réel = Dépense\ engagée\ maximale\ \times 80\% - Remboursement\ de\ la\ SS$$

Une garantie égale à 80% des dépassements s'écrit de la même manière :

$$80\%\ Dépassement = 80\%\ (Dépense\ engagée\ maximale - Tarif\ de\ convention)$$

Ces hypothèses peuvent paraitre fortes mais elles sont nécessaires pour montrer le caractère haut de gamme de ces garanties.

## 4.2) RETRAITEMENT DU POSTE VERRES

Le remboursement des verres est une garantie très importante en tarification santé collective. C'est un élément qui permet de dire si un contrat est de niveau entrée, milieu ou haut de gamme. Depuis la mise en place du contrat responsable, l'intitulé des garanties en fonction des types de verres s'uniformise par le biais de la grille LPP (ANNEXE 4). Mais avant cela il y avait de nombreuses façons de présenter cette garantie. Les plus courantes étaient composées d'un forfait global verres et monture, d'une distinction selon le type de correction, selon la puissance de la correction ou tout simplement par verre quelle que soit la correction.

Pour uniformiser les remboursements, nous utilisons les statistiques de consommation de chaque type de verres pour donner finalement une estimation de la garantie sous le même format.

Dans un souci de simplification des modèles, nous ne nous sommes pas intéressés à la garantie verres pour les enfants.

| | Consommation |
|---|---|
| Verres unifocaux | 64,30% |
| Verres multifocaux | 35,20% |
| Verres hypercomplexes | 0,50% |
| Verre faible correction | 62,40% |
| Verre moyenne correction | 37,10% |
| Verres forte correction | 0,50% |

*Tableau 4 : Répartition de la consommation de verres*

Ainsi, on obtient un montant de remboursement global pour les verres ce qui facilitera le travail.

Exemple : Forfait au global verres et monture

Si l'on s'intéresse à un forfait regroupant les garanties verres et monture, un coefficient de répartition doit être pris en considération pour la tarification.

Garantie : 15% PMSS en complément de la SS

$$Equivalent\ €\ Verres = 15\% \times 3170 \times coefficient\ de\ répartition$$

Le coefficient de répartition correspond à la part du forfait attribuée à la garantie verres. La plupart du temps ce coefficient est fixé à 50% mais il peut être modifié en fonction du dossier.

Exemple : Différentiation par type de verres

Garantie : 3,50%PMSS en complément de la Sécurité Sociale pour les verres unifocaux et 6,50% PMSS en complément de la Sécurité Sociale pour les verres multifocaux et hypercomplexes.

$$Equivalent\ €\ Verres = (3{,}50\% \times 64{,}30\% + 6{,}50\% \times 35{,}20\% + 6{,}50\% \times 0{,}50\%) \times 3170 = 144{,}90\ €$$

Grâce au calcul des équivalents euro pour chacune des garanties, il est possible de comparer directement l'ensemble des garanties de toutes les études de la base de données.

## B) ALGORITHME CART

L'algorithme CART (*Classification And Regression Tree*) est un algorithme non paramétrique fonctionnant par apprentissage statistique. Il permet entre autre d'établir quelles sont les variables les plus importantes dans la modélisation. Dans la suite, nous allons nous intéresser exclusivement au traitement de variables continues et donc aux arbres de régression.

# 1) ARBRE DE RÉGRESSION

## 1.1) BUT

Par la suite on considèrera que l'on souhaite expliquer une variable continue $Y$ en fonction de $m$ variables continues $X_1, X_2, \dots, X_m$.

Il existe une fonction $\bar{f}$ mesurable telle que :

$$\mathbb{E}[Y|X_1, \dots, X_m] = \bar{f}(X_1, \dots, X_m)$$

Ainsi on peut théoriquement prédire la variable $Y$ en fonction des données fournies en entrée. De ce fait, on obtient une estimation de la fonction $\bar{f}$ comme étant :

$$\bar{f} = \underset{f \in \mathcal{M}(\mathbb{R}^m, \mathbb{R})}{\operatorname{Argmin}} \mathbb{E}[(Y - \bar{f}(X))^2]$$

Le but principal sera donc de déterminer une fonction approximative de $\bar{f}$ notée $\hat{f}$ qui permette d'expliquer au mieux la variable d'entrée $Y$. L'algorithme CART va permettre de déterminer une estimation de cette fonction par le biais d'arbres.

## 1.2) LECTURE D'UN ARBRE

La lecture d'arbre est très aisée ce qui explique pourquoi cette méthode est appréciée. Chaque arbre est composé de nœuds et de branches. A chaque nœud, on retrouve une condition sur une des variables explicatives puis l'apparition de deux branches qui vont segmenter la base de données en fonction de la condition présente dans le nœud. Lorsqu'une branche ne se termine pas par un nœud, on trouve ce que l'on appelle une feuille.

Ci-dessous un exemple pour mieux visualiser :

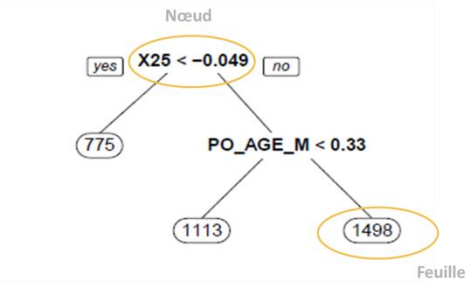



*Figure 6 : Arbre de régression avec deux nœuds*

Cet arbre s'analyse de la façon suivante, la variable principale segmentant la base de données est celle du premier nœud et correspond à la variable X25 centrée et réduite. Ce code correspond plus précisément à la garantie Orthodontie remboursée par la Sécurité Sociale. Une fois cette condition sur la garantie effectuée on se retrouve avec deux groupes de variables. Le groupe ne respectant pas la condition du premier nœud est soumis à une autre condition (cette fois sur l'âge moyen centré réduit de l'entreprise) ce qui permet d'aboutir à une nouvelle partition.

## 1.3) THÉORIE DE L'ALGORTIHME CART

L'algorithme CART permet de créer des régions de l'espace où la valeur de la fonction recherchée est constante. Cette méthode est utile pour deux raisons. La première est qu'elle permet d'avoir un aperçu de comment la variable à prédire dépend des variables explicatives. La deuxième est que quand il y a beaucoup de variables explicatives, comme dans la détermination d'un tarif en santé par exemple, cette méthode permet de les classer par ordre d'importance. Dans ce mémoire, l'utilité principale de cet algorithme sera de donner les bonnes variables à intégrer dans un algorithme d'apprentissage.

Pour plus de détails sur la théorie de l'algorithme CART, le lecteur intéressé pourra se référer à HASTIE T & AL [2009].

### 1.3.1) CONSTRUCTION DE L'ARBRE

Dans un premier temps, pour chacune des variables explicatives $X_i$ on détermine un point $x_i^1$ qui partitionne l'espace de cette variable en deux. Le choix de $x_i^1$ est réalisé de la manière suivante, pour chacune des valeurs de l'espace de cette variable, on détermine deux constantes $c_i^{1,1}$ et $c_i^{1,2}$ telles que :

$$f(x) = c_i^{1,1} \text{ si } x_i^1 > x \text{ et } f(x) = c_i^{1,2} \text{ si } x_i^1 < x$$

et on conserve le $x_i^1$ qui permet une meilleure estimation des données.

Dans la suite, par amélioration des estimations, on entendra diminution du RMSE (*Root mean square error)* avec :

$$RMSE(\hat{\theta}) = \sqrt{\mathbb{E}((\hat{\theta} - \theta)^2)}$$

En notant $\theta$ la valeur exacte de la fonction et $\hat{\theta}$ son estimation.

Cette recherche de $x_i^1$ est réalisée pour les $i$ variables explicatives. Le premier critère de séparation sera donc celui qui parmi toutes les variables explicatives minimise la valeur du RMSE. On obtient ainsi une première partition de l'espace.

On réalise ensuite la même étude sur chacun des deux sous espaces créés pour obtenir la prochaine segmentation.

### 1.3.2) NOTATION

Voici les différentes notations utilisées dans la suite :

- $R$ : nœud principal comportant l'ensemble des données
- $A$ : l'arbre modélisé
- $n$ : la taille de l'échantillon initial
- $K$ : le nombre de classes
- $N_k$ le nombre d'observations de la classe k
- $N(t)$ le nombre d'observations dans le nœud t
- $N_k(t)$ le nombre d'observations de la classe k dans le nœud t

Si l'on s'intéresse aux probabilités conditionnelles, plus particulièrement à celle d'appartenir à la classe $k$ sachant qu'on appartient au nœud $t$, on a la relation suivante :

$$p(k|t) = \frac{N_k(t)}{N(t)}$$

### 1.3.3) FONCTIONNEMENT DE L'ALGORITHME

L'objectif final est de créer un nombre $K$ de classes en fonction des valeurs de l'ensemble des variables explicatives. Pour pouvoir faire cela il faut définir plusieurs critères qui montrent qu'une partition de l'espace apporte de l'information supplémentaire.

#### 1.3.3.1) HÉTÉROGÉNITÉ DU MODÈLE

Pour déterminer l'hétérogénéité du modèle, on fait appel à une fonction $h$ appelée fonction d'hétérogénéité. C'est une fonction réelle définie sur un ensemble fini de probabilités discrètes telle que :

$$h : (p_1, p_2, \dots, p_k) \to h(p_1, p_2, \dots, p_k)$$

Une fonction d'hétérogénéité est symétrique par rapport à chacune des variables $p_i$, son maximum est atteint lorsqu'il y a équiprobabilité

$$\text{Argmax}(p_1, \dots, p_k) = (\frac{1}{k}, \dots, \frac{1}{k})$$

Et son minimum est atteint pour les éléments de la base canonique, c'est-à-dire lorsqu'il n'y a plus qu'une observation par partition ou plusieurs observations identiques.

$$\text{Argmin}(p_1, \dots, p_k) = (e_1, \dots, e_k)$$

avec $(e_1, \dots, e_k)$ les éléments de la base canonique de $\mathbb{R}^k$. Grâce à cette fonction, on va pouvoir déterminer la qualité de division d'un nœud en calculant son hétérogénéité.

L'hétérogénéité d'un nœud $t$ est définie à partir de la fonction d'hétérogénéité :

$$\text{hét(t)} = h(p(1|t), \dots, p(k|t))$$

L'hétérogénéité de l'arbre $A$ est définie comme étant la somme de l'hétérogénéité pondérée par les probabilités d'appartenance à un nœud de l'ensemble des nœuds de cet arbre.

$$\text{Hét(t)} = \sum_{t \in A} p(t)\ \text{hét(t)}$$

Plusieurs fonctions d'hétérogénéité existent pour la construction d'un arbre de régression. Les plus connues sont l'entropie de Shannon et l'indice d'inégalité de Gini définies comme suit :

Entropie de Shannon :

$$\text{hét}_{Shannon}(t) = -\sum_{k=1}^{K} p(k|t) \times \ln(p(k|t))$$

Indice de Gini :

$$\text{hét}_{Gini} = \sum_{k \neq i}^{K} p(j|t) \times p(k|t) = 1 - \sum_{i=1}^{K} p(i|t)^2$$

Par la suite, nous conserverons l'indice de Gini comme fonction d'hétérogénéité du modèle.

### 1.3.3.2) CRÉATION D'UN NŒUD

A chaque nœud, l'algorithme doit déterminer la variable explicative qui va servir de critère de séparation et la valeur correspondante. Plus simplement, pour chaque variable explicative, l'algorithme va déterminer le seuil permettant de réduire au maximum l'hétérogénéité de l'arbre.

Puis pour choisir finalement la condition de séparation du nœud, on conserve la variable qui minimise le plus l'hétérogénéité.

Mathématiquement parlant, cela revient à chercher :

$$\delta_t = \text{Argmax}\ (\partial\text{hét}(\delta, t))$$

Ainsi réduire l'hétérogénéité au niveau d'un nœud revient à réduire l'hétérogénéité au niveau de l'arbre entier. On obtient également grâce à ce processus une classification des variables en fonction de leur importance dans l'explication de la variable réponse.

### 1.3.3.3) ARRÊT DE L'ALGORITHME

Si l'on ne s'intéresse qu'à la réduction de l'hétérogénéité, l'algorithme aboutit finalement à une partition totale de l'ensemble des variables, c'est-à-dire que chaque classe finale ne contient qu'une seule observation. Pour résoudre cela, des critères d'arrêt peuvent être mis en place. Les plus courants sont :

- La mise en place d'une limite dans le nombre de feuilles
- La mise en place d'un nombre d'éléments minimal dans chacune des classes
- Lorsque pour tout nœud $t$ on a :

$$\max\big(\partial\text{hét}(\delta, t)\big) = c$$

avec $c$ une constante réelle.

La mise en place de l'un de ces critères est donc nécessaire pour pouvoir obtenir des classes avec suffisamment d'observations en sortie.

### 1.3.3.4) ERREUR DE PRÉDICTION

Plusieurs estimateurs de l'erreur de prédiction sont à notre disposition, le plus simple est celui qui fait appel à l'estimation par substitution.

Cette méthode consiste à calculer la proportion des observations de l'échantillon qui est mal classée dans la procédure.

C'est-à-dire la valeur :

$$\tau(y) = \mathbb{P}(y(X, E) \neq Y)$$

Si l'on note $y(., E)$ une procédure de classification de l'ensemble des éléments de l'espace $E$. L'estimation par substitution consiste à estimer la probabilité d'être mal classé $\tau$ par :

$$\hat{\tau}(y) = \frac{1}{N} \sum_{(x_n,k_n)\in E} 1_{\{y(x_n,E)\neq k_n\}}$$

Le problème de cette méthode est qu'elle utilise le même échantillon que celui qui a servi à la classification et de ce fait une surestimation des performances peut avoir lieu.

Une autre méthode pour pallier cela est l'estimation par utilisation d'un échantillon témoin. On compare les différentes proportions des observations dans l'échantillon témoin et dans l'arbre final pour définir l'erreur de prédiction. Dans notre cas particulier, cette méthode ne peut pas être mise en œuvre car elle nécessite une autre base de données comparative possédant plus de lignes.

### 1.3.3.5) ÉLAGAGE

L'élagage est une phase nécessaire pour réduire le nombre de feuilles d'un arbre en supprimant les sous arbres se trouvant après un nœud $t$.

Par la suite on notera $A^{Elag}$ l'arbre élagué. Pour qu'il soit plus intéressant de conserver l'arbre élagué, il faut que

$$\widehat{\tau(t)} > \widehat{\tau(A^{Elag})}$$

Pour défavoriser les arbres ayant un nombre trop important de feuilles, l'idée est d'ajouter un paramètre de complexité pour prendre en compte la taille de l'arbre, ainsi on introduit un coefficient positif $\alpha$ tel que

$$\widehat{\tau_\alpha(A)} = \widehat{\tau(A)} + \alpha \times \#(A)$$

Avec $\#(A)$ le nombre de feuilles de l'arbre $A$.

Ainsi un élagage est réalisé pour un certain nœud $t$ si l'erreur en tant que nœud terminal de $t$ est plus faible que si l'on conservait les branches sortant du nœud $t$.

$$\widehat{\tau_\alpha(A_t^{Elag})} < \widehat{\tau_\alpha(A_t)}$$

Ce qui implique :

$$e(t) = \frac{\widehat{\tau(A_t^{Elag})} - \widehat{\tau(A_t)}}{\#(A_t) - 1} \leq \alpha$$

Pour déterminer le meilleur arbre, on utilise ce qu'on appelle l'algorithme de coupe du maillon faible. Cet algorithme fonctionne en plusieurs étapes.

La première consiste à créer l'arbre maximum noté $A^{max}$ , chacune des feuilles de cet arbre ne contient qu'une seule observation.

Ensuite, on calcule la fonction d'élagage $e(t)$ pour chacun des nœuds $t$. Le premier élagage aura lieu pour le nœud qui minimise la fonction $e(t)$. On obtient ainsi un arbre réduit.

On réitère ensuite le processus jusqu'à ce que l'erreur de prédiction ne se réduise plus ou trop faiblement à partir d'un critère défini à l'avance. On obtient ainsi l'arbre élagué.

#### 1.3.3.6) VALIDATION CROISÉE

Pour choisir le meilleur arbre, une validation sur un échantillon de test est nécessaire. On recherche l'arbre qui minimise l'erreur de prédiction sur cet échantillon, c'est-à-dire en réduisant la proportion de données mal classées.

Statistiquement, on cherche à minimiser la valeur suivante qui correspond à la proportion d'individus mal classés :

$$R^{El}(A_l) = \sum_{i,j} 1_{\{i \neq j\}} \frac{N_{i,j}}{N}$$

Avec $N$ le nombre total de données dans l'échantillon et $N_{i,j}$

Cette validation consiste à retirer une partie de la base de données avant la phase d'entrainement. Ensuite, une fois l'entrainement réalisé et l'arbre créé, les données non utilisées servent à tester la performance du modèle sur de nouvelles données.

D'autres méthodes de validation croisée existent. Parmi elles, il y a plus particulièrement la méthode dite *Leave-one-out cross validation*. Cette méthode est un cas extrême de la méthode précédente. En effet, au lieu de séparer la base de données en deux selon une certaine proportion, l'entrainement aura lieu sur toutes les valeurs sauf une. La prédiction sera ensuite uniquement réalisée sur le point n'ayant pas servi à l'apprentissage. L'erreur moyenne obtenue sur l'ensemble des points de la base de données permet de connaitre la qualité finale du modèle. Cette méthode permet d'obtenir des résultats plus fiables mais demande un temps de calcul beaucoup plus long que la précédente.

### 1.4) IMPORTANCE RELATIVE DES VARIABLES

Il peut être utile de pouvoir connaitre les variables ayant le plus d'importance dans la classification des observations. En effet, cette connaissance peut donner des informations sur la contribution ou encore sur la nature de la relation entre une variable explicative et la variable à expliquer.

On rappelle que $\hat{f}$ est l'approximation de la fonction à expliquer, on note $\boldsymbol{x}$ un vecteur de variables explicatives et $x_j$ la j$^{ième}$ variable explicative, alors d'après J. FRIEDMAN [2001] on peut écrire l'importance relative de la variables $x_j$ pour un arbre de régression de la manière suivante :

$$I_j = \left( \mathbb{E}_x \left[ \frac{\partial \hat{f}(x)}{\partial x_j} \right] . Var_x(x_j) \right)^{\frac{1}{2}}$$

Une des caractéristiques des arbres de régression est de fournir des estimations constantes sur des parties de l'espace des variables explicatives. De ce fait, la fonction précédente n'existe pas réellement et doit être approximée par :

$$\hat{I}_j^2(Arbre) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 \, 1_{\{v_t = j\}}$$

$J$ indique le nœud terminal de l'arbre, $v_t$ correspond au critère de différentiation du nœud $t$ et $\hat{\imath}_t^2$ l'amélioration empirique de l'erreur quadratique définie de la sorte :

$$\hat{\imath}_t^2 = \frac{\omega_d \omega_g}{\omega_d + \omega_g} (\bar{y}_d - \bar{y}_g)^2$$

$\omega_d$ correspond à la probabilité de se trouver dans la branche droite du nœud $t$ du sous arbre étudié et $\omega_g$ celle de se trouver dans la branche gauche.

$\bar{y}_d$ représente la valeur de la feuille droite de l'arbre étudié et $\bar{y}_g$la valeur de la feuille gauche. Ainsi $(\bar{y}_d - \bar{y}_g)^2$ est un moyen de mesurer l'amélioration de l'estimation des variables par le biais de la création d'une nouvelle séparation de l'espace.

Le schéma suivant permet de mieux visualiser le calcul de l'importance des variables. La valeur $\bar{y}_g$ correspond à la valeur de la feuille gauche du sous arbre $t$ avant la nouvelle séparation.

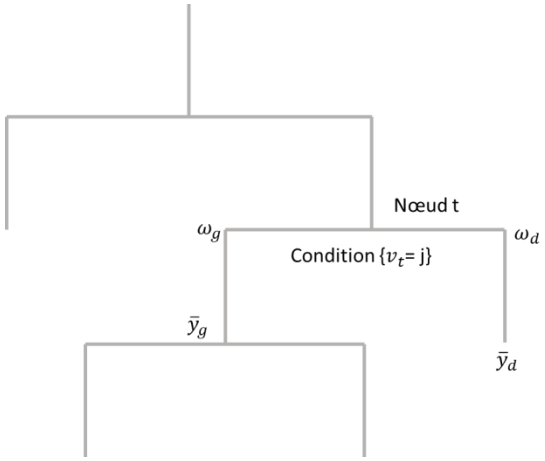



*Figure 7 : Schématisation du calcul de l'importance des variables dans un arbre de régression*

## 1.5) AVANTAGES ET INCONVÉNIENTS DE L'ALGORITHME CART

L'avantage principal de l'algorithme CART est sa simplicité de mise en place et sa capacité à s'adapter à de nombreux types de données. Son principal défaut est son manque de robustesse. En effet, la prédiction dépend largement de l'échantillon d'apprentissage. La prédiction sur des données très éloignées de celles ayant servies à l'apprentissage entraine des erreurs d'estimation importantes. D'autre part, si l'on modifie une seule variable, c'est l'ensemble de l'arbre qui va être modifié.

## 1.6) APPLICATION DE L'ALGORITHME CART AUX DONNÉES SANTÉ

Nous allons utiliser cette méthode sur les données dont nous disposons non pas pour réaliser des prédictions sur les tarifs, mais pour déterminer quelles sont les variables les plus importantes pour la tarification d'un régime complémentaire santé sur-mesure, que ce soit au niveau des garanties ou au niveau des informations relatives à la population couverte.

Pour cela, nous allons utiliser le logiciel R pour obtenir les différentes modélisations.

Les arbres créés devront retrouver la valeur en euros du tarif uniforme hors taxe et hors chargement. En entrée, on retrouve l'ensemble des équivalents euro des garanties et les caractéristiques démographiques des entreprises. La fonction d'hétérogénéité choisie est l'indice de Gini. Au vu de la taille des arbres, le détail des différentes conditions dans chaque nœud n'a pas été affiché sur les graphiques. Dans un premier temps 70% de la base de données servira à l'apprentissage et 30% à la phase de test.

On obtient ainsi l'arbre maximal suivant sur la base d'apprentissage.

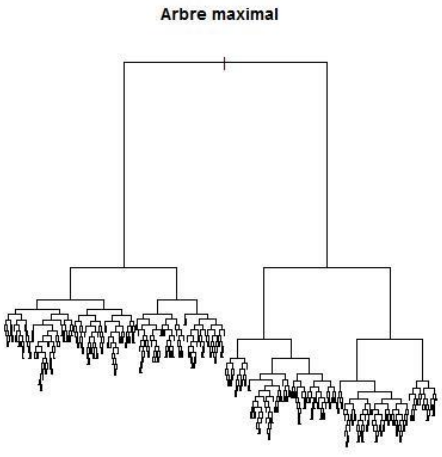
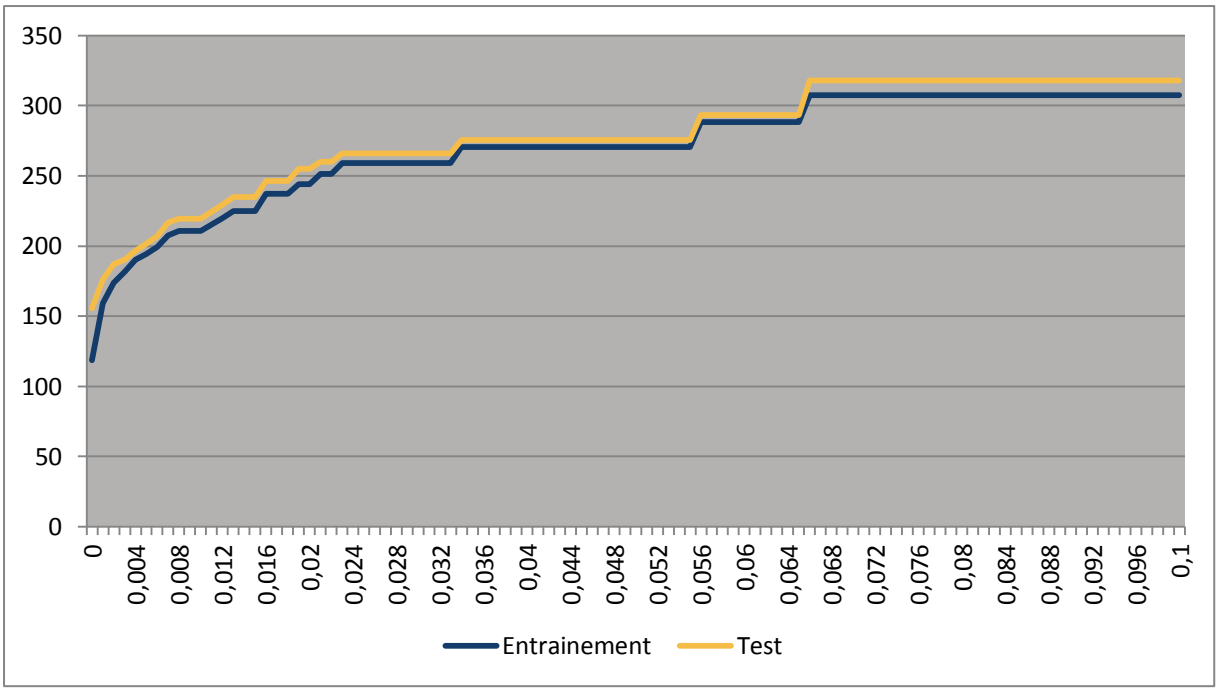



*Figure 8 : Représentation de l'arbre maximal*

L'idée est maintenant de réduire la taille de cet arbre en incorporant un poids pour chacun des nœuds créés. La valeur optimale pour le paramètre de complexité correspond à celle qui minimise le critère d'erreur conservé, ici la racine carrée de l'erreur quadratique moyenne. Pour obtenir ce minimum, on trace l'évolution de l'erreur en fonction du paramètre de complexité. Pour chacun des modèles, on étudie également l'erreur de prédiction sur la base de test pour voir s'il y a un surapprentissage ou non. Le paramètre de complexité optimal est celui qui minimise l'erreur de prédiction sur la base de test.




*Graphique 7 : Evolution de l'erreur du modèle en fonction de la valeur du paramètre de complexité*

En abscisse se trouve la valeur du paramètre de complexité qui est inversement proportionnel à la taille de l'arbre, en ordonnée se trouve la valeur du RMSE pour chacune des prédictions.

On peut remarquer sur ce graphique que les erreurs sur la base de test et sur la base d'apprentissage suivent la même évolution. Ceci montre qu'il n'y a pas de surapprentissage. En effet, s'il y en avait, l'erreur de la base de test devrait augmenter à partir d'une certaine taille d'arbre. L'absence de surapprentissage pouvait être anticipée du fait de l'absence de données aléatoires et de la présence de données utilisées par un modèle tarifaire complet.

Ainsi, le choix de la valeur du paramètre de complexité ne peut pas se faire directement par une lecture du minimum de la courbe précédente. En effet, en résonnant ainsi on donnerait la valeur 0 pour le paramètre de complexité et on conserverait l'arbre maximal ce qui n'est pas dans notre intérêt car nous recherchons les variables les plus pertinentes pour l'estimation du tarif.

Nous allons donc limiter la valeur minimale du paramètre de complexité à 0,001 pour ne pas avoir un arbre contenant trop de variables explicatives.

On obtient ainsi l'arbre élagué suivant :

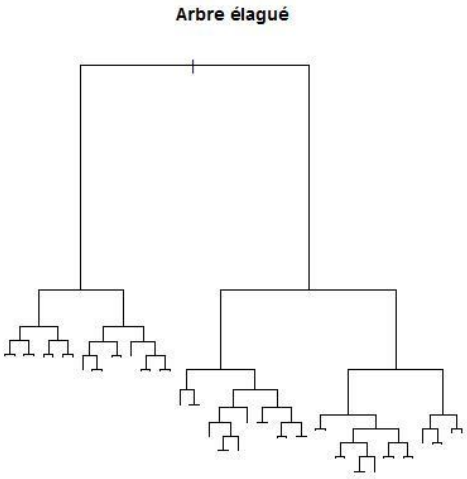



*Figure 9 : Arbre élagué avec un paramètre de complexité égal à 0,001*

L'arbre élagué est beaucoup plus lisible que l'arbre maximal. C'est celui-ci que nous conserverons pour le choix des variables principales. Pour plus de lisibilité, les différentes conditions présentes dans chaque nœud n'ont pas été affichées. L'ensemble de ces conditions ainsi que le tableau regroupant les codifications des variables se trouvent en ANNEXES 2 et 3.

En réalité, il ne s'agit que d'un arbre potentiel. En effet, si l'on relance le processus, en fonction des données mises en entrée nous obtiendrons potentiellement une autre segmentation.

Ainsi les variables conservées par l'algorithme CART sont les suivantes :

| | | |
|---|---|---|
| Honoraires - Pharmacie | 1 | Honoraires - Pharmacie Actes courants Radiologie |
| | 2 | Honoraires - Pharmacie Consultations - Visites Conventionnées Consultations de généralistes Médecin correspondant (PS) |
| | 3 | Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste (Hors PS) |
| | 4 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ADA |
| | 5 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ADC |
| | 6 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ATM |
| Hospitalisation | 7 | Hospitalisation - Chirurgie Accessoires Chambre particulière |
| | 8 | Hospitalisation - Chirurgie Accessoires Forfait journalier |
| | 9 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ADE |
| | 10 | Hospitalisation - Chirurgie Non conventionnées Frais de séjour |
| Dentaire | 11 | Dentaire Hors nomenclature Implantologie |
| | 12 | Dentaire Orthodontie Acceptée |
| | 13 | Dentaire Prothèses Acceptées Fixes métal non visibles |
| | 14 | Dentaire Prothèses Acceptées Fixes métal visibles |
| Optique | 15 | Optique Lentille ( unité) Refusée |
| | 16 | Optique Montures |
| | 17 | Optique Verres |
| Autres | 18 | Autres Cures thermales Honoraires |
| | 19 | Autres Maternité Chambre particulière |
| Démographie | 20 | % Hommes |
| | 21 | % Marié |
| | 22 | Age moyen |
| | 23 | Nombre moyen d'enfant |

*Tableau 5 : Récapitulatif des variables conservées par l'algorithme CART*

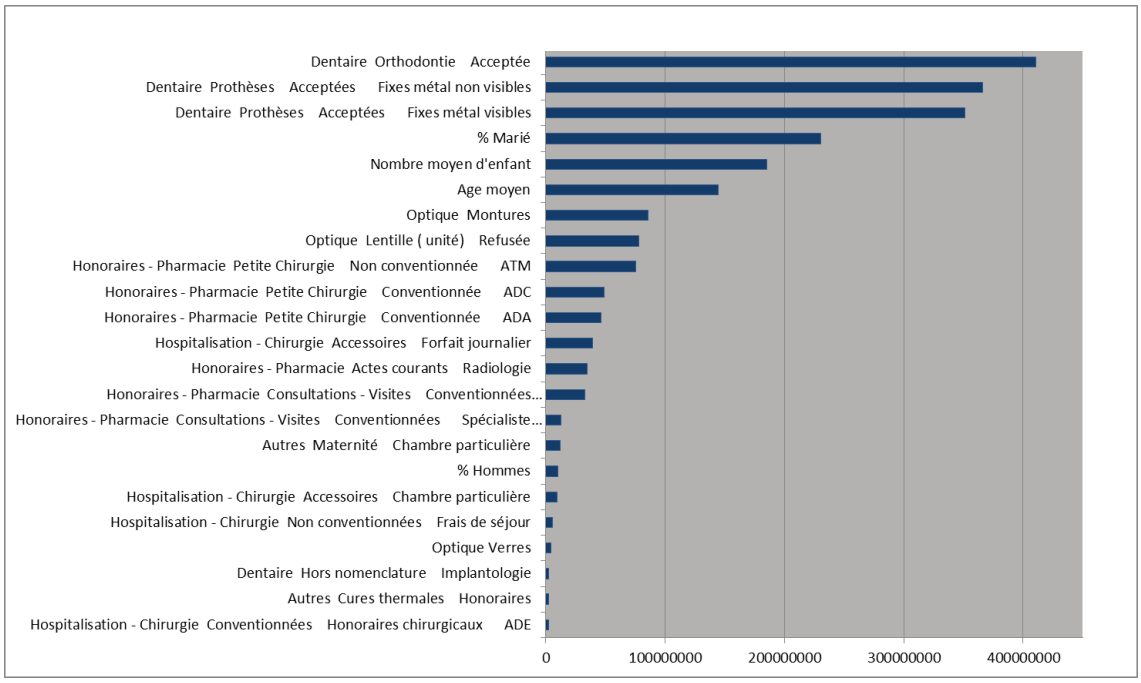
Les variables conservées concernent l'ensemble des postes couverts. Pour les informations sur la démographie de l'entreprise, on peut voir que le pourcentage de cadres ou encore l'effectif total de l'entreprise sont deux variables qui ne sont pas retenues par le modèle.

Au niveau des garanties, on observe la disparition des garanties hors nomenclature comme par exemple l'acupuncture ou le sevrage tabagique.

La garantie forfait journalier a été conservée, cela peut sembler étrange car elle ne possède qu'une seule valeur pour toutes les garanties : 18 €. De ce fait, l'utilité de cette variable est à remettre en question.

Par contre des garanties extrêmement proches apparaissent dans la segmentation comme les garanties différenciées en fonction de la lettre clé. Peu de contrats sur le marché ont une segmentation aussi fine, aussi la question de l'utilité de ce type de variables se pose. Ce sont des variables très corrélées qui apportent plus ou moins la même quantité d'information. La sélection de variables très corrélées est due à l'algorithme qui va conserver la variable diminuant au maximum l'erreur lors de la segmentation, il ne prend pas du tout en compte le nombre total de variables conservées dans les nœuds mais seulement le nombre de feuilles.

Ci-dessous, un graphique représentant l'importance de chacune des variables conservées dans la détermination des différentes classes par l'algorithme CART.




*Graphique 8 : Importance des variables dans la segmentation de l'algorithme CART*

Ce graphique donne des informations sur l'importance d'une variable sur la segmentation finale. On peut observer que les garanties d'un même domaine ont environ la même importance.

Ainsi, les variables les plus importantes pour la tarification uniforme d'un contrat complémentaire sont les garanties dentaires, puis les spécificités démographiques de l'entreprise et dans un troisième temps les garanties honoraires - pharmacie. Ainsi la répartition entre les hommes et les femmes, l'âge moyen, le pourcentage de mariés et le nombre moyen d'enfants sont importants pour la tarification, mais pas autant qu'une bonne détermination des garanties dentaires.

Ces résultats sont surprenants car on aurait pu s'attendre à ce que les caractéristiques démographiques, notamment l'âge moyen, aient un rôle prépondérant dans la détermination du tarif. Par contre, ils montrent bien l'importance de l'obtention de données démographiques correctes pour tarifer.

Pour réduire encore d'avantage le nombre de variables, nous allons faire une étude des corrélations dans un premier temps, puis nous allons confirmer l'utilité des variables choisies par l'algorithme CART par l'utilisation de méthodes GLM.

## C) ÉTUDES DES CORRÉLATIONS

L'objectif de cette partie est double. D'une part, l'étude des corrélations pourra permettre de retenir moins de variables pour la tarification santé en ne prenant pas en considération celles qui sont trop proches et d'autre part cette partie est obligatoire pour l'utilisation de modèles GLM par la suite.

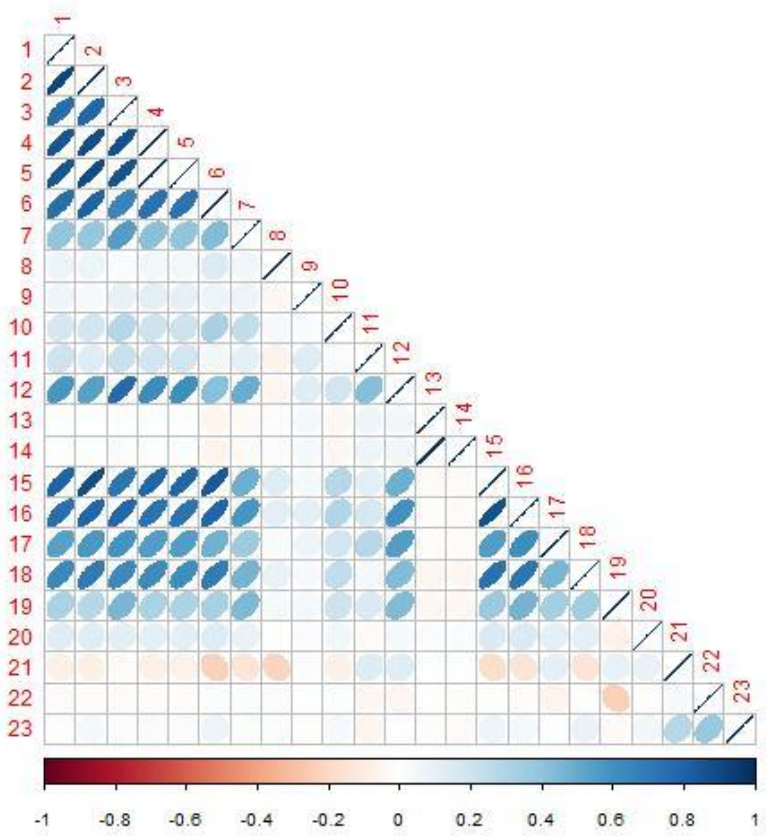



*Figure 10 : Corrélations entre les différentes variables conservées*

La numérotation des variables est celle du TABLEAU 5 indiquant les variables conservées par l'algorithme CART. Ce graphique nous permet de voir la forme de la copule entre les différentes variables du modèle. La couleur permet de repérer si la corrélation est positive ou négative.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,000 | 0,917 | 0,770 | 0,853 | 0,856 | 0,757 | 0,393 | 0,070 | 0,067 | 0,177 | 0,203 | 0,586 | 0,007 | 0,009 | 0,809 | 0,755 | 0,547 | 0,642 | 0,316 | 0,126 | -0,076 | -0,028 | 0,005 |
| 2 | 0,917 | 1,000 | 0,783 | 0,885 | 0,890 | 0,808 | 0,381 | 0,073 | 0,036 | 0,184 | 0,136 | 0,550 | 0,001 | 0,004 | 0,885 | 0,790 | 0,572 | 0,703 | 0,279 | 0,124 | -0,084 | -0,015 | 0,040 |
| 3 | 0,770 | 0,783 | 1,000 | 0,871 | 0,862 | 0,656 | 0,556 | 0,024 | 0,094 | 0,283 | 0,217 | 0,774 | 0,009 | 0,010 | 0,726 | 0,789 | 0,600 | 0,639 | 0,451 | 0,120 | -0,022 | -0,024 | 0,013 |
| 4 | 0,853 | 0,885 | 0,871 | 1,000 | 0,999 | 0,736 | 0,411 | 0,058 | 0,111 | 0,206 | 0,181 | 0,629 | 0,003 | 0,006 | 0,791 | 0,739 | 0,555 | 0,621 | 0,310 | 0,111 | -0,070 | -0,014 | 0,030 |
| 5 | 0,856 | 0,890 | 0,862 | 0,999 | 1,000 | 0,737 | 0,400 | 0,056 | 0,109 | 0,201 | 0,182 | 0,618 | 0,003 | 0,006 | 0,793 | 0,733 | 0,552 | 0,618 | 0,300 | 0,111 | -0,069 | -0,015 | 0,030 |
| 6 | 0,757 | 0,808 | 0,656 | 0,736 | 0,737 | 1,000 | 0,438 | 0,142 | 0,070 | 0,321 | 0,057 | 0,410 | -0,055 | -0,055 | 0,847 | 0,791 | 0,470 | 0,692 | 0,321 | 0,136 | -0,229 | 0,014 | 0,079 |
| 7 | 0,393 | 0,381 | 0,556 | 0,411 | 0,400 | 0,438 | 1,000 | 0,055 | 0,071 | 0,246 | 0,104 | 0,497 | -0,029 | -0,029 | 0,484 | 0,586 | 0,353 | 0,462 | 0,450 | 0,082 | -0,137 | 0,017 | 0,020 |
| 8 | 0,070 | 0,073 | 0,024 | 0,058 | 0,056 | 0,142 | 0,055 | 1,000 | -0,037 | 0,033 | -0,066 | -0,044 | 0,017 | 0,017 | 0,131 | 0,120 | 0,025 | 0,085 | 0,038 | -0,006 | -0,210 | 0,016 | 0,019 |
| 9 | 0,067 | 0,036 | 0,094 | 0,111 | 0,109 | 0,070 | 0,071 | -0,037 | 1,000 | 0,029 | 0,132 | 0,129 | 0,047 | 0,046 | 0,049 | 0,117 | 0,069 | 0,032 | 0,047 | 0,029 | 0,014 | 0,002 | -0,006 |
| 10 | 0,177 | 0,184 | 0,283 | 0,206 | 0,201 | 0,321 | 0,246 | 0,033 | 0,029 | 1,000 | 0,023 | 0,185 | -0,031 | -0,033 | 0,275 | 0,299 | 0,188 | 0,242 | 0,207 | 0,060 | -0,079 | 0,001 | 0,036 |
| 11 | 0,203 | 0,136 | 0,217 | 0,181 | 0,182 | 0,057 | 0,104 | -0,066 | 0,132 | 0,023 | 1,000 | 0,421 | 0,061 | 0,065 | 0,118 | 0,162 | 0,275 | 0,046 | 0,155 | -0,028 | 0,139 | -0,045 | -0,055 |
| 12 | 0,586 | 0,550 | 0,774 | 0,629 | 0,618 | 0,410 | 0,497 | -0,044 | 0,129 | 0,185 | 0,421 | 1,000 | 0,058 | 0,058 | 0,482 | 0,608 | 0,558 | 0,435 | 0,437 | 0,036 | 0,122 | -0,049 | 0,004 |
| 13 | 0,007 | 0,001 | 0,009 | 0,003 | 0,003 | -0,055 | -0,029 | 0,017 | 0,047 | -0,031 | 0,061 | 0,058 | 1,000 | 0,997 | -0,028 | -0,019 | -0,013 | -0,041 | -0,040 | 0,022 | 0,008 | -0,009 | -0,018 |
| 14 | 0,009 | 0,004 | 0,010 | 0,006 | 0,006 | -0,055 | -0,029 | 0,017 | 0,046 | -0,033 | 0,065 | 0,058 | 0,997 | 1,000 | -0,026 | -0,021 | -0,013 | -0,041 | -0,040 | 0,021 | 0,009 | -0,012 | -0,021 |
| 15 | 0,809 | 0,885 | 0,726 | 0,791 | 0,793 | 0,847 | 0,484 | 0,131 | 0,049 | 0,275 | 0,118 | 0,482 | -0,028 | -0,026 | 1,000 | 0,876 | 0,556 | 0,743 | 0,368 | 0,151 | -0,173 | -0,001 | 0,067 |
| 16 | 0,755 | 0,790 | 0,789 | 0,739 | 0,733 | 0,791 | 0,586 | 0,120 | 0,117 | 0,299 | 0,162 | 0,608 | -0,019 | -0,021 | 0,876 | 1,000 | 0,615 | 0,722 | 0,464 | 0,144 | -0,133 | -0,013 | 0,047 |
| 17 | 0,547 | 0,572 | 0,600 | 0,555 | 0,552 | 0,470 | 0,353 | 0,025 | 0,069 | 0,188 | 0,275 | 0,558 | -0,013 | -0,013 | 0,556 | 0,615 | 1,000 | 0,459 | 0,334 | 0,115 | 0,104 | -0,053 | 0,003 |
| 18 | 0,642 | 0,703 | 0,639 | 0,621 | 0,618 | 0,692 | 0,462 | 0,085 | 0,032 | 0,242 | 0,046 | 0,435 | -0,041 | -0,041 | 0,743 | 0,722 | 0,459 | 1,000 | 0,342 | 0,109 | -0,138 | -0,004 | 0,088 |
| 19 | 0,316 | 0,279 | 0,451 | 0,310 | 0,300 | 0,321 | 0,450 | 0,038 | 0,047 | 0,207 | 0,155 | 0,437 | -0,040 | -0,040 | 0,368 | 0,464 | 0,334 | 0,342 | 1,000 | -0,068 | 0,094 | -0,237 | -0,021 |
| 20 | 0,126 | 0,124 | 0,120 | 0,111 | 0,111 | 0,136 | 0,082 | -0,006 | 0,029 | 0,060 | -0,028 | 0,036 | 0,022 | 0,021 | 0,151 | 0,144 | 0,115 | 0,109 | -0,068 | 1,000 | 0,082 | -0,002 | 0,060 |
| 21 | -0,076 | -0,084 | -0,022 | -0,070 | -0,069 | -0,229 | -0,137 | -0,210 | 0,014 | -0,079 | 0,139 | 0,122 | 0,008 | 0,009 | -0,173 | -0,133 | 0,104 | -0,138 | 0,094 | 0,082 | 1,000 | 0,035 | 0,270 |
| 22 | -0,028 | -0,015 | -0,024 | -0,014 | -0,015 | 0,014 | 0,017 | 0,016 | 0,002 | 0,001 | -0,045 | -0,049 | -0,009 | -0,012 | -0,001 | -0,013 | -0,053 | -0,004 | -0,237 | -0,002 | 0,035 | 1,000 | 0,372 |
| 23 | 0,005 | 0,040 | 0,013 | 0,030 | 0,030 | 0,079 | 0,020 | 0,019 | -0,006 | 0,036 | -0,055 | 0,004 | -0,018 | -0,021 | 0,067 | 0,047 | 0,003 | 0,088 | -0,021 | 0,060 | 0,270 | 0,372 | 1,000 |

*Tableau 6 : Corrélations entre les variables conservées*

De façon générale, on peut remarquer qu'il existe une corrélation positive entre les variables dépendant du poste Honoraires & Pharmacie, celles dépendant de l'Hospitalisation, l'Optique et le Dentaire sauf les prothèses dentaires. Cela confirme l'existence de niveaux de couverture au global, si une garantie couvre bien un poste alors elle a plus de chance de bien couvrir l'ensemble des postes.

L'optique est corrélé avec les prothèses dentaires contrairement aux autres postes. Cela induit que le remboursement du dentaire et de l'optique joue un rôle particulier dans la création de produit santé.

L'étude de la copule entre les garanties verres et prothèses dentaires montre l'existence de paliers. Ces paliers sont représentés par les rectangles présents dans la copule. Ils indiquent que dans la majorité des cas si le niveau de l'optique est élevé alors le niveau du dentaire l'est aussi.

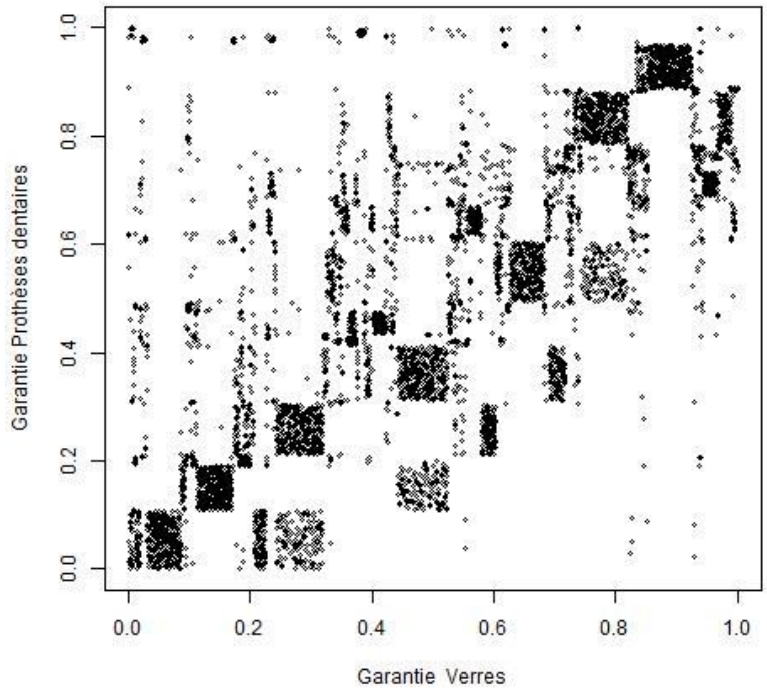



*Figure 11 : Copule des garanties prothèses dentaires remboursées par la SS et la garantie verres de lunette*

En ce qui concerne la démographie, on remarque une corrélation négative entre l'âge moyen de l'entreprise et les garanties honoraires & pharmacie et hospitalisation. Par contre, il y a une corrélation positive avec les garanties sur les verres de lunettes, l'implantologie et l'orthodontie.

Finalement, si l'on prend comme seuil d'exclusion une corrélation à 0,9, on ne conserve plus les variables suivantes dans le modèle :

- Honoraires – Pharmacie Consultations – Visites Conventionnées Consultations de généralistes Médecin correspondant (PS)
- Honoraires – Pharmacie Petite Chirurgie Conventionnée ADC
- Dentaire Prothèses Acceptées Fixes métal visibles

Nous allons maintenant tester les 20 variables restantes à l'aide d'un modèle GLM pour confirmer ou non l'utilité de celles conservées dans le modèle.

# D) CONFIRMATION DES RÉSULTATS PAR L'UTILISATION DE MÉTHODES GLM

Le but de cette partie est d'utiliser les modèles GLM pour confirmer ou infirmer la conservation de variables dans la suite de la modélisation.

Le modèle GLM permet d'analyser les différents liens existants entre une variable aléatoire $Y$ et une liste de $p$ variables explicatives $X_1, \dots, X_p$. Ici la variable à expliquer sera le tarif obtenu à l'aide de la méthode fréquence * coût et les différentes variables explicatives seront celles conservées par l'algorithme CART en enlevant les variables les plus corrélées entre elles.

Le but de cette partie n'est pas d'obtenir un modèle mais de vérifier la pertinence du choix des variables. C'est pour cela que nous n'étudierons pas en détail les différents principes d'estimations des paramètres ou de choix de modèle. Le lecteur intéressé pourra se référer à DENUIT & CHARPENTIER [2005].

## 1) THÉORIE

### 1.1) MODÈLES LINÉAIRES

Le but de ce modèle est d'expliquer un ensemble de $n$ variables $Y_1, Y_2, \dots, Y_n$ par un ensemble de $p$ variables explicatives $X_1, \dots, X_p$.

Si l'on note :

- $Y = (Y_1, \dots, Y_n)^t$ l'ensemble des variables à expliquer
- $X = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_p \end{pmatrix}$ le vecteur des variables explicatives
- $\beta = (\beta_1, \dots, \beta_p)^t$ le vecteur représentant les paramètres du modèle
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ le vecteur des erreurs du modèle

Le modèle linéaire peut s'écrire sous la forme suivante :

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

Avec $Y \sim N(\mu, \sigma^2)$ et $\mathbb{E}[\epsilon] = 0$

Ce modèle est très facile à mettre en œuvre mais il repose sur des hypothèses très fortes, notamment sur la normalité des variables à expliquer. Cette hypothèse n'est pas plausible dans notre cas, en effet, les tarifs ne suivent pas une loi normale car cela impliquerait une probabilité non nulle d'avoir des tarifs négatifs.

Pour ces raisons, nous allons nous intéresser au modèle linéaire généralisé.

## 1.2) MODÈLE LINÉAIRE GÉNÉRALISÉ

Le modèle linéaire généralisé permet de lever l'hypothèse de normalité de la variable à expliquer. La nouvelle contrainte de ce type de modélisation est que la variable à expliquer appartienne à la famille exponentielle.

Plus précisément, il faut que la densité de la variable à expliquer puisse s'écrire sous la forme :

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

Avec :

- $y \in \mathbb{N}$ ou $y \in \mathbb{R}$
- $a$ une fonction de $\mathbb{R}$ non nulle
- $b$ une fonction de $\mathbb{R}$ dérivable deux fois
- $c$ une fonction de $\mathbb{R}^2$

$\theta$ est le paramètre naturel du modèle et $\phi$ le paramètre de forme.

Les lois les plus connues appartenant à cette famille de fonction sont les lois Normales, Poisson, Gamma et Exponentielle.

Si l'on considère les variables $Y_1, \dots, Y_n$ comme étant indépendantes et identiquement distribuées, alors leur densité de probabilité jointe s'écrit sous la forme :

$$f(y|\theta,\phi) = \prod_{i=1}^{n} f(y_i|\theta_i,\phi) = exp\left\{\frac{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} b(\theta_i)}{\frac{\phi}{\omega_i}} + \sum_{i=1}^{n} c(y_i,\phi)\right\}$$

$\omega_i$ correspond à l'effectif ou au poids des données.

En notant $\mu$ l'espérance de $Y$, le modèle linéaire généralisé s'écrit :

$$\mathrm{g[E(Y)]} = \mathrm{g(\mu)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \xi$$

$$\mathrm{E(Y)} = \mu = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \xi)$$

$$\mathrm{E(Y_i)_{i=1,\dots,n}} = \mu_i = g^{-1}\left(\sum_{j=0}^{p} \beta_j x_{ij} + \xi_i\right)$$

La fonction lien $g$ représente le lien entre les variables explicatives et l'espérance mathématique de Y, cette fonction doit être bijective et dérivable.

$\xi$ représente la variable offset du modèle. Ce type de variable permet de calibrer et de tarer les modèles GLM. Un exemple simple est l'utilisation de la période d'exposition dans un processus de mise en place de norme tarifaire par le biais de GLM.

### 1.2.1) ESTIMATION DES PARAMÈTRES DU MODÈLE

Pour estimer les paramètres d'un modèle GLM, il faut utiliser la méthode dite du maximum de vraisemblance.

La vraisemblance $L$ du modèle est :

$$L = \prod_{i=1}^{n} f(y_i|\theta_i, \phi)$$

Soit en passant au log :

$$\ln(L) = \sum_{i=1}^{n} \ln(f(y_i|\theta_i, \phi)$$

$$\ln(L) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + \sum_{i=1}^{n} c(y_i, \phi)$$

Le maximum de vraisemblance s'obtient en annulant la dérivée de la log-vraisemblance par rapport à l'ensemble des paramètres $\beta_j$.

C'est-à-dire :

$$\frac{\partial \ln(L)}{\partial \beta_j} = 0 \,, \forall\, j = 0 \dots p$$

On obtient ainsi :

$$\frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \beta_j} = \frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \beta_j}$$

$$\frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\frac{\phi}{\omega_i}} = \frac{y_i - \mu_i}{\frac{\phi}{\omega_i}}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_j} \times \frac{\partial \eta_j}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_j} X_{ij}$$

Avec $g(\mu_i) = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij} = \eta_i$

$$\frac{\partial \ln(f(y_i|\theta_i, \phi)}{\partial \beta_j} = \frac{(y_i - \mu_i) x_{ij} \times \frac{\partial \mu_i}{\partial \eta_j}}{\frac{\phi}{\omega_i} b''(\theta_i)}$$

Finalement on obtient les équations de vraisemblance suivantes :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij} \times \frac{1}{g'(\mu_i)}}{Var(Y_i)} = 0\,, \quad \forall\, j = 0 \dots p$$

C'est en résolvant ces équations par le biais de méthodes itératives comme l'algorithme de Newton Raphson ou Fischer Scoring que l'on peut obtenir une évaluation des $\beta_i$.

### 1.2.2) MODÈLE LOG GAMMA

Dans la suite, nous nous intéresserons plus particulièrement au modèle log gamma.

La fonction de densité d'une loi gamma s'écrit sous la forme :

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha - 1} \exp(-\beta y)\,, y \geq 0$$

En réécrivant la densité de la loi gamma, on retrouve la forme de la famille exponentielle en prenant comme paramètres :

$$b(\theta) = -\ln(-\theta)\ ;\ \phi = \frac{1}{\beta}\ ;\ \theta = -\frac{1}{\mu}\ ; c(y; \phi) = \beta \times \ln(\beta y) - \ln(y) - \ln(\Gamma(\beta))$$

Nous utiliserons par la suite la fonction de lien log car elle permet d'obtenir une forme multiplicative qui facilite l'analyse des résultats.

Ce modèle permet entre autre de modéliser les coûts des sinistres en assurance non vie.

### 1.2.3) TEST DE TYPE III

Le test de type III permet de sélectionner les variables les plus pertinentes pour expliquer une autre variable. Il est basé sur un test de nullité de $q$ coefficients libres, $q$ représentant le nombre de modalités d'une variable. Une de ces spécificités est qu'il permet d'effectuer des tests sur des

modèles emboités. De ce fait, pour chacune des variables conservées dans un modèle, le test va comparer la qualité du modèle actuel avec celle du sous modèle excluant une variable explicative.

Les hypothèses dans ce modèle s'écrivent sous la forme :

$$H_0 : \beta_{j1} = \beta_{j2} = \cdots = \beta_{jq} = 0$$

$$H_1 : \exists k \in \{1, \dots, q\} \; tel \; que \; \beta_{jk} \neq 0$$

Pour savoir quelle hypothèse est vérifiée, on peut utiliser le test de Wald ou le test de LRT (*Likelihood Ratio Test*).

Le test de Wald permet de tester si l'écart entre l'estimateur du modèle sans contrainte $\hat{\beta}$ et l'estimateur sous $H_0$ noté $\beta_{H0}$ est nul ou non.

Il est caractérisé par l'équation suivante :

$$S = \left(Q\hat{\beta}\right)' \left[Q\hat{I}_n^{-1}(\hat{\beta})Q'\right]^{-1} Q\hat{\beta}$$

Où $Q$ est une matrice de taille $q \times (p+1)$ telle que $Q\beta = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jq})'$ et $\hat{I}_n(\hat{\beta})$ est la matrice d'information de Fisher exposant $n$.

Le test LRT va lui tester si l'écart de vraisemblance entre les points $\hat{\beta}$ et $\beta_{H0}$ est nul ou non.

Sous $H_0$ les deux tests suivent asymptotiquement une loi du $\chi^2$ à $q$ degrés de liberté. Ainsi en fonction de la p-valeur associée à ce test, on pourra conserver ou non la variable dans le modèle.

### 1.2.4) DÉTERMINATION DE LA QUALITÉ DU MODÈLE

Plusieurs indicateurs existent pour déterminer la qualité d'un modèle. Les plus connus sont la déviance, l'AIC, l'AICC et le BIC.

#### 1.2.4.1) DÉVIANCE

La déviance utilise une comparaison entre le modèle saturé et le modèle actuel pour pouvoir déterminer sa qualité. Par modèle saturé, on entend le modèle qui possède autant de paramètres que d'observations. Ce modèle décrit donc parfaitement les données.

Si l'on considère la notion de vraisemblance, cela implique que la vraisemblance du modèle étudié notée $L$ soit très proche de la vraisemblance du modèle saturé notée $L^{sat}$.

La statistique de la Déviance se définit de la façon suivante :

$$D = 2\ln\left(\frac{L^{sat}}{L}\right)$$

### 1.2.4.2) LE CRITÈRE D'INFORMATION D'AKAIKE : AIC

$$AIC = -2FLL + 2p$$

$p$ représente le nombre de paramètres estimés dans le modèle et $FLL$ la log vraisemblance calculée sur le nombre de paramètres à estimer.

### 1.2.4.3) LE CRITÈRE D'INFORMATION D'AKAIKE CORRIGÉ : AICC

Ce critère prend également en compte le nombre total d'observations noté $n$.

$$AICC = -2FLL + 2p\frac{n}{n-p-1}$$

### 1.2.4.4) LE CRITÈRE D'INFORMATION DE BAYES : BIC

$$BIC = -2FLL + pln(n)$$

Le meilleur des modèles sera celui qui minimisera la valeur de ces critères.

### 1.2.4.5) RÉSIDUS DE DÉVIANCE

De nombreux types de résidus existent pour l'analyse des GLM, nous nous focaliserons sur les résidus de déviance. En effet, ils permettent de corriger l'asymétrie des résidus de Pearson et approchent mieux la loi normale.

Pour chaque observation $i$, les résidus de déviance sont calculés de la façon suivante :

$$r_{D_i} = (y_i - \hat{y}_i) * \sqrt{d_i}$$

$d_i$ est l'apport de l'observation $i$ à la déviance du modèle, c'est-à-dire

$$D = \sum_i d_i$$

$\hat{y}_i$ est l'estimation de l'observation $y$ par le modèle, plus précisément :

$$\hat{y}_i = g^{-1}\left(\sum_{j=0}^{p} \hat{\beta}_j x_{ij} + \xi_i\right)$$

Les résidus de déviance normalisés s'écrivent sous la forme

$$r_{D_i}^{*} = \frac{r_{D_i}}{\sqrt{\phi(1-h_i)}}$$

$h_i$ correspond à l'effet levier et est le i$^{ème}$ élément diagonale de la matrice :

$$\left(\left(\frac{\phi V(\mu)}{\omega}\right)^{-1}[g'(\mu)]^{-2}\right)^{\frac{1}{2}} X\left(X^T\left(\left(\frac{\phi V(\mu)}{\omega}\right)^{-1}[g'(\mu)]^{-2}\right)X\right)^{-1} X^T\left(\left(\frac{\phi V(\mu)}{\omega}\right)^{-1}[g'(\mu)]^{-2}\right)^{\frac{1}{2}}$$

Le modèle convient si la représentation graphique des résidus est celle d'une loi normale.

## 2) APPLICATIONS AUX DONNÉES

Nous allons maintenant appliquer la théorie des GLM et plus précisément une procédure *stepwise* sur des tests de type III pour obtenir les variables les plus pertinentes dans l'explication du tarif uniforme.

Actuellement, il nous reste 20 variables explicatives sur les 149 initiales :

- La garantie radiologie
- La garantie spécialiste conventionné hors parcours de soins
- La garantie petite chirurgie conventionnée ADA
- La garantie petite chirurgie non conventionnée ATM
- La garantie chambre particulière en hospitalisation
- La garantie forfait journalier
- La garantie honoraires chirurgicaux conventionnés ADE
- La garantie frais de séjour non conventionné
- La garantie implantologie
- La garantie orthodontie acceptée
- La garantie prothèses dentaires acceptées en métal non visible
- La garantie lentille refusée
- La garantie monture
- La garantie verres
- La garantie honoraires de cures thermales
- La garantie chambre particulière en cas de maternité
- Le pourcentage d'hommes
- Le pourcentage de mariés
- L'âge moyen

- Le nombre moyen d'enfants

Pour savoir si elles sont toutes nécessaires à la tarification, nous allons utiliser un modèle log gamma à l'aide du logiciel ADDACTIS Pricing®.

Les variables en entrée pour expliquer le tarif sont continues. Pour pouvoir les utiliser, nous allons diviser leur intervalle de réalisation en cinq segments d'exposition équivalente. Cela a pour but de ne pas avoir une certaine classe d'une variable avec trop peu de réalisations pour pouvoir faire converger le modèle.

Nous allons également utiliser l'option de recherche automatique de modèle appelée *stepwise*. Cette procédure va chercher à ajouter au modèle initial le facteur disponible le plus significatif. Par plus significatif, on entend la variable qui permet d'obtenir une p-valeur au test de type III inférieure à un critère d'inclusion. Le critère d'inclusion choisi est de 5%.

Une fois cette nouvelle variable significative ajoutée, le même test est effectué sur l'ensemble des variables restantes pour savoir si une autre variable ajoute de l'information au modèle.

Quand il n'y a plus de variable à inclure, on va chercher à exclure les variables actuellement présentes dans le modèle et qui sont le moins significatives en réalisant un test de type III avec un seuil d'exclusion qui est ici aussi calibré à 5%.

Une fois qu'il n'y a plus de variables à exclure, le processus s'arrête.

Lorsque l'on utilise un modèle log gamma sur les données on obtient les résultats suivants :

| Criteria | Degrees of freedom | Chi-2 | Pr > Chi-2 |
|---|---|---|---|
| Optique Lentille ( unité) Refusée | 4 | 69,98069257 | 2,29E-14 |
| Dentaire Orthodontie Acceptée | 4 | 82,46042478 | 5,24E-17 |
| Hospitalisation - Chirurgie Accessoires Forfait journalier | 0 | 0 | 0 |
| Honoraires - Pharmacie Petite Chirurgie Conventionnée ADA | 4 | 93,29567214 | 2,63E-19 |
| Optique Verres | 4 | 90,34651224 | 1,11E-18 |
| Dentaire Hors nomenclature Implantologie | 4 | 130,2848661 | 3,38E-27 |
| Dentaire Prothèses Acceptées Fixes métal non visibles | 4 | 326,398229 | 2,18E-69 |
| Nombre moyen d'enfant | 4 | 371,6904209 | 3,63E-79 |
| % Hommes | 4 | 174,1958676 | 1,31E-36 |
| Age moyen | 4 | 2192,028553 | 0 |
| % Marié | 4 | 1943,972115 | 0 |
| Hospitalisation - Chirurgie Non conventionnées Frais de séjour | 4 | 121,2898496 | 2,83E-25 |
| Autres Maternité Chambre particulière | 4 | 52,79240568 | 9,42E-11 |
| Honoraires - Pharmacie Actes courants Radiologie | 4 | 36,09252385 | 2,77E-07 |
| Autres Cures thermales Honoraires | 4 | 32,4614953 | 1,54E-06 |
| Hospitalisation - Chirurgie Accessoires Chambre particulière | 4 | 32,44320121 | 1,55E-06 |
| Honoraires - Pharmacie Petite Chirurgie Non conventionnée ATM | 4 | 29,48861683 | 6,22E-06 |
| Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste (Hors PS) | 4 | 14,31817426 | 0,006345846 |
| Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ADE | 4 | 9,809717712 | 0,043757891 |

*Tableau 7 : P-valeurs pour le test de type III sur les variables incluses dans le modèle*

| Criteria | Degrees of freedom | Chi-2 | Pr > Chi-2 |
|---|---|---|---|
| Optique Montures | 4 | 2,885675029 | 0,57713606 |

*Tableau 8 : P-valeurs pour le test de type III sur la variable exclue du modèle*

Le résultat de ce modèle est le suivant, toutes les variables apportent de l'information sur le tarif sauf la garantie monture. En effet, sa p-valeur à 58% ne respecte pas la contrainte d'inclusion de la procédure *stepwise* et donc n'est pas importante pour la modélisation. L'information qu'elle apporte est en fait déjà fournie par une autre variable présente dans le modèle.

L'utilité de la variable Hospitalisation – Chirurgie Conventionnées Honoraires chirurgicaux ADE peut aussi être remise en cause avec une p-valeur à 4,4%.

Le but final de cette étude étant de réduire au maximum le nombre de variables, nous allons comparer le modèle incluant cette variable et celui ne l'incluant pas.

## 2.1) CHOIX DU MODÈLE

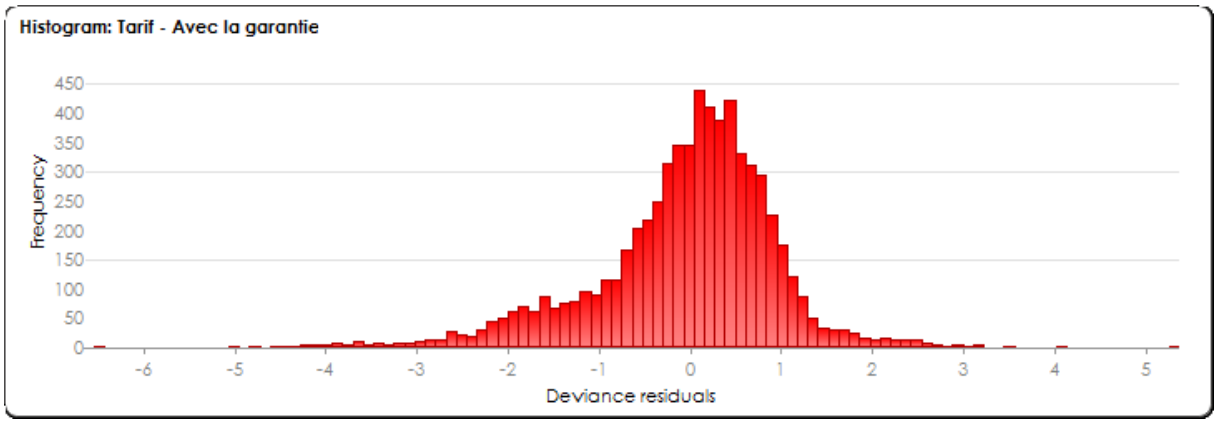
Le choix entre le modèle comportant la variable représentant la garantie honoraires chirurgicaux conventionnés et celui qui ne l'intègre pas se fera par rapport aux différents indicateurs statistiques ainsi que par rapport aux résidus des modèles.

| | Avec la variable | Sans la variable | Comparaison |
|---|---|---|---|
| **DoF** | 6446 | 6450 | 4,00E+00 |
| **Deviance** | 166,2928093 | 169,4023376 | 3,11E+00 |
| **Pr>ChiSq** | | | 2,83231E-25 |
| **AIC** | 84082,85196 | 84196,14181 | 1,13E+02 |
| **BIC** | 84584,7552 | 84670,91515 | 8,62E+01 |

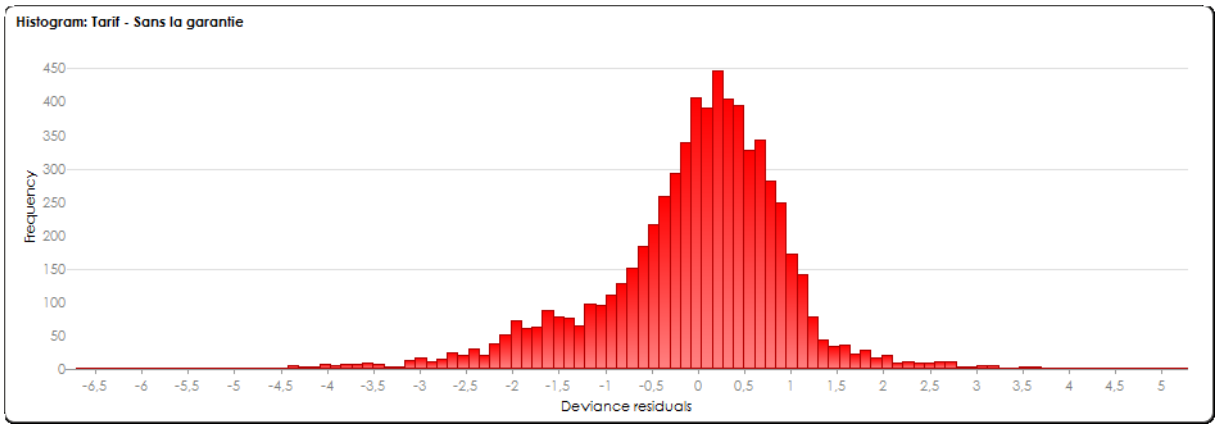
*Tableau 9 : Récapitulatif des indicateurs statistiques pour les modèles GLM*

Le modèle intégrant la garantie honoraires chirurgicaux conventionnés est meilleur que celui ne l'intégrant pas, par contre cette amélioration est très faible. En effet, on peut voir que l'AIC ne diminue que de 1% en intégrant cette variable.

Si l'on s'intéresse aux différents graphiques des résidus des modèles, on obtient :



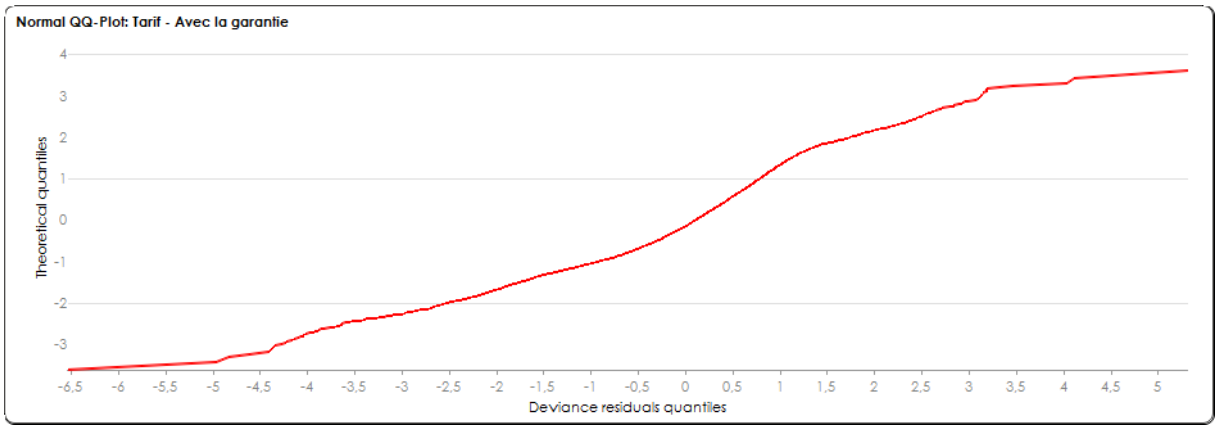

*Graphique 9 : Histogramme des résidus de déviance pour le modèle conservant la garantie honoraires chirurgicaux conventionnés*



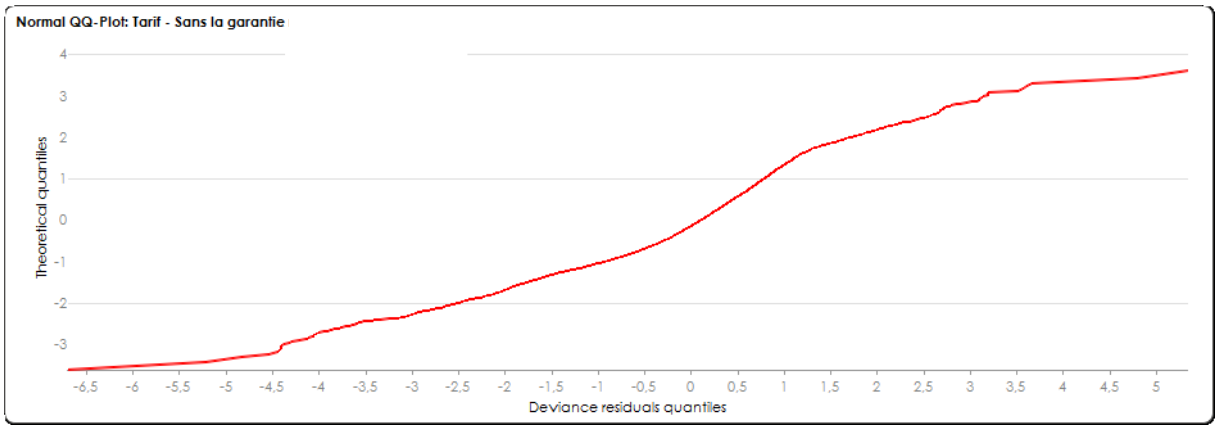

*Graphique 10 : Histogramme des résidus de déviance pour le modèle sans la garantie honoraires chirurgicaux conventionnés*

Les deux histogrammes ont l'aspect d'une loi normale et sont très peu différents, cela montre qu'ils décrivent approximativement de la même façon les données.

Voici les QQ-plots qui permettent de faciliter la lecture :



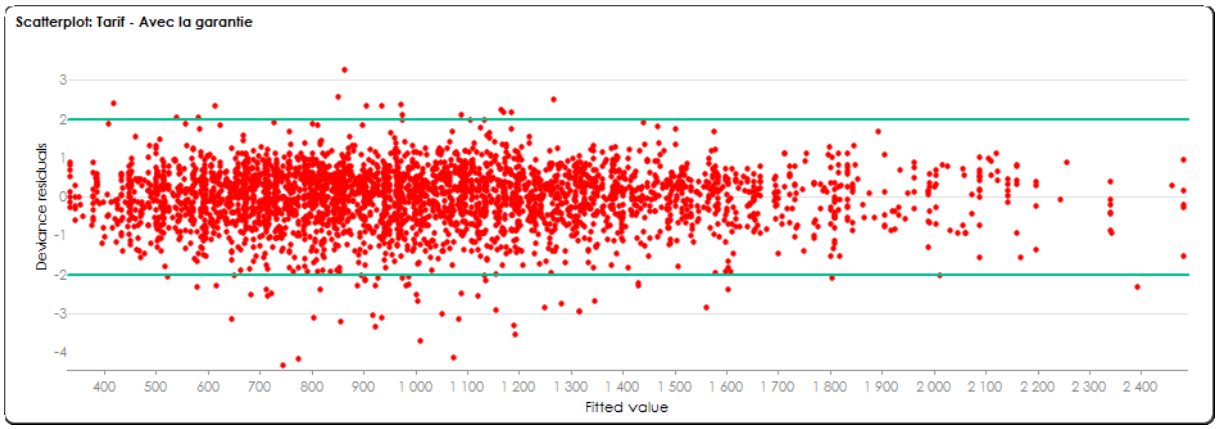

*Graphique 11 : QQ-plot des résidus de déviance pour le modèle conservant la garantie honoraires chirurgicaux conventionnés*



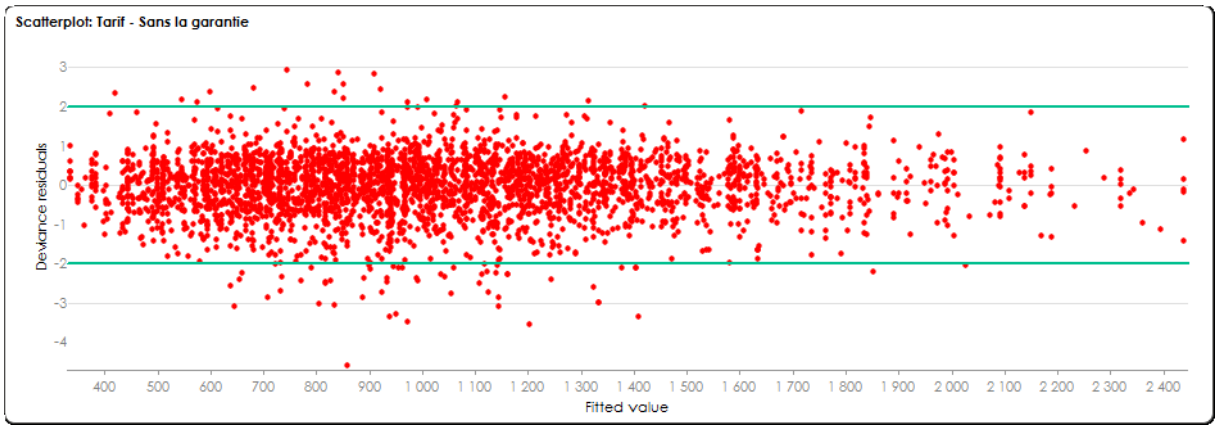

*Graphique 12 : QQ-plot des résidus de déviance pour le modèle sans la garantie honoraires chirurgicaux conventionnés*

Les courbes sont presque identiques, les modèles sont donc très proches l'un de l'autre.

Pour finir, voici le nuage de points des résidus de déviance pour les deux modèles :




*Graphique 13 : Nuage de points des résidus de déviance pour le modèle conservant la garantie honoraires chirurgicaux conventionnés*




*Graphique 14 : Nuage de points des résidus de déviance pour le modèle sans la garantie honoraires chirurgicaux conventionnés*

Les nuages de points des résidus de déviance pour un modèle GLM parfaitement calibré doivent être centrés autour de 0, tous les points doivent être compris en -2 et 2 et ils ne doivent pas montrer une forme particulière comme une courbe ou un cône.

Si une ou des courbes apparaissent, alors cela signifie que les variables explicatives conservées ne décrivent pas entièrement la variable à expliquer. Si un cône apparait cela peut signifier une erreur dans le choix de la fonction lien ou de la fonction de répartition.

Ici, on peut voir que les résidus sont bien centrés en 0, que la majorité des points se trouvent entre -2 et 2 et que les résidus ne présentent pas de forme particulière. De ce fait, on peut conclure que le modèle est bien calibré. En ce qui concerne la différence entre les deux modèles, on peut voir plus de points à l'extérieur de l'intervalle [-2 ; 2] pour le deuxième modèle n'incluant pas la variable.

Finalement, nous avons pu voir par cette étude que toutes les variables conservées par l'algorithme CART sont significatives pour l'explication du tarif uniforme sauf la garantie monture. Une ambivalence quant à l'utilité de la variable reprenant l'équivalent euro de la garantie honoraires chirurgicaux conventionnés existe. Cependant, l'analyse des modèles intégrant ou non cette variable montrent qu'ils sont extrêmement proches et que l'apport de cette garantie n'est pas très important. De ce fait, nous conserverons l'ensemble des variables gardées par l'algorithme CART à l'exception de la garantie monture et de la garantie honoraires chirurgicaux conventionnés.

Dans la suite, nous allons utiliser ces variables pour obtenir une estimation du tarif.

Le modèle GLM présenté ci-dessus peut bien sûr être amélioré par l'utilisation d'approximations polynomiales sur les variables ou encore par l'exclusion d'observations augmentant de façon trop importante la déviance globale pour obtenir de meilleurs résidus. Par contre, cela ne changera pas fondamentalement les résultats obtenus dans le choix des variables explicatives.

Finalement les variables conservées dans la suite de l'étude sont les suivantes par ordre d'importance :

- La garantie prothèses dentaires acceptées en métal non visibles
- Le pourcentage de mariés
- L'âge moyen
- La garantie petite chirurgie non conventionnée ATM
- La garantie radiologie
- Le nombre moyen d'enfants
- La garantie orthodontie acceptée
- La garantie lentille refusée
- La garantie spécialiste conventionné hors parcours de soins
- La garantie chambre particulière en hospitalisation
- La garantie verres
- La garantie frais de séjour non conventionnés
- La garantie implantologie
- Le pourcentage d'hommes
- La garantie petite chirurgie conventionnée ADA
- La garantie chambre particulière en cas de maternité
- La garantie honoraires de cures thermales
- La garantie forfait journalier

# III) APPRENTISSAGE PAR RÉSEAUX DE NEURONES ARTIFICIELS

## A) PRÉSENTATION DES RÉSEAUX DE NEURONES

### 1) LE NEURONE

La théorie des réseaux de neurones est basée sur l'étude de ces réseaux en biologie. Chaque neurone autrement appelé perceptron reçoit une ou plusieurs stimulations en entrée notées $x_1 \dots x_n$. Chacune de ces stimulations va passer par les différentes entrées du neurone qui vont pondérer ce signal avec un certain poids noté $\omega_i$. Finalement, il va fournir un signal $y$ en sortie. On peut le représenter de la façon suivante :

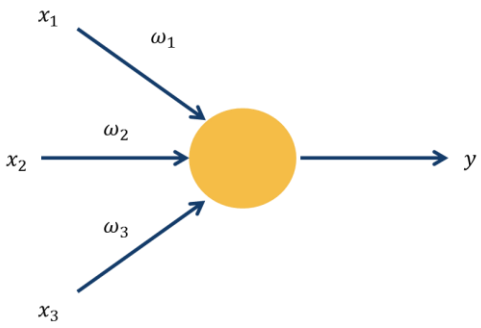



*Figure 12 : Neurone artificiel*

Mathématiquement parlant, un neurone se caractérise de la sorte :

$$y = f\left(\sum_i \omega_i x_i\right)$$

La fonction $f$ est appelée fonction d'activation. La fonction d'activation doit être de préférence croissante, bornée et suffisamment régulière. Les principales fonctions d'activation sont :

- Le pas unitaire : $f(x) = \begin{cases} 0 \; si \; x < 0 \\ 1 \; si \; x \geq 0 \end{cases}$
- La fonction sigmoïde : $f(x) = \frac{1}{1+e^{-\beta x}}$
- La fonction normale
- La fonction identité

L'interaction de plusieurs neurones représente un réseau de neurones. Le choix de la fonction d'activation est important car c'est elle qui fournit les différentes propriétés des réseaux de neurones.

## 2) RÉSEAUX DE NEURONES

Plusieurs types de réseaux de neurones existent. Le plus simple est le perceptron monocouche autrement appelé Perceptron de Rosenblatt. Il est composé d'une couche d'entrée reprenant l'ensemble des variables du modèle et d'une couche en sortie nous fournissant les estimations nécessaires. Son point fort est sa simplicité de mise en place mais son point faible est qu'il ne peut résoudre que des problèmes de type linéaire.

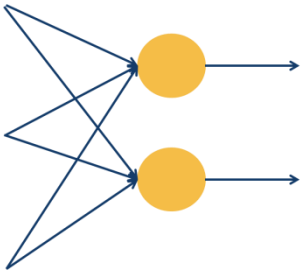


*Figure 13 : Perceptron monocouche (3 entrées, 2 sorties)*

Une amélioration de ce modèle est le perceptron multicouche. Contrairement au perceptron monocouche, il intègre une couche cachée entre la couche d'entrée et la couche de sortie. Le fait que les variables d'entrée et de sortie ne soient pas directement reliées permet de résoudre des problématiques non linéaires. Par la suite, nous nous focaliserons sur ce type de réseau.

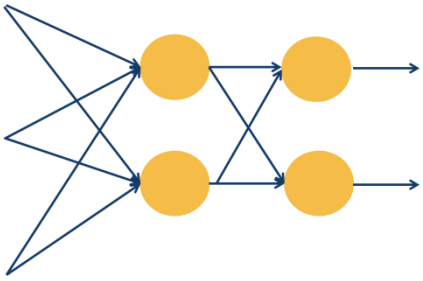


*Figure 14 : Perceptrons multicouches (3 entrées, 2 neurones cachés, 2 sorties)*

Du bruit peut être ajouté en entrée des neurones de la couche cachée ou en entrée des neurones de sortie. Ce bruit généralement gaussien permet de mieux prendre en considération le caractère aléatoire des données en entrée.

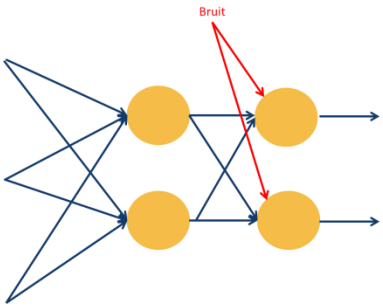



*Figure 15 : Perceptrons multicouches avec bruits (3 entrées, 2 neurones cachés, 2 sorties)*

Par la suite, on appellera synapse le lien entre deux neurones.

## 3) PROPRIÉTÉS DU PERCEPTRON MULTICOUCHE

Cybenko et Funahashi ont prouvé la propriété d'approximation universelle des perceptrons multicouches. Cette propriété peut s'écrire sous la forme :

*Toute fonction bornée suffisamment régulière peut être approchée uniformément avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire.*

Une deuxième propriété des réseaux de neurones est la propriété de parcimonie. En effet, si l'approximation par le réseau n'est pas une fonction linéaire des paramètres, alors l'estimation du réseau est plus parcimonieuse que si elle était une fonction linéaire de ces paramètres. Cela justifie entre autre l'utilisation de certaines fonctions d'activation autre que la fonction linéaire. De plus, si la fonction d'activation est la fonction sigmoïde, alors l'erreur d'approximation décroit lorsque le nombre de neurones cachés augmente. Ainsi, il est possible de trouver un nombre de neurones permettant d'avoir une erreur minimale.

Par la suite, nous conserverons un réseau de neurones de type perceptron multicouche avec une fonction d'activation sigmoïde et un neurone de sortie avec une fonction d'activation linéaire.

## 4) APPRENTISSAGE

Ce que l'on appelle apprentissage pour les réseaux de neurones est la modification de la valeur des poids $\omega_i$ en fonction des données en entrée. Il existe trois types d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage renforcé.

### 4.1) APPRENTISSAGE SUPERVISÉ

Ce type d'apprentissage propose de mettre directement en relation les variables d'entrée et la valeur souhaitée en sortie.

Les valeurs en entrée se propagent à travers le réseau en activant les différents neurones et cela fournit une valeur à la sortie. Celle-ci est comparée à la valeur attendue. Si la différence entre les deux valeurs est trop importante alors les poids du réseau sont modifiés pour que lorsqu'une donnée semblable arrive dans le réseau alors sa réponse soit plus proche de celle attendue.

### 4.2) APPRENTISSAGE NON SUPERVISÉ

L'apprentissage non supervisé ne donne au réseau de neurones que les variables en entrée. Le réseau ne peut donc pas s'adapter en fonction des valeurs attendues. Pour pouvoir fournir des résultats cohérents, il va segmenter les données en fonction de leur ressemblance ou en fonction de leur redondance pour pouvoir fournir une estimation en sortie.

### 4.3 APPRENTISSAGE RENFORCÉ

L'apprentissage renforcé est une variante des deux précédentes méthodes. Toutes les données sont insérées dans le réseau de neurones. La valeur à la sortie sera comparée avec celle attendue et le réseau ne saura que si la valeur à la sortie correspond ou non à cette valeur, il ne saura pas vers quelle valeur converger. Si la réponse est non, alors les poids sont adaptés en conséquence. L'avantage de cette méthode est qu'elle réduit la part de sur-apprentissage. En effet, en ne sachant pas vers quelle valeur se rapprocher, le réseau n'est pas trop influencé par les valeurs de la base d'apprentissage.

## 5) MINIMISATION DE L'ERREUR D'APPRENTISSAGE

### 5.1) FONCTION DE COÛT

Pour clarifier les notations, on notera $N$ le nombre d'éléments de la base d'apprentissage, $x_i$ un vecteur de taille $n$ représentant un exemple de la base d'apprentissage ($n$ correspond au nombre de variables en entrée), $\omega$ représente le vecteur des poids du réseau, $t_i$ la valeur attendue à la sortie et enfin la sortie du neurone est notée $y_i$.

Deux principales fonctions existent pour quantifier l'erreur d'un réseau de neurones. La première est l'erreur quadratique moyenne et la deuxième l'entropie croisée. L'entropie croisée étant utile pour

résoudre des problèmes de classification, nous conserverons donc l'erreur quadratique moyenne comme fonction de coût.

La fonction de coût noté $J$ correspond à :

$$J(\omega) = \frac{1}{N}\sum_{i=1}^{N}(y_i(\omega) - t_i)^2$$

## 5.2) MINIMISATION DE L'ERREUR DU MODÈLE

Le but de la minimisation de l'erreur est de déterminer les poids optimaux pour chacune des synapses du réseau.

La méthode utilisée est celle de la descente du gradient.

A l'itération $k$, on recalcule la valeur des poids par la formule suivante :

$$\omega_k = \omega_{k-1} + \alpha_{k-1} d_{k-1}$$

Avec $\alpha_k$ le pas et $d_k$ la direction souhaitée pour minimiser l'erreur.

La méthode de la descente du gradient nous permet de poser :

$$d_k = -\nabla J(\omega_k)$$

En ce qui concerne la détermination du pas, il est déterminé par rapport à la valeur du gradient de la façon suivante :

$$\alpha_k = \frac{\alpha_0}{1 + \|\nabla J(\omega_k)\|}$$

Avec $\alpha_0$ une constante qui est arbitrairement fixée à 0,01.

D'autres méthodes que la descente du gradient peuvent être utilisées comme par exemple la méthode de Newton ou la méthode de Quasi Newton. Ces deux dernières permettent d'éviter de converger vers un minimum local.

# 6) DÉGRADATION DES PONDÉRATIONS

Comme pour les arbres de régression, le risque de sur-apprentissage existe. La méthode de dégradation des pondérations permet de le réduire.

Cette méthode consiste à ajouter une pénalité en fonction de l'importance des poids entre les synapses. En effet, plus les poids sont importants, moins le réseau sera généralisable à d'autres données. Si un poids est trop important il peut même engendrer des discontinuités dans la fonction en sortie du réseau et une grande variance.

De ce fait, on rajoute à la fonction d'erreur choisie un coefficient $p$ prenant en compte l'importance de chacun des poids. Ainsi si l'on choisit l'erreur quadratique moyenne comme valeur pour comparer les modèles, ce critère devient avec la dégradation des pondérations :

$$RMSE + \frac{p}{2} \sum \omega_j^2$$

### 7) ALGORITHME DE GARSON

L'algorithme de Garson permet d'identifier les variables ayant le plus d'importance dans un réseau de neurones. L'idée principale est que l'importance relative d'une variable explicative peut être déterminée par l'identification de la valeur absolue des poids synaptiques entre deux perceptrons par lesquels transite l'information de la variable. Autrement dit, on prend en considération tous les poids de connexion d'un neurone d'entrée spécifique qui passent à travers la couche cachée pour finir sur le neurone de sortie de la variable à expliquer.

Cette méthode est répétée pour l'ensemble des variables d'entrée. Pour plus de précisions sur cet algorithme, le lecteur intéressé pourra se référer à GARSON [1991].

## B) APPLICATIONS AUX DONNÉES

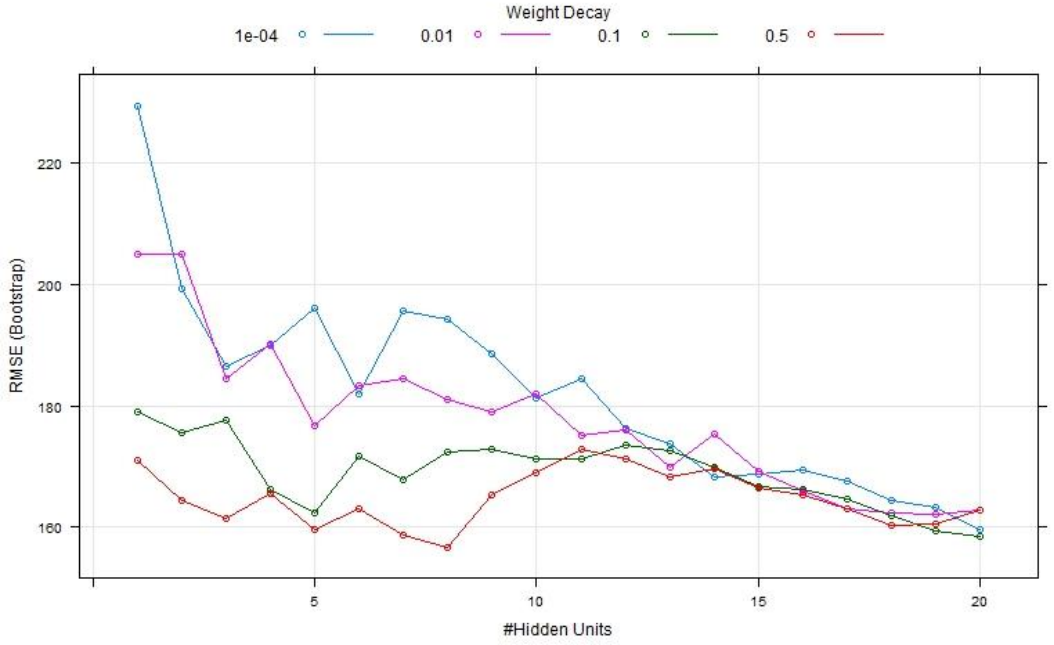
Nous allons appliquer un apprentissage supervisé à notre base de données initiale. En effet, le tarif uniforme se calculait simplement en utilisant l'ensemble des 149 variables. Maintenant que leur nombre est divisé par 7, réobtenir le tarif est plus complexe. C'est là qu'interviennent les réseaux de neurones.

Nous utiliserons ici un perceptron multicouche à une couche de neurones cachées. L'apprentissage sera réalisé sur une sous base correspondant à 70% de la base initiale et les tests sur la partie de la base restante.

Comme toutes les variables explicatives sont continues, seule une normalisation des données est à réaliser comme retraitement. Cette normalisation est nécessaire notamment lorsque les valeurs insérées en entrée du réseau sont très différentes. Par exemple, lorsque dans un premier temps on présente une garantie au ticket modérateur et dans un second temps on présente une autre garantie aux frais réels.

## 1) DÉTERMINATION DU RÉSEAU DE NEURONES OPTIMAL

La détermination d'un réseau optimal se fait par comparaison des différentes combinaisons possibles. Nous allons donc réaliser plusieurs réseaux dont le nombre de neurones cachés va de 1 à 20 et avec des coefficients de dégradations différents. Nous allons ensuite comparer la valeur de leur RMSE respectif pour choisir le réseau optimal.




*Graphique 15 : RMSE en fonction du nombre de neurones cachés et du coefficient de dégradation*

On peut remarquer que de 1 à 13 neurones cachés le coefficient de dégradation joue un rôle important dans la réduction de l'erreur. En effet, l'erreur est 13% plus élevée entre une modélisation avec un coefficient de dégradation proche de 0 et un coefficient égal à 0,5. Plus on augmente le nombre de neurones, plus la différence entre les diverses pondérations se réduit. Cela s'explique par le fait que le nombre de synapses augmente et que le poids de chaque connexion devient plus faible. D'autre part, il est recommandé de ne pas avoir trop de neurones dans la couche cachée pour éviter tout sur-apprentissage.

De ce fait, nous allons conserver comme réseau de neurones un perceptron multicouche à 8 neurones cachés avec un coefficient de dégradation des pondérations de 0,5 et une fonction d'activation sigmoïde.

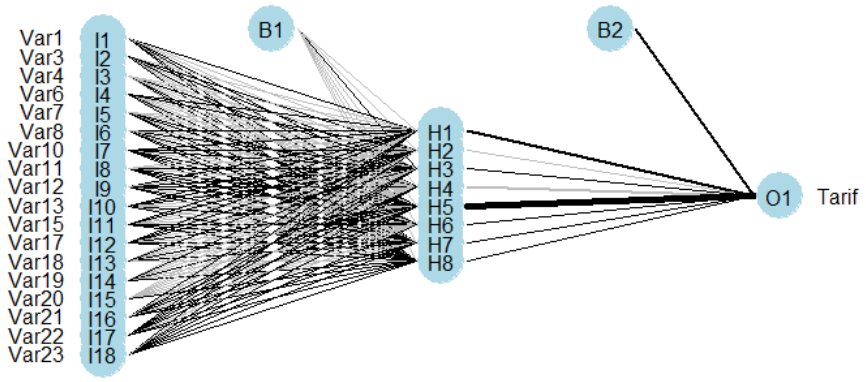
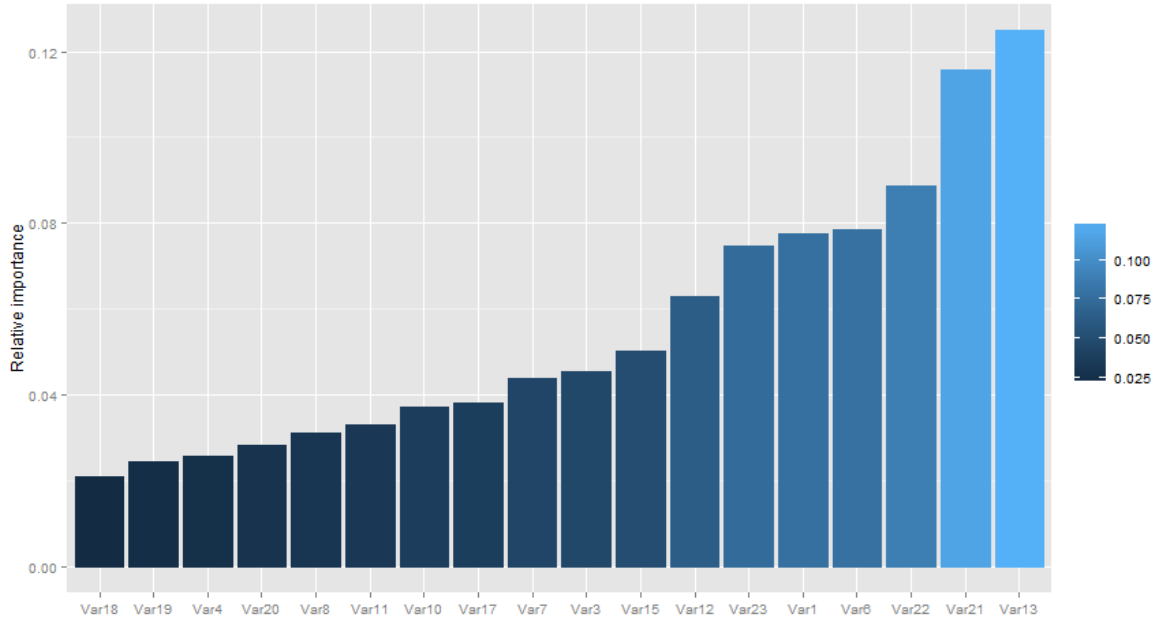



*Figure 16 : Représentation graphique du réseau de neurones obtenu*

La numérotation des variables correspond à celle du TABLEAU 5. Il s'agit ici d'une représentation d'un réseau obtenu avec une initialisation des poids aléatoire. Plus le trait est épais, plus le poids est important en valeur absolue. Si le lien entre deux neurones est noir cela signifie que le poids est positif, si ce lien est gris, cela signifie que le poids est négatif. Les éléments notés B correspondent aux bruits ajoutés au modèle.

## 2) IMPORTANCE RELATIVE DES VARIABLES

L'utilisation de l'algorithme de Garson nous permet d'obtenir les variables les plus importantes pour la prédiction.




*Graphique 16 : Importance relatives des variables du réseau de neurones par l'algorithme de Garson*

La numérotation des variables correspond à celle présente dans le TABLEAU 5.

On peut remarquer que les variables les plus importantes dans le réseau sont tout d'abord la garantie prothèse dentaire, puis le pourcentage de mariés dans l'entreprise et l'âge moyen. On retrouve ici comme pour l'algorithme CART l'importance de la garantie prothèse dentaire dans la détermination du tarif uniforme. Les caractéristiques démographiques sont également des variables importantes dans le réseau.

Parmi toutes les variables conservées, on observe que la garantie Cures thermales est celle ayant le moins d'impact avec une importance relative six fois plus faible par rapport à celle des garanties prothèses dentaires. Cette spécificité se retrouve bien actuellement dans les modèles tarifaires en fréquence * coût où l'augmentation de la garantie Cures thermales n'a pas un impact fort sur le tarif final.

## 3) DISGRÉTION SUR LES PROTHÈSES DENTAIRES EN TARIFICATION SANTÉ

Le remboursement des prothèses dentaires en assurance santé collective et individuelle peut représenter la majeure partie du prix de la garantie pour les garanties haut de gamme. Cette réalité se retrouve dans l'ordre d'importance des variables dans le réseau de neurones. Pourtant, c'est une des garanties les plus difficiles à calibrer, à comparer et même à tarifer. En effet, sous le nom global de prothèses dentaires se cache une multitude d'actes codifiés de façons différentes par la

Classification Commune des Actes Médicaux dentaires. La méconnaissance ou le non-respect des règles qui ont servi à calibrer le modèle tarifaire peut engendrer des écarts tarifaires importants entre le coût réel du risque assuré et la prime pure proposée. La spécificité de cette garantie rend très difficile la retranscription des garanties dentaires à l'identique d'un assureur à l'autre, chacun ayant ses spécificités en fonction de la calibration de son modèle tarifaire. Ainsi, il n'est pas rare d'observer des différences entre les remboursements des bridges, des inters de bridges, des couronnes sur implant ou encore sur les réparations de prothèses alors que l'intitulé global de la garantie est le même. Ce poste pouvant représenter une part importante dans le prix d'une complémentaire santé, une erreur dans son analyse pour la reprise des garanties d'un concurrent lors d'une prospection commerciale entrainera soit une sous tarification et donc un risque de perte future importante si l'affaire est réalisée soit l'échec des négociations à cause d'un prix trop élevé.

## 4) PRÉDICTION DES TARIFS

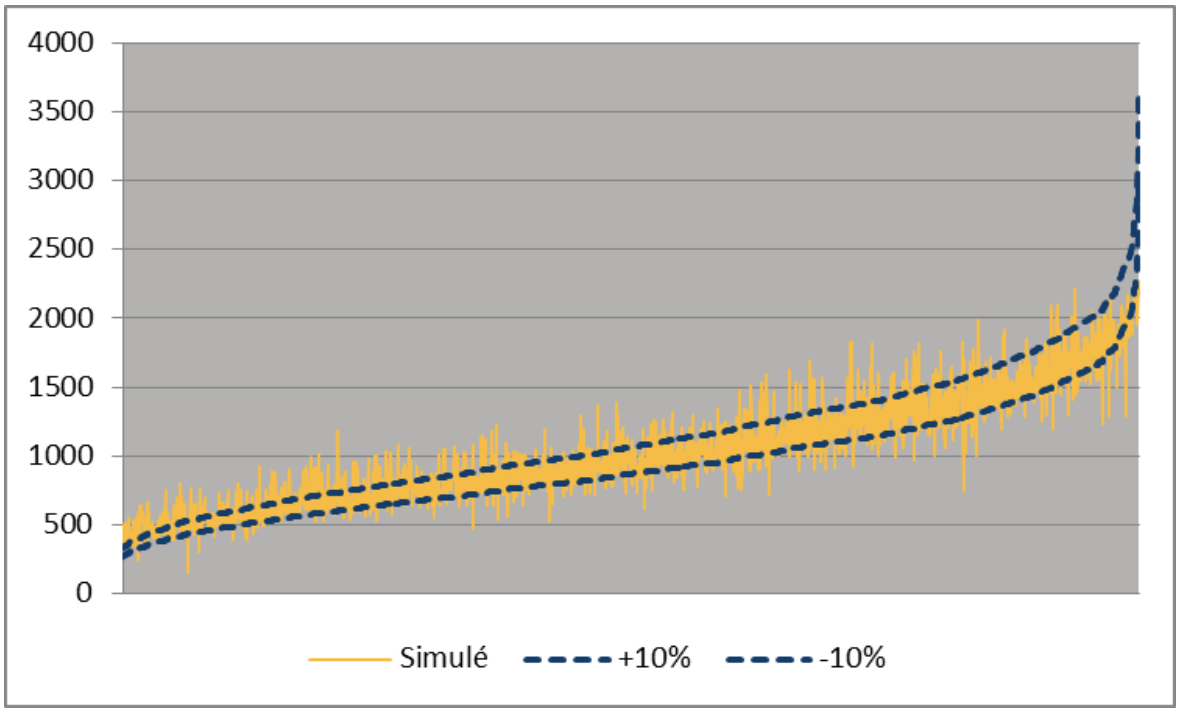
L'utilisation du réseau de neurones sur les données de la base de test nous permet de connaître des indicateurs statistiques sur la qualité du modèle.

L'utilisation d'un réseau de neurones à 8 neurones cachés permet d'avoir les résultats suivants :

$$RMSE = 148{,}32 \;\; VAR = 1{,}23\%$$

Le RMSE permet de connaître la qualité du modèle, en revanche il est difficile d'analyser sa valeur.

Pour résoudre cela, un graphique représentant un intervalle de plus ou moins 10% par rapport aux tarifs réels et les tarifs à la sortie du réseau a été réalisé.



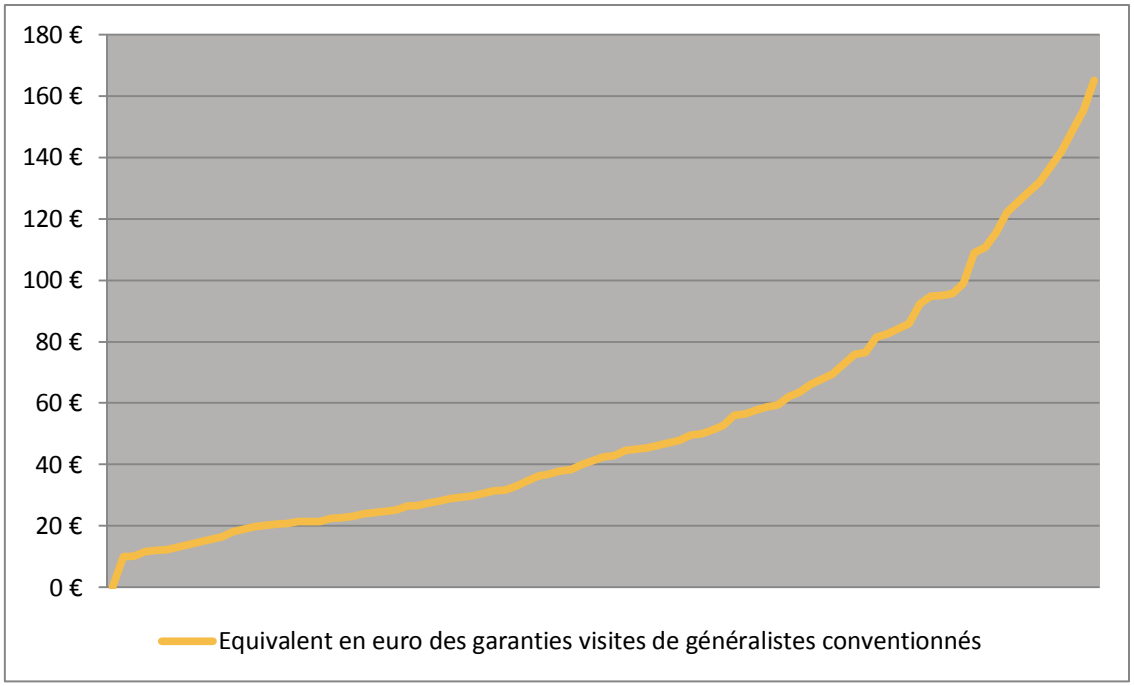

*Graphique 17 : Estimation des tarifs en sortie du réseau de neurones*

En jaune se trouve l'estimation du tarif en sortie du réseau de neurones, les pointillés représentent l'intervalle entre +10% et -10% par rapport au tarif réel.

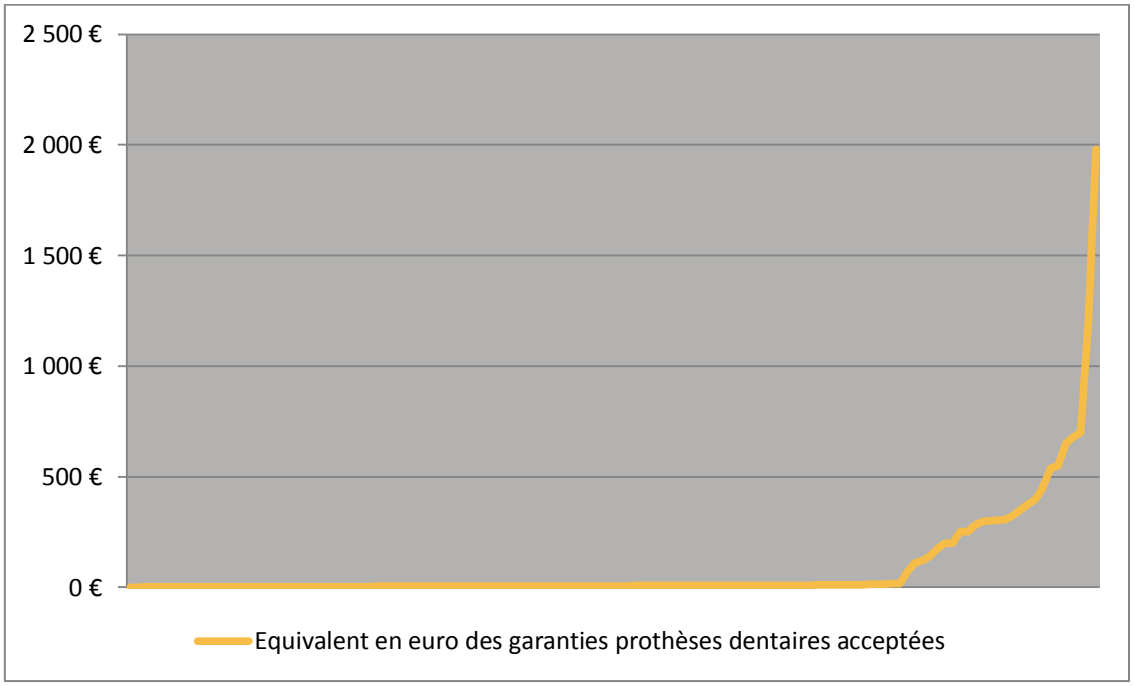
Dans un premier temps, on peut remarquer que les résultats du réseau suivent bien la même tendance que les prix réels. Il y a cependant un endroit où les tarifs sont mal estimés, il s'agit des tarifs très élevés représentant des garanties « très haut de gamme ».

Une explication à ce phénomène peut se trouver dans le fait que l'on a réduit de plus de trois quarts le nombre de variables dans la modélisation. En effet, en réduisant autant le nombre, deux études peuvent apparaître comme étant identiques en entrée du réseau et donc ressortir avec le même tarif alors que certaines vont avoir des garanties supplémentaires non renseignées dans le réseau et être en réalité plus chères. Une autre explication peut se trouver dans le fait que les tarifs extrêmes ne sont pas beaucoup représentés dans la base de données d'entrainement et de ce fait le réseau n'a pas pu apprendre correctement les caractéristiques conduisant à des tarifs élevés. Pour finir, les garanties « haut de gamme » sont celles qui présentent des garanties aux frais réels, or une hypothèse quant à la valeur de l'équivalent euro des frais réels a été réalisée. Cette hypothèse a été nécessaire mais elle peut expliquer le manque de qualité des estimations.

Les graphiques suivants permettent de mieux comprendre la problématique.




*Graphique 18 : Evolution de l'équivalent euro des garanties visites de généralistes conventionnés*



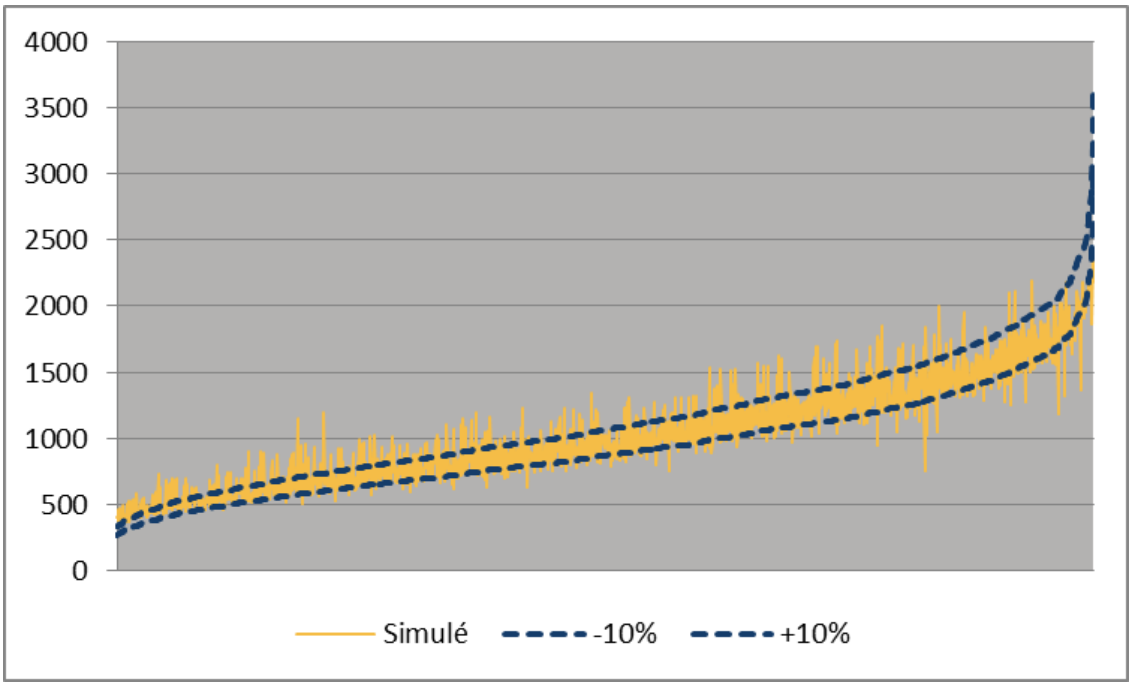

*Graphique 19 : Evolution de l'équivalent euro des garanties prothèses dentaires acceptées*

Pour la garantie visites de généralistes conventionnés, l'hypothèse faite pour les garanties en frais réels et pour les dépassements est dans la continuité de l'évolution de l'équivalent euro de la garantie. Par contre, pour les prothèses dentaires, on peut remarquer une forte augmentation de l'équivalent euro à partir du moment où ces types de quantification des garanties sont utilisés.

En moyenne, la modélisation par réseau de neurones en n'utilisant que 18 variables explicatives permet de retrouver le tarif initial avec plus ou moins 10% d'erreur.

Une variabilité dans les estimations peut se retrouver par le fait que les poids inter-synaptiques ont été initialisés aléatoirement. Pour tenter de la réduire, 20 réseaux de neurones ont été initialisés avec les données d'apprentissage et des poids synaptiques différents, puis pour chaque réseau créé, une prédiction a été réalisée. En réalisant la moyenne des prédictions, on peut réduire l'erreur causée par l'initialisation des poids.

$$RMSE = 134{,}48 \;\; VAR = 0{,}95\%$$




*Graphique 20 : Moyenne des estimations des tarifs en sortie des 20 réseaux de neurones*

L'utilisation de la moyenne de 20 prédictions permet de faire passer le RMSE de 148,32 à 134,48.

La moyenne des erreurs dans ce cas passe à plus ou moins 9% par rapport aux tarifs originaux. Par contre, cela ne résout pas la problématique des tarifs très élevés.

## 5) MODIFICATION DE LA BASE D'APPRENTISSAGE

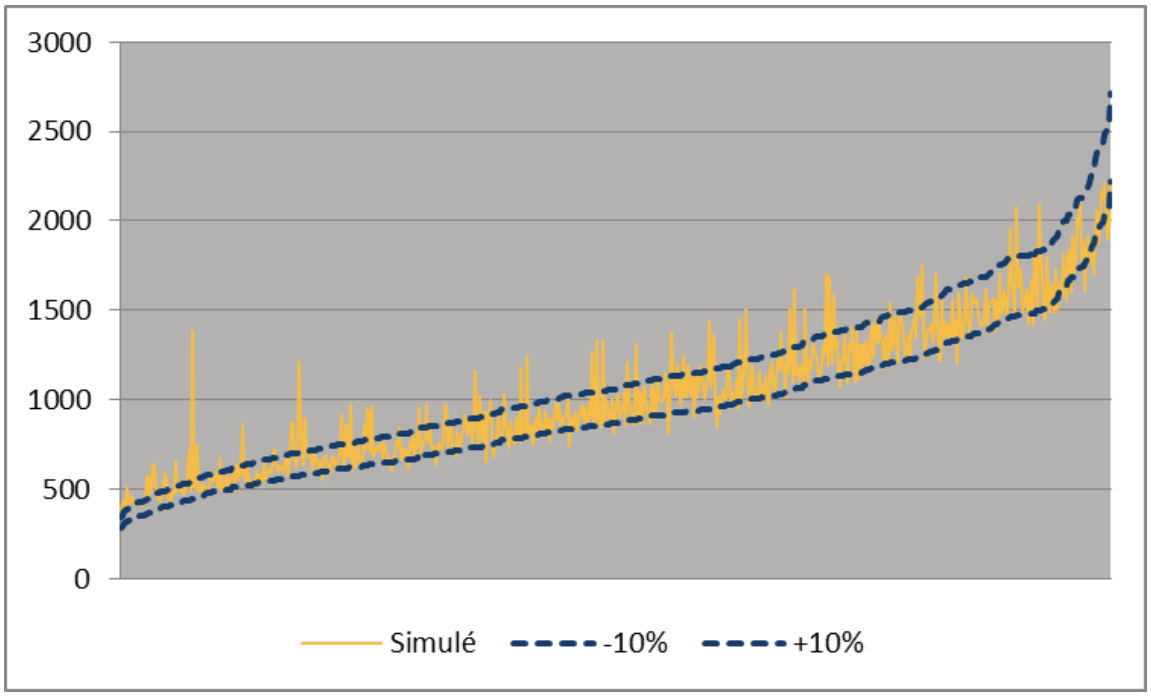
Il a été vu avec l'algorithme CART que les données n'étaient pas soumises au sur-apprentissage car il ne s'agit pas de variables aléatoires. De ce fait, la répartition entre 70% de la base de données pour l'apprentissage et 30% pour la base de test peut être modifiée. L'augmentation de la taille de la base d'apprentissage a pour but de permettre au réseau de mieux apprendre sur les tarifs extrêmes pour finalement diminuer son erreur de prédiction.

Par la suite, nous allons utiliser le même type de réseau de neurones que dans la partie précédente, cela signifie un perceptron multicouche à 8 neurones cachés et une fonction d'activation sigmoïde. Par contre la base d'apprentissage va passer à 90% et la base de test à 10%.

En réalisant la moyenne des prédictions de 20 réseaux de neurones, on obtient les résultats suivants :

$$RMSE = 129{,}19 \;\; VAR = 1{,}36\%$$

Graphiquement on a :




*Graphique 21 : Moyenne des estimations des tarifs en sortie des 20 réseaux de neurones*

L'utilisation d'une base d'apprentissage plus importante permet d'améliorer légèrement l'estimation des tarifs élevés. En revanche, l'erreur moyenne commise reste à plus ou moins 9% du tarif réel. Le modèle n'a que très légèrement été amélioré.

Ainsi, à partir d'un nombre réduit de variables, on peut réobtenir la valeur du tarif uniforme d'une complémentaire santé avec plus ou moins 9% d'erreur.

En conclusion, les réseaux de neurones permettent d'obtenir une approximation du tarif uniforme à partir d'un nombre réduit de variables. Par contre cette estimation n'est pas des plus précises. Pour contourner cela d'autres méthodes d'apprentissage peuvent être mises en place pour réduire l'erreur. Dans la suite, les méthodes boostées seront abordées pour tenter d'améliorer les prédictions.

# IV) MÉTHODES BOOSTÉES

Les réseaux de neurones ont servi à obtenir un modèle permettant d'avoir une estimation des tarifs uniformes en sur-mesure à partir d'un nombre réduit de variables. Cependant, l'erreur d'estimation est en moyenne de 9%. De ce fait, il parait utile de trouver une nouvelle modélisation permettant de réduire cette erreur.

Pour parvenir à cela, on s'intéressera plus particulièrement aux méthodes boostées et dans un premier temps à la méthode du gradient boosté. Cette partie se réfère aux travaux de J. FRIEDMAN [2001].

## A) MÉTHODE DU GRADIENT BOOSTÉ

### 1) MINIMISATION DE L'ERREUR

Le but de cette méthode est d'obtenir une estimation des fonctions dans le but de prédictions. Comme pour les réseaux de neurones, son principe est d'expliquer une certaine variable réponse par le biais de variables explicatives en utilisant une base de données d'apprentissage.

Pour plus de clarté, on notera $x = \{x_1, \dots, x_n\}$ le vecteur des variables explicatives conservées et $y$ le vecteur des tarifs uniformes. Le but est donc de trouver une fonction notée $\hat{F}(x)$ qui approxime la fonction réelle reliant les variables explicatives et le tarif noté $F^*(x)$.

Le principe est de minimiser l'erreur d'estimation modélisée par une certaine fonction $L\big(y, F(x)\big)$. Comme pour les réseaux de neurones, cette fonction restera le RMSE pour pouvoir comparer plus facilement les méthodes. Mathématiquement, on peut écrire le problème de minimisation de l'erreur de la façon suivante :

$$F^* = \underset{F}{\operatorname{Argmin}}\, \mathbb{E}_{y,x}\big[L\big(y, F(x)\big)\big] = \underset{F}{\operatorname{Argmin}}\, \mathbb{E}_x\big[\mathbb{E}_y\big(L\big(y, F(x)\big)\big|x\big)\big]$$

### 2) OPTIMISATION

Au vu du grand nombre de paramètres inclus dans la modélisation, nous supposerons que la fonction réelle reliant les variables explicatives et les tarifs s'écrit sous la forme suivante :

$$F^*(x) = \sum_{m=0}^{M} f_m(x)$$

Les différents $f_m$ sont obtenus par l'utilisation de la théorie de la descente du gradient.

$$f_m(x) = -\rho_m g_m(x)$$

Avec :

$$g_m(x) = \mathbb{E}_y\left[\frac{\partial L\big(y\,, F(x)\big)}{\partial F(x)} \mid x\right]_{F(x)=F_{m-1}(x)}$$

$$F_{m-1}(x) = \sum_{i=0}^{m-1} f_i(x)$$

$$\rho_m = \underset{\rho}{\text{Argmin}}\, \mathbb{E}_{y,x}[L(y, F_{m-1}(x) - \rho g_m)]$$

Pour plus de simplicité, nous supposerons que $F$ est une fonction de type additive et qu'elle peut s'écrire sous la forme suivante :

$$F(x;\, \{\beta_m\,;\, a_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(x\,;a_m)$$

Avec $a = \{a_1, \dots, a_M\}$ un vecteur de paramètres et $h$ une fonction à paramètre des variables d'entrée.

Le problème de minimisation revient donc à :

$$\{\beta_m\,;\, a_m\}_1^M = \underset{\{\beta'_m\,;\, a'_m\}_1^M}{\text{Argmin}} \sum_{i=1}^{N} L\left(y_i\,; \sum_{m=1}^{M} \beta'_m h(x_i\,;\, a'_m)\right)$$

$h$ est un classifieur faible qui va permettre de réaliser les prédictions. $F_{m-1}$ peut être assimilée à une première estimation qui va ensuite être améliorée par le biais d'une nouvelle estimation.
On obtient ainsi :

$$a_m = \underset{a\,;\,\beta}{\text{Argmin}} \sum_{i=1}^{N} \big(-g_m(x_i) - \beta h(x_i\,;a)\big)^2$$

$$\rho_m = \underset{\rho}{\text{Argmin}} \sum_{i=1}^{N} L\big(y_i\,;\, F_{m-1}(x_i) + \rho h(x_i\,;\, a_m)\big)$$

$g_m$ est également appelé pseudo résidu du modèle.
Finalement, l'utilisation des méthodes boostées dans l'estimation de la fonction reliant les variables explicatives à la variable à expliquer est :

$$F_m(x) = F_{m-1}(x) + \rho_m h(x\,;\, a_m)$$

## B) ARBRES DE RÉGRESSION BOOSTÉS

On s'intéresse ici au cas où les classifieurs faibles sont en fait des arbres de régression. En effet, un arbre de régression permet d'obtenir des séparations de l'espace où la valeur de la fonction à approximer est constante. De ce fait, on peut écrire mathématiquement un arbre de régression sous la forme suivante :

$$h\left(x\,;\,\{b_j\,;\,R_j\}_1^J\right) = \sum_{j=1}^{J} b_j \mathbf{1}_{\{x\,\in\,R_j\}}$$

$\{R_j\}_1^J$ correspond à un sous espace où la valeur est commune à chacune des prédictions. Il s'agit tout simplement d'une feuille terminale d'un arbre dont $b_j$ est la valeur estimée.

En reprenant la théorie du gradient boosté et en l'appliquant au cas particulier des arbres de régression, on obtient l'évolution de l'estimation de la fonction reliant les variables explicatives à la variable à expliquer.

$$F_m(x) = F_{m-1}(x) + \rho_m \sum_{j=1}^{J} b_{jm} \mathbf{1}_{\{x\,\in\,R_{jm}\}}$$

$\{R_{jm}\}_1^J$ correspond à un sous espace défini par un arbre de régression à la m$^{\text{ième}}$ itération et $b_{jm}$ la valeur associée.
En posant $\gamma_{jm} = \rho_m b_{jm}$ on obtient l'évolution de l'estimation de la fonction à prédire suivante :

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} \mathbf{1}_{\{x\,\in\,R_{jm}\}}$$

Avec :

$$\gamma_{jm} = \underset{\gamma}{\text{Argmin}} \sum_{x_i\,\in\,R_{jm}} L(y_i\,;\,F_{m-1}(x_i) + \gamma)$$

Cette méthode permet d'obtenir les constantes optimales sur chacune des régions et d'améliorer la précision au fur et à mesure des itérations.

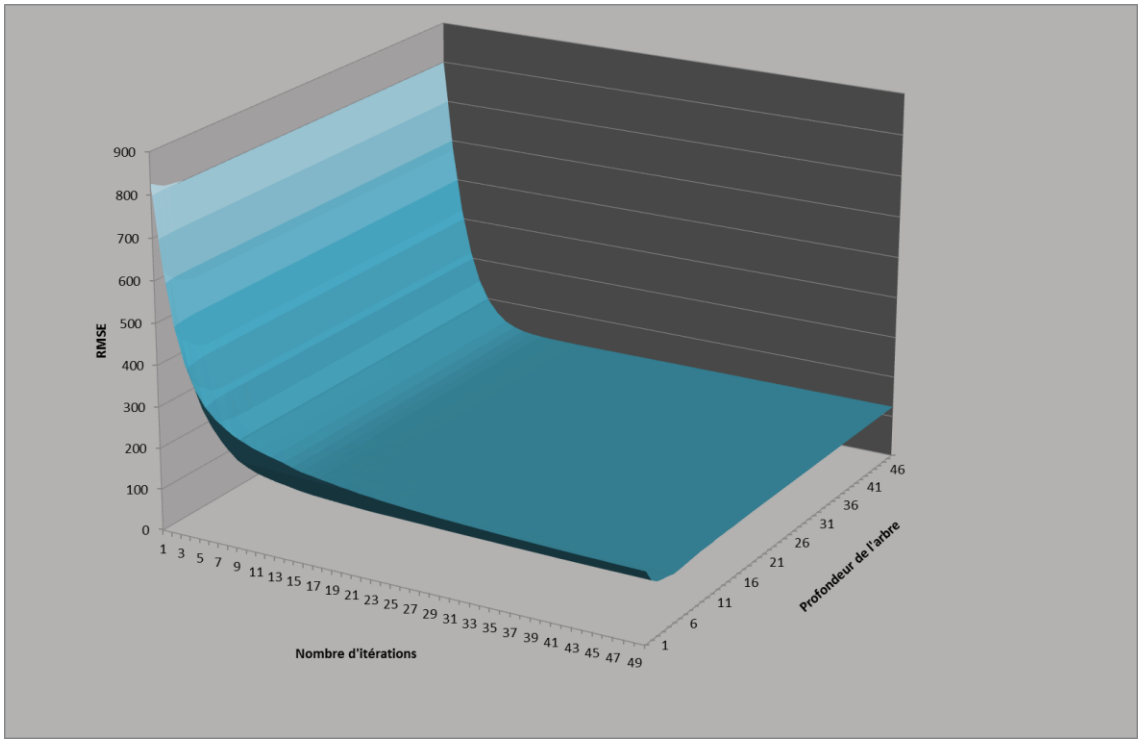
## C) APPLICATION AUX DONNÉES

Nous allons utiliser des arbres de régression boostés pour améliorer la prédiction des tarifs à partir d'un nombre réduit de variables.

Dans un premier temps, la base de données est séparée en deux parties. Comme il ne s'agit pas de données présentant un caractère aléatoire, on conservera 90% de la base de données en tant que base d'apprentissage et 10% de la base pour tester les prédictions.

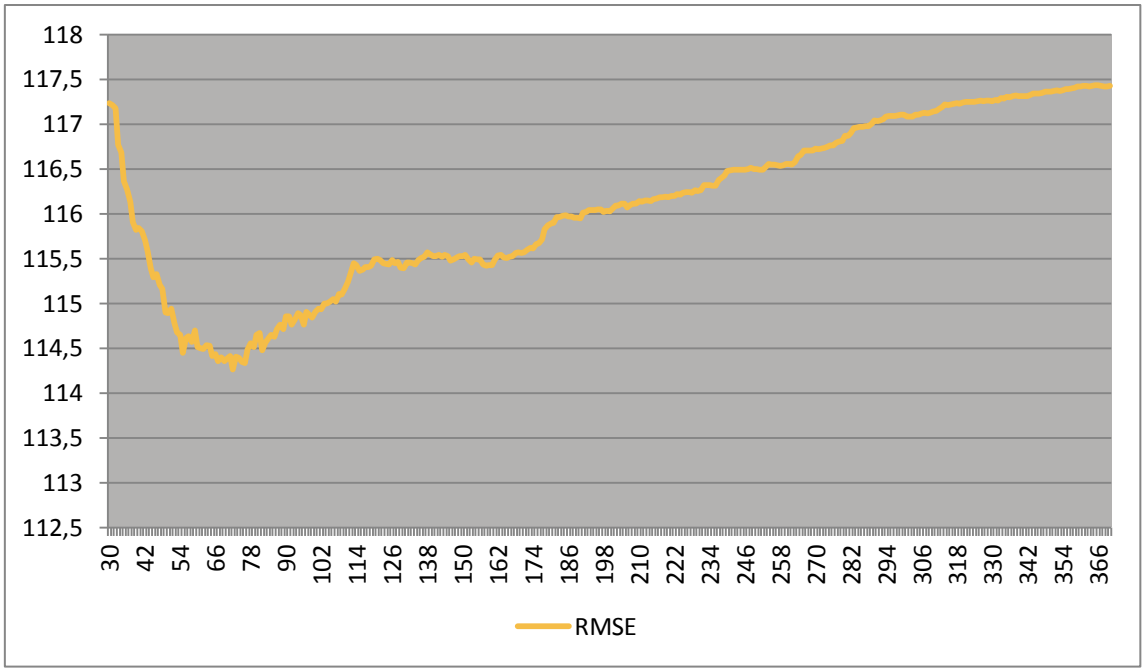
Deux paramètres sont importants pour déterminer le meilleur modèle. Il y a d'une part la taille maximale de l'arbre de régression et le nombre d'itérations voulues. Pour réduire au maximum

l'erreur du modèle, nous allons réaliser des prédictions pour un nombre de nœuds allant de 1 à 50 et un nombre d'itérations de 1 à 50 également. On obtient l'évolution de l'erreur suivante :




*Graphique 22 : Evolution de l'erreur de prédiction pour l'arbre de régression boosté*

On peut remarquer que plus le nombre d'itérations augmente, plus l'erreur diminue. D'autre part, la profondeur de l'arbre à moins d'impact à partir de 20 nœuds. Finalement, au vu des scénarii testés, il semblerait que le meilleur modèle soit celui défini pour une profondeur égale à 6 et pour 50 itérations. Pour s'assurer qu'il s'agisse du meilleur modèle, nous allons augmenter le nombre d'itérations à profondeur constante égale à 6.

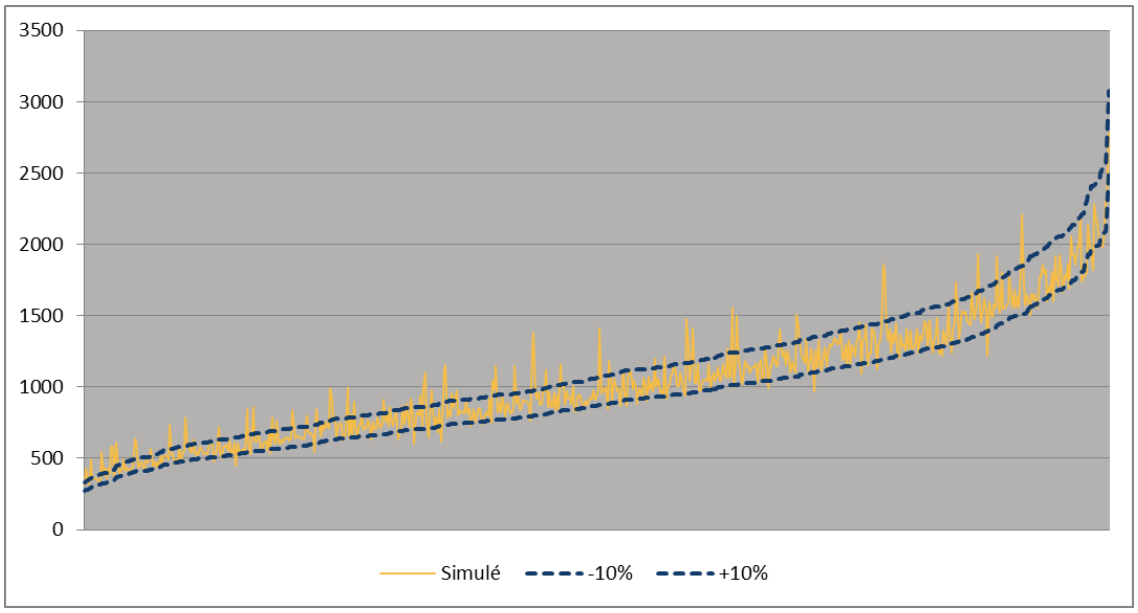



*Graphique 23 : Evolution du RMSE en fonction du nombre d'itérations pour une profondeur de 6*

La valeur la plus faible est obtenue pour un nombre d'itérations égal à 72. Ensuite, on peut observer une augmentation de la valeur du RMSE. Ceci est dû au phénomène de surapprentissage. En effet, en augmentant autant le nombre d'itérations, le modèle prédit parfaitement les données d'apprentissage mais devient de moins en moins généralisable.

Ainsi, grâce à ce modèle on obtient les résultats suivants :

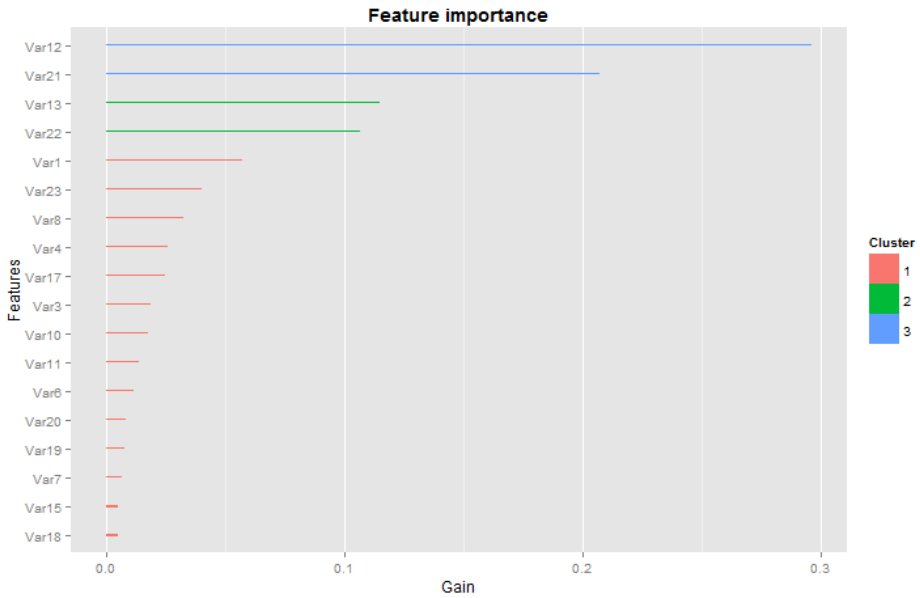
$$RMSE = 112{,}92 \;\; VAR = 0{,}85\%$$




*Graphique 24 : Estimations des tarifs par l'utilisation d'arbres boostés*

Finalement l'erreur moyenne du modèle est passée à plus ou moins 8%. Les arbres boostés ont aussi permis de réduire le RMSE d'environ 14%. On peut également observer une amélioration de l'estimation des tarifs élevés comparé aux réseaux de neurones.

L'utilisation d'arbres boostés nous permet de connaître l'apport de chaque variable dans le modèle. On peut les séparer en trois catégories comme le montre le graphique ci-dessous :




*Graphique 25 : Apport de chacune des variables dans la prédiction*

Les variables les plus importantes lorsque l'on utilise les arbres boostés sont donc la garantie orthodontie acceptée et le pourcentage de mariés, viennent ensuite la garantie prothèses dentaires et l'âge moyen.

## D) COMPARAISON DES RÉSULTATS

Dans ce mémoire, quatre modélisations ont été utilisées pour obtenir une estimation du tarif uniforme à partir d'un nombre réduit de variables conservées par l'algorithme CART. Les informations principales de ces modèles sont regroupées dans le tableau suivant :

| | Variables principales | | | RMSE |
|---|---|---|---|---|
| | 1ère Variable | 2ème Variable | 3ème Variable | |
| **Algorithme CART** | Orthodontie acc. | Prothèse dent. Acc. | % Mariés | |
| **Réseau de neurones Base d'apprentissage 70%** | Prothèse dent. Acc. | % Mariés | Age moyen | 148,32 |
| **20 Réseaux de neurones Base d'apprentissage 70%** | | | | 134,48 |
| **20 Réseaux de neurones Base d'apprentissage 90%** | | | | 129,19 |
| **Arbres boostés Base d'apprentissage 90%** | Orthodontie acc. | % Mariés | Prothèse dent. Acc. | 112,92 |

*Tableau 10 : Récapitulatif des différents modèles*

Dans un premier temps, on peut observer qu'il est possible d'améliorer la prédiction des réseaux de neurones en réalisant une moyenne des estimations de plusieurs réseaux ayant les mêmes caractéristiques. L'augmentation de la taille de la base d'apprentissage participe également à l'amélioration des prédictions. Les arbres boostés permettent d'obtenir une meilleure estimation avec une réduction de 30% du RMSE par rapport à un réseau de neurones classique.

En ce qui concerne les variables les plus importantes dans chacun des modèles, on remarque que parmi celles-ci, les réseaux de neurones conservent deux variables démographiques (âge moyen et pourcentage de mariés) ainsi qu'une variable concernant des garanties dentaires. Les arbres boostés ainsi que l'algorithme CART en conservent deux concernant les garanties dentaires et une seule démographique.

Ce résultat est intéressant pour plusieurs raisons. Tout d'abord, il montre l'importance d'avoir une démographie propre et complète pour l'obtention d'un tarif. Les hypothèses démographiques en cas de données manquantes doivent être prises en ne perdant pas de vue leur impact sur les tarifs finaux. Ainsi, les simples règles de souscription indiquant de prendre un pourcentage de mariés à 70% ou encore un nombre moyen d'enfant à 1 sont à proscrire au profit de résultats issus de statistiques démographiques plus poussées qui se basent notamment sur le secteur d'activité de l'entreprise.

D'autre part, un résultat plus inattendu est apparu dans cette étude. Il s'agit de l'importance des garanties dentaires dans l'obtention du tarif. Les garanties dentaires représentent une part importante du coût d'une protection santé mais ce n'est pas l'unique poste de dépense. On aurait pu s'attendre à une place plus importante des variables optique au vu de leurs coûts élevés.

Finalement, le meilleur modèle mathématique est celui composé d'arbres boostés. Il est le meilleur car il permet une réduction du RMSE par rapport aux réseaux de neurones. Ainsi, si l'on reprend la problématique initiale qui était de permettre une vérification simple ou l'obtention d'une estimation, la méthode des arbres boostés semble la plus appropriée. En effet, il est très facile de retranscrire les résultats de l'arbre dans une maquette de calcul en important les conditions entre chaque branche. Ainsi, il suffira au souscripteur ou même au commercial en charge du dossier d'utiliser cette maquette pour voir plus rapidement si le tarif obtenu après une tarification sur-mesure ne présente pas d'erreur importante ou encore si un concurrent a réalisé du *dumping* tarifaire pour une entreprise en particulier.

# CONCLUSION GÉNÉRALE

Dans ce mémoire, nous nous sommes intéressés à la tarification de complémentaires santé collectives et plus particulièrement à la réduction du nombre de variables nécessaires à la tarification.

La première étape de l'étude a consisté à déterminer les variables les plus importantes pour la tarification. Pour cela, une base de données correspondant à l'historique des études santé tarifées par le cabinet ACTUARIS depuis 2009 regroupant les données démographiques et les garanties santé a été utilisée. Grâce à l'algorithme CART, il a été possible de réduire une première fois le nombre important de variables tarifaires. L'étude des corrélations des variables conservées a montré que l'on pouvait réduire de nouveau ce nombre en ne prenant pas en considération les variables très corrélées. Pour finir, des méthodes GLM ont été utilisées pour confirmer et infirmer le choix de conserver telle ou telle variable grâce aux tests de type III.

Finalement, le modèle est passé d'une dimension 149 à une dimension 18. Toutes les variables concernant les caractéristiques démographiques de l'entreprise ont été conservées à l'exception de l'effectif global et du pourcentage de cadres. La liste des variables conservées est la suivante (par ordre d'importance) :

- Dentaire Orthodontie Acceptée
- % Marié
- Dentaire Prothèses Acceptées Fixes métal non visibles
- Age moyen
- Honoraires – Pharmacie Actes courants Radiologie
- Nombre moyen d'enfants
- Hospitalisation – Chirurgie accessoire Forfait journalier
- Honoraires – Pharmacie Petite Chirurgie conventionnée ADA
- Optique Verres
- Honoraires – Pharmacie Consultations – Visites conventionnées Spécialiste Hors PS
- Hospitalisation – Chirurgie Non conventionnées Frais de séjour
- Dentaire Hors nomenclature Implantologie
- Honoraires – Pharmacie Petite Chirurgie Non conventionnée ATM
- % Hommes
- Autre maternité Chambre particulière
- Hospitalisation – Chirurgie Accessoires Chambre particulière
- Optique Lentille (unité) Refusée
- Autres Cures thermales Honoraires

En réalité, il n'y a que 17 variables conservées pour la modélisation car la garantie forfait journalier est toujours égale à 100% des frais réels.

Dans un deuxième temps, un modèle permettant d'estimer le tarif uniforme à partir de ces variables a été créé par le biais de réseaux de neurones. Le réseau conservé est un perceptron multicouche à 8 neurones cachés avec un coefficient de dégradation des pondérations à 0,5 et des fonctions d'activation sigmoïdes. L'utilisation de ce réseau dans le cadre d'un apprentissage supervisé nous permet de retrouver les tarifs initiaux avec une erreur moyenne de 9%.

Un modèle plus performant permet de réduire l'erreur moyenne. Il s'agit de l'utilisation d'arbres boostés par la méthode du gradient boosté. En effet, cette méthode permet de réduire le RMSE de 14% par rapport aux réseaux de neurones. Ce modèle basé sur des arbres de régression présente également l'intérêt de pouvoir être facilement retranscrit dans une maquette de calcul utilisable par les souscripteurs pour vérifier leur tarification et même par les commerciaux pour éviter la demande d'une tarification sur-mesure complète lorsqu'un concurrent s'est positionné de manière très compétitive. Il ne faut cependant pas perdre de vue que la tarification d'un contrat complémentaire santé reste un exercice complexe et que la connaissance de l'ensemble des variables tant au niveau des garanties qu'au niveau démographique est nécessaire pour obtenir un tarif en adéquation totale avec le risque couvert.

Dans ce mémoire plusieurs méthodes ont été utilisées pour réduire l'erreur d'estimation. Il serait intéressant de voir l'intérêt de l'utilisation d'autres méthodes comme l'algorithme MARS ou encore de l'appliquer à un modèle tarifaire de type barème pour l'estimation du tarif uniforme à partir d'un nombre réduit de variables.

# BIBLIOGRAPHIE

AOUIZERATE J-M [2010], « Alternative neuronale en tarification santé », *Mémoire d'actuaire*

DENUIT M., CHARPENTIER A. [2005] « Mathématiques de l'assurance non-vie, Tome II : tarification et provisionnement », *Economica*

FRIEDMAN J. [2001], « Greedy Function Approximation A Gradient Boosting Machine »

FRIEDMAN J. [1999], « Stochastic Gradient Boosting »

FYFE C. [2000], « Artificial neural networks and information theory », *Edition 1.2*

GARSON G. [1991], « Interpreting neural network connection weights », *Artificial Intelligence Expert*

HASTIE T., TIBSHIRANI R., RIEDMAN J. [2009], « The element of statistical learning », *Data mining, inference and prediction, Springer series in statistics 2nd edition*

HSIEM W. [2009], « Machine learning method in the environmental sciences », *Neural networks and Kernels, Cambridge*

MERDRIGNAC M. [2011], « Modélisation de la consommation santé en fonction de variables exogènes », *Mémoire d'actuaire*

NOZACH J., PARADIS C. [2015], « Deux ans après l'ANI où en sont les accords de branches ? », *Infotech Actuaris*

NOZACH J. [2015], « Contrats responsables : décrypter et comprendre la circulaire du 30 janvier 2015 », Infotech Actuaris

ROUSVOAL J. [2014], « Etude de la consommation en assurance santé pour les garanties « entrée de gamme » et influence de la localisation », *Mémoire d'actuaire*

STRICKER M. [2000], « Réseaux de neurones pour le traitement automatique de langage : conception et réalisation de filtres d'informations », *Thèse*

---

AMELI : Base CCAM activité bucco-dentaire :
http://www.ameli.fr/fileadmin/user_upload/documents/Base_actes_transposes_dentistes.pdf

# ANNEXES

**ANNEXE 1** : Remboursements de la Sécurité Sociale utilisés pour le calcul des équivalents euro

## Régime de base santé

| | |
|---|---|
| Montant de la contribution forfaitaire | 1,00 € |
| Montant du dépassement d'honoraires non pris en charge | 8,00 € |

| | Conv. | Taux | Assiette | (enfants) | TC Rec. | Contrib. Forf. |
|---|---|---|---|---|---|---|
| Honoraires – Pharmacie | | | | | | |
| Consultations – Visites | | | | | | |
| Conventionnées | | | | | | |
| Consultations de généralistes | | | | | | |
| Médecin traitant (PS) | VRAI | 70,00% | 23,00 | | | VRAI |
| Médecin correspondant (PS) | VRAI | 70,00% | 25,55 | | | VRAI |
| Généraliste (Hors PS) | VRAI | 70,00% | 23,00 | | | VRAI |
| Visites de généraliste | VRAI | 70,00% | 33,00 | | | VRAI |
| Spécialiste | | | | | | |
| Spécialiste traitant (PS) | VRAI | 70,00% | 26,00 | | | VRAI |
| Spécialiste régulier (PS) | VRAI | 70,00% | 44,00 | | | VRAI |
| Spécialiste ponctuel (PS) | VRAI | 70,00% | 26,25 | | | VRAI |
| Spécialiste (Hors PS) | VRAI | 70,00% | 26,00 | | | VRAI |
| NeuroPsychiatre | VRAI | 70,00% | 39,70 | | | VRAI |
| Non conventionnées | FAUX | 70,00% | 1,07 | | 23,00 | VRAI |
| Petite Chirurgie | | | | | | |
| Conventionnée | | | | | | |
| ATM | VRAI | 70,00% | 2,09 | | | FAUX |
| ADC | VRAI | 70,00% | 2,09 | | | FAUX |
| ACO | VRAI | 70,00% | 2,09 | | | FAUX |
| ADA | VRAI | 70,00% | 2,09 | | | FAUX |
| ADE | VRAI | 70,00% | 2,09 | | | FAUX |
| Non conventionnée | | | | | | |
| ATM | FAUX | 70,00% | 0,38 | | 2,09 | FAUX |
| ADC | FAUX | 70,00% | 0,38 | | 2,09 | FAUX |
| ACO | FAUX | 70,00% | 0,38 | | 2,09 | FAUX |
| ADA | FAUX | 70,00% | 0,38 | | 2,09 | FAUX |
| ADE | FAUX | 70,00% | 0,38 | | 2,09 | FAUX |
| Actes courants | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Radiologie | VRAI | 70,00% | 1,62 | | | VRAI |
| Analyses médicales | VRAI | 60,00% | 0,27 | | | VRAI |
| Auxiliaires médicaux | VRAI | 60,00% | 2,04 | | | FAUX |
| Pharmacie | | | | | | |
| Vignettes blanches 65% | VRAI | 65,00% | 152,45 | | | FAUX |
| Vignettes bleues 30% | VRAI | 30,00% | 152,45 | | | FAUX |
| Vignettes oranges 15% | VRAI | 15,00% | 152,45 | | | FAUX |
| Hospitalisation - Chirurgie | | | | | | |
| Conventionnées | | | | | | |
| Honoraires chirurgicaux | | | | | | |
| ATM | VRAI | 80,00% | 2,09 | | | FAUX |
| ADC | VRAI | 80,00% | 2,09 | | | FAUX |
| ACO | VRAI | 80,00% | 2,09 | | | FAUX |
| ADA | VRAI | 80,00% | 2,09 | | | FAUX |
| ADE | VRAI | 80,00% | 2,09 | | | FAUX |
| Frais de séjour | VRAI | 80,00% | 156,86 | | | FAUX |
| Non conventionnées | | | | | | |
| Honoraires chirurgicaux | | | | | | |
| ATM | FAUX | 80,00% | 0,29 | | 2,09 | FAUX |
| ADC | FAUX | 80,00% | 0,29 | | 2,09 | FAUX |
| ACO | FAUX | 80,00% | 0,29 | | 2,09 | FAUX |
| ADA | FAUX | 80,00% | 0,29 | | 2,09 | FAUX |
| ADE | FAUX | 80,00% | 0,29 | | 2,09 | FAUX |
| Frais de séjour | FAUX | 80,00% | 22,87 | | 156,86 | FAUX |
| Accessoires | | | | | | |
| Chambre particulière | | | | | | FAUX |
| Frais d'accompagnant | | | | | | FAUX |
| Forfait journalier | | | 18,00 | | | FAUX |
| Transport | VRAI | 65,00% | 53,36 | | | FAUX |
| Dentaire | | | | | | |
| Soins dentaires | VRAI | 70,00% | 2,41 | | | FAUX |
| Prothèses | | | | | | |
| Acceptées | | | | | | |
| Fixes métal visibles | VRAI | 70,00% | 2,15 | | | FAUX |
| Fixes métal non visibles | VRAI | 70,00% | 2,15 | | | FAUX |
| Fixes esthétiques visibles | VRAI | 70,00% | 2,15 | | | FAUX |
| Fixes esthétiques non visibles | VRAI | 70,00% | 2,15 | | | FAUX |
| Amovibles | VRAI | 70,00% | 2,15 | | | FAUX |
| Actes divers | VRAI | 70,00% | 2,15 | | | FAUX |
| Refusées | FAUX | 0,00% | 0,00 | | 2,15 | FAUX |
| Inlays / Onlays acceptés | VRAI | 0,00% | 0,00 | | | FAUX |
| Orthodontie | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Acceptée | VRAI | 100,00% | 2,15 | | | FAUX |
| Refusée | FAUX | 0,00% | 0,00 | | 2,15 | FAUX |
| Hors nomenclature | | | | | | |
| Parodontologie | FAUX | 0,00% | 0,00 | | 0,00 | FAUX |
| Implantologie | FAUX | 0,00% | 0,00 | | 0,00 | FAUX |
| Prophylaxie bucco-dentaire | FAUX | 0,00% | 0,00 | | 0,00 | FAUX |
| Optique | | | | | | |
| Verres (la paire) | | | | | | |
| Verres (la paire) | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Unifocaux | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Multifocaux | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Hypercomplexes | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Faible correction | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Moyenne correction | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Forte correction | VRAI | 60,00% | 15,24 | 45,73 | | FAUX |
| Montures | VRAI | 60,00% | 2,84 | 30,49 | | FAUX |
| Lentille ( unité) | | | | | | |
| Acceptée | VRAI | 60,00% | 39,48 | | | FAUX |
| Refusée | FAUX | 0,00% | 0,00 | | 39,48 | FAUX |
| Autres | | | | | | |
| Prothèses – Orthopédie | | | | | | |
| Prothèses auditives ( unité ) | VRAI | 60,00% | 199,71 | | | FAUX |
| Autres prothèses | VRAI | 60,00% | 152,45 | | | FAUX |
| Orthopédie | VRAI | 60,00% | 152,45 | | | FAUX |
| Cures thermales | | | | | | |
| Honoraires | VRAI | 70,00% | 64,03 | | | FAUX |
| Forfait | VRAI | 65,00% | 226,22 | | | FAUX |
| Maternité | | | | | | |
| Chambre particulière | FAUX | | | | | FAUX |
| Forfait | FAUX | | | | | FAUX |
| Allocation frais d'obsèques | FAUX | | | | | FAUX |

**ANNEXE 2** : Listes des variables présentes dans le modèle

| Code variable | Intitulé |
|---|---|
| X2 | Honoraires - Pharmacie Consultations - Visites Conventionnées Consultations de généralistes Visites de généraliste |
| X4 | Honoraires - Pharmacie Consultations - Visites Conventionnées NeuroPsychiatre |
| X5 | Honoraires - Pharmacie Consultations - Visites Conventionnées Non conventionnées |
| X8 | Honoraires - Pharmacie Actes courants Radiologie |
| X9 | Honoraires - Pharmacie Actes courants Analyses médicales |
| X10 | Honoraires - Pharmacie Actes courants Auxiliaires médicaux |
| X11 | Honoraires - Pharmacie Pharmacie Vignettes blanches 65% |
| X12 | Honoraires - Pharmacie Pharmacie Vignettes bleues 35% |
| X14 | Hospitalisation - Chirurgie Conventionnées Frais de séjour |
| X16 | Hospitalisation - Chirurgie Non conventionnées Frais de séjour |
| X18 | Hospitalisation - Chirurgie Accessoires Chambre particulière |
| X19 | Hospitalisation - Chirurgie Accessoires Frais d'accompagnant |
| X20 | Hospitalisation - Chirurgie Accessoires Forfait journalier |
| X21 | Hospitalisation - Chirurgie Accessoires Transport |
| X22 | Dentaire Soins dentaires |
| X24 | Dentaire Prothèses refusées |
| X25 | Dentaire Orthodontie Acceptée |
| X26 | Dentaire Orthodontie Refusée |
| X27 | Optique Verres (la paire) |
| X28 | Optique Montures |
| X29 | Optique Lentille ( unité) Acceptée |
| X30 | Optique Lentille ( unité) Refusée |
| X32 | Autres Prothèses - Orthopédie Prothèses auditives ( unité ) |
| X33 | Autres Prothèses - Orthopédie Autres prothèses |
| X34 | Autres Prothèses - Orthopédie Orthopédie |
| X35 | Autres Cures thermales Honoraires |
| X38 | Autres Maternité Chambre particulière |
| X39 | Autres Maternité Forfait |
| X40 | Autres Allocation frais d'obsèques |
| X44 | Spécifique / Assistance VIVACTIV |
| X45 | Spécifique / Spécifique2 |
| X46 | Spécifique / Spécifique3 |
| X47 | Spécifique / Spécifique4 |
| X48 | Spécifique / Spécifique5 |
| X49 | Spécifique / Spécifique6 |
| X50 | Spécifique / Spécifique7 |
| X51 | Spécifique / Spécifique8 |

| | |
|---|---|
| X52 | Spécifique / Spécifique9 |
| X53 | Spécifique / Spécifique10 |
| X54 | Spécifique / Bien Etre VIVACTIV |
| X55 | Spécifique / Spécifique11 |
| X56 | Spécifique / Spécifique12 |
| X57 | Spécifique / Spécifique13 |
| X58 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ATM |
| X59 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ADC |
| X60 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ACO |
| X61 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ADA |
| X62 | Honoraires - Pharmacie Petite Chirurgie Conventionnée ADE |
| X63 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ATM |
| X64 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ADC |
| X65 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ACO |
| X66 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ADA |
| X67 | Honoraires - Pharmacie Petite Chirurgie Non conventionnée ADE |
| X68 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ATM |
| X69 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ADC |
| X70 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ACO |
| X71 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ADA |
| X72 | Hospitalisation - Chirurgie Conventionnées Honoraires chirurgicaux ADE |
| X73 | Hospitalisation - Chirurgie Non conventionnées Honoraires chirurgicaux ATM |
| X74 | Hospitalisation - Chirurgie Non conventionnées Honoraires chirurgicaux ADC |
| X75 | Hospitalisation - Chirurgie Non conventionnées Honoraires chirurgicaux ACO |
| X76 | Hospitalisation - Chirurgie Non conventionnées Honoraires chirurgicaux ADA |
| X77 | Hospitalisation - Chirurgie Non conventionnées Honoraires chirurgicaux ADE |
| X78 | Dentaire Prothèses Acceptées Fixes métal visibles |
| X79 | Dentaire Prothèses Acceptées Fixes métal non visibles |
| X80 | Dentaire Prothèses Acceptées Fixes esthétiques visibles |
| X81 | Dentaire Prothèses Acceptées Fixes esthétiques non visibles |
| X82 | Dentaire Prothèses Acceptées Amovibles |
| X83 | Dentaire Prothèses Acceptées Actes divers |
| X84 | Dentaire Hors nomenclature Parodontologie |
| X85 | Dentaire Hors nomenclature Implantologie |

| | |
|---|---|
| X86 | Dentaire Hors nomenclature Prophylaxie bucco-dentaire |
| X87 | Optique Unifocaux |
| X88 | Optique Multifocaux |
| X89 | Optique Hypercomplexes |
| X90 | Optique Faible correction |
| X91 | Optique Moyenne correction |
| X92 | Optique Forte correction |
| X93 | Autres Cures thermales Forfait |
| X94 | Hors nomenclature / Kératochirurgie (par œil) |
| X95 | Hors nomenclature / Vaccins non remboursés |
| X96 | Hors nomenclature / Diététicien |
| X97 | Hors nomenclature / Ostéo / Chiro |
| X98 | Hors nomenclature / Pilule 3ème génération |
| X99 | Hors nomenclature / Vaccin antigrippe |
| X100 | Hors nomenclature / FIV |
| X101 | Hors nomenclature / Amniocentèse |
| X102 | Hors nomenclature / Acupuncteur |
| X103 | Hors nomenclature / Pilules contraceptives |
| X104 | Hors nomenclature / Ostéodensiométrie |
| X105 | Hors nomenclature / Médicaments prescrits |
| X106 | Hors nomenclature / Sevrage Tabac |
| X107 | Hors nomenclature / Pédicure/Podologue |
| X108 | Hors nomenclature / Laboratoire HN |
| X109 | Hors nomenclature / Transport refusé |
| X110 | Hors nomenclature / Pharmacie prescrite NR |
| X111 | Hors nomenclature / Pharmacie automédication |
| X112 | Hors nomenclature / Tests de grossesse |
| X113 | Hors nomenclature / Renfort HOSPI |
| X114 | Hors nomenclature / Prothèses capillaires |
| X115 | Hors nomenclature / Spécifique36 |
| X116 | Hors nomenclature / Spécifique37 |
| X117 | Hors nomenclature / Spécifique38 |
| X118 | Hors nomenclature / Spécifique39 |
| X119 | Hors nomenclature / Spécifique40 |
| X120 | Hors nomenclature / Spécifique41 |
| X121 | Hors nomenclature / Acte spécifique 42 |
| X122 | Hors nomenclature / Acte spécifique 43 |
| X123 | Hors nomenclature / Acte spécifique 44 |
| X124 | Hors nomenclature / Acte spécifique 45 |
| X125 | Hors nomenclature / Acte spécifique 46 |
| X126 | Hors nomenclature / Acte spécifique 47 |
| X127 | Hors nomenclature / Acte spécifique 48 |

| | |
|---|---|
| X128 | Hors nomenclature / Acte spécifique 49 |
| X129 | Hors nomenclature / Acte spécifique 50 |
| X130 | Honoraires - Pharmacie Consultations - Visites Conventionnées Consultations de généralistes Médecin traitant (PS) |
| X131 | Honoraires - Pharmacie Consultations - Visites Conventionnées Consultations de généralistes Médecin correspondant (PS) |
| X132 | Honoraires - Pharmacie Consultations - Visites Conventionnées Consultations de généralistes Généraliste (Hors PS) |
| X133 | Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste traitant (PS) |
| X134 | Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste régulier (PS) |
| X135 | Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste ponctuel (PS) |
| X136 | Honoraires - Pharmacie Consultations - Visites Conventionnées Spécialiste Spécialiste (Hors PS) |
| X137 | Honoraires Pharmacie / Consultations - Visite / Conventionnées / Consultations de généralistes / Autre consultation traitant PS |
| X138 | Honoraires Pharmacie / Consultations - Visite / Conventionnées / Spécialiste / Secteur II sans Cardiologues |
| X139 | Honoraires Pharmacie / Consultations - Visite / Conventionnées / Spécialiste / Cardiologue |
| X140 | Honoraires Pharmacie / Consultations - Visite / Conventionnées / Spécialiste / Autres consultations |
| X141 | Honoraires Pharmacie / Consultations - Visite / Conventionnées / Autres neuropsychiatres |
| X142 | Honoraires - Pharmacie Pharmacie Vignettes oranges 15% |
| X143 | Dentaire Inlays / Onlays acceptés |
| X144 | Dentaire Hors nomenclature Inlays / Onlays refusés |
| X157 | Optique / Verres (la paire) / Verres unifocaux classe 1 |
| X158 | Optique / Verres (la paire) / Verres unifocaux classe 2 |
| X159 | Optique / Verres (la paire) / Verres unifocaux classe 3 |
| X160 | Optique / Verres (la paire) / Verres unifocaux classe 4 |
| X161 | Optique / Verres (la paire) / Verres multifocaux classe 1 |
| X162 | Optique / Verres (la paire) / Verres multifocaux classe 2 |
| X163 | Optique / Verres (la paire) / Verres multifocaux classe 3 |
| X164 | Optique / Verres (la paire) / Verres multifocaux classe 4 |
| X165 | Optique / Verres (la paire) / Verres hypercomplexes classe 1 |
| X166 | Optique / Verres (la paire) / Verres hypercomplexes classe |
| X167 | Optique / Verres (la paire) / Verres hypercomplexes classe 3 |
| X168 | Optique / Verres (la paire) / Verres hypercomplexes classe 4 |
| X170 | Optique / Kératotomie |
| PO_EFFECTIF | Effectif de la société |
| PO_AGE_MOYEN | Age moyen des salariés de la société |
| PO___MARIES | Pourcentage de marié dans la société |

| | |
|---|---|
| PO_NOMBRE_MOYEN_D_ENFANTS | Nombre d'enfants moyen des salariés de la société |
| PO___HOMMES | Pourcentage d'hommes dans la société |
| PO__CADRES | Pourcentage de cadres dans la société |

**ANNEXE 3** : Conditions dans les nœuds de l'arbre élagué

1) root 6519 1176720000.0 1028.9820
  2) X25< 0.1542114 3319 236278000.0 784.2453
    4) PO___MARIES< 0.4998599 1957 82204200.0 667.2511
      8) X59< -0.6676609 1030 26629520.0 577.3437
        16) PO_AGE_MOYEN< -0.6705064 428 4729163.0 482.9462
          32) PO_NOMBRE_MOYEN_D_ENFANTS< -1.517063 156 517741.2 401.3116 *
          33) PO_NOMBRE_MOYEN_D_ENFANTS>=-1.517063 272 2575558.0 529.7659 *
        17) PO_AGE_MOYEN>=-0.6705064 602 15374970.0 644.4569
          34) X16< -0.07434867 437 9144191.0 605.3727 *
          35) X16>=-0.07434867 165 3795233.0 747.9709 *
      9) X59>=-0.6676609 927 37997910.0 767.1482
        18) PO_AGE_MOYEN< -0.3949488 430 8371967.0 640.7694
          36) PO_NOMBRE_MOYEN_D_ENFANTS< -1.075119 171 1289855.0 535.4756 *
          37) PO_NOMBRE_MOYEN_D_ENFANTS>=-1.075119 259 3934575.0 710.2878 *
        19) PO_AGE_MOYEN>=-0.3949488 497 16816210.0 876.4900
          38) X16< 0.0440754 318 7864812.0 815.7295 *
          39) X16>=0.0440754 179 5691730.0 984.4332 *
    5) PO___MARIES>=0.4998599 1362 88798440.0 952.3492
      10) X61< -0.6812033 724 23746600.0 820.4983
        20) PO_AGE_MOYEN< 0.2976766 404 10289560.0 742.4098
          40) X20< -2.157688 38 566128.9 483.2297 *
          41) X20>=-2.157688 366 6905776.0 769.3192
            82) X18< -0.8424767 175 1569615.0 683.6508 *
            83) X18>=-0.8424767 191 2875072.0 847.8112 *
        21) PO_AGE_MOYEN>=0.2976766 320 7883337.0 919.0850
          42) X72< -0.1580117 198 2905086.0 855.0634 *
          43) X72>=-0.1580117 122 2849578.0 1022.9890 *
      11) X61>=-0.6812033 638 38182240.0 1101.9730
        22) X20< -2.157688 54 1466158.0 670.9367 *
        23) X20>=-2.157688 584 25755610.0 1141.8290
          46) PO_AGE_MOYEN< 0.2907556 303 7845739.0 1035.5900
            92) X18< -0.2420851 200 3329310.0 982.9081 *
            93) X18>=-0.2420851 103 2883544.0 1137.8850 *
          47) PO_AGE_MOYEN>=0.2907556 281 10802320.0 1256.3860
            94) X28< -0.1528974 180 4321380.0 1188.7340 *
            95) X28>=-0.1528974 101 4188923.0 1376.9540 *
  3) X25>=0.1542114 3200 535459100.0 1282.8200
    6) PO___MARIES< 0.5763014 1700 163717700.0 1084.9140
      12) PO_NOMBRE_MOYEN_D_ENFANTS< -1.238496 353 13493340.0 793.1500
        24) X78< -0.07626057 153 2996548.0 680.8795 *

```
  25) X78>=-0.07626057 200    7092962.0  879.0370
    50) PO___MARIES< 0.09927029 193    5158653.0  863.6990 *
    51) PO___MARIES>=0.09927029 7     637054.4 1301.9270 *
 13) PO_NOMBRE_MOYEN_D_ENFANTS>=-1.238496 1347  112299800.0 1161.3750
  26) X8< 0.494638 674   32584800.0 1007.4790
    52) PO_AGE_MOYEN< 0.6020688 525   21705070.0  955.4130
     104) X20< -2.157688 44     858663.5  656.0110 *
     105) X20>=-2.157688 481   16541380.0  982.8012
      210) PO___MARIES< 0.1994622 358    9273550.0  931.1050
        420) X38< 0.2167163 214    4634415.0  881.3255 *
        421) X38>=0.2167163 144    3320769.0 1005.0830 *
      211) PO___MARIES>=0.1994622 123    3526384.0 1133.2660 *
    53) PO_AGE_MOYEN>=0.6020688 149    4441824.0 1190.9340 *
  27) X8>=0.494638 673   47765270.0 1315.5000
    54) PO_AGE_MOYEN< -0.5548613 246    8594129.0 1129.0010
     108) PO___HOMMES>=1.224923 72    1369003.0  986.4782 *
     109) PO___HOMMES< 1.224923 174    5157426.0 1187.9760 *
    55) PO_AGE_MOYEN>=-0.5548613 427   25685420.0 1422.9440
     110) PO_AGE_MOYEN< 1.110643 283   14521770.0 1348.1220
      220) X85< -0.03610372 128    4395945.0 1245.0340 *
      221) X85>=-0.03610372 155    7642220.0 1433.2530 *
     111) PO_AGE_MOYEN>=1.110643 144    6465662.0 1569.9900
      222) X35>=1.8844 72    1802833.0 1457.1950 *
      223) X35< 1.8844 72    2830733.0 1682.7860 *
7) PO___MARIES>=0.5763014 1500  229697300.0 1507.1130
 14) X28< 1.243883 1053   86423660.0 1354.0550
  28) X20< -2.633373 113    4757458.0  942.3096
    56) X79< -0.08883829 57    1334552.0  823.0858 *
    57) X79>=-0.08883829 56    1788007.0 1063.6620 *
  29) X20>=-2.633373 940   60205840.0 1403.5520
    58) X136< -0.2061748 395   16435760.0 1265.3930
     116) PO_AGE_MOYEN< -0.1083067 155    3964972.0 1166.9070
      232) X30< -0.5708761 85    1366314.0 1080.7630 *
      233) X30>=-0.5708761 70    1201945.0 1271.5110 *
     117) PO_AGE_MOYEN>=-0.1083067 240    9996438.0 1328.9980
      234) Verres< 0.3416538 197    6483881.0 1289.2340
        468) PO_AGE_MOYEN< 0.5678927 152    3616442.0 1241.5920 *
        469) PO_AGE_MOYEN>=0.5678927 45    1357065.0 1450.1600 *
      235) Verres>=0.3416538 43    1774065.0 1511.1690 *
    59) X136>=-0.2061748 545   30765660.0 1503.6860
     118) PO_AGE_MOYEN< 0.1239552 292    9839524.0 1394.4800
      236) X25< 1.359881 200    5380854.0 1335.9790 *
```

```
      237) X25>=1.359881 92    2286190.0 1521.6570 *
    119) PO_AGE_MOYEN>=0.1239552 253   13424590.0 1629.7260
      238) X131< 0.4840661 202    8557099.0 1585.5180 *
      239) X131>=0.4840661 51    2909048.0 1804.8260 *
  15) X28>=1.243883 447   60494560.0 1867.6710
   30) PO_AGE_MOYEN< 0.1835545 195   15934590.0 1659.2840
     60) X78< -0.04009216 116    5937707.0 1545.7510 *
     61) X78>=-0.04009216 79    6306180.0 1825.9910
      122) PO___MARIES< 0.9141624 66    4845964.0 1771.3770 *
      123) PO___MARIES>=0.9141624 13     263939.5 2103.2600 *
   31) PO_AGE_MOYEN>=0.1835545 252   29539580.0 2028.9220
     62) X63< 1.974209 149   10353470.0 1895.2460 *
     63) X63>=1.974209 103   12671940.0 2222.2990 *
```

**ANNEXE 4** : Grille LPP

| Code | Dénomination |
|---|---|
| **Montures** | |
| **2210546** | MONTURE < 18 ANS |
| **2223342** | MONTURE >= 18 ANS |
| **2227908** | SUPPLEMENT POUR MONTURE DE LUNETTES A COQUE, < 6 ANS |
| **Verres blancs unifocaux** | |
| **2200393** | VERRE BLANC SIMPLE FOYER, < 18 ANS, CYLINDRE <= +4,00, SPHERE DE -6,00 A +6,00 |
| **2238941** | VERRE BLANC SIMPLE FOYER, < 18 ANS, CYLINDRE > +4,00, SPHERE DE -6,00 A +6,00 |
| **2245036** | VERRE BLANC SIMPLE FOYER, < 18 ANS, CYLINDRE > +4,00, SPHERE H.Z -6,00 A +6,00 |
| **2243304** | VERRE BLANC SIMPLE FOYER, < 18 ANS, SPHERE DE +6,25 A +10,00 |
| **2261874** | VERRE BLANC SIMPLE FOYER, < 18 ANS, SPHERE DE -6,00 A +6,00 |
| **2243540** | VERRE BLANC SIMPLE FOYER, < 18 ANS, SPHERE DE -6,25 A -10,00 |
| **2273854** | VERRE BLANC SIMPLE FOYER, < 18 ANS, SPHERE H.Z DE -10,00 A +10,00 |
| **2283953** | VERRE BLANC SIMPLE FOYER, < 18 ANS,CYLINDRE <= +4,00, SPHERE H.Z -6,00 A +6,00 |
| **2288519** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, CYLINDRE >+4,00, SPHERE H.Z -6,00 A +6,00 |
| **2212976** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, CYLINDRE >+4,00,SPHERE DE -6,00 A +6,00 |
| **2280660** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, SPHERE DE +6,25 A +10,00 |
| **2203240** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, SPHERE DE -6,00 A +6,00 |
| **2282793** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, SPHERE DE -6,25 A -10,00 |
| **2235776** | VERRE BLANC SIMPLE FOYER, >= 18 ANS, SPHERE H.Z DE -10,00 A +10,00 |
| **2259966** | VERRE BLANC SIMPLE FOYER, >= 18 ANS,CYLINDRE <= +4,00, SPHERE -6,00 A +6,00 |
| **2284527** | VERRE BLANC SIMPLE FOYER, >= 18 ANS,CYLINDRE <= +4,00, SPHERE H.Z -6,00 A +6,00 |
| **Verres blancs multifocaux** | |
| **2259245** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE DE -4,00 A +4,00 |
| **2240671** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE DE -8,00 A +8,00 |
| **2238792** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE H.Z DE -4,00 A +4,00 |
| **2234239** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE H.Z DE -8,00 A +8,00 |
| **2227038** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE DE -8,00 A +8,00 |
| **2290396** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE DE -4,00 A +4,00 |
| **2245384** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE H.Z DE -4,00 A +4,00 |
| **2202239** | VERRE BLANC MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE H.Z DE -8,00 A +8,00 |
| **Verres teintés unifocaux** | |
| **2270413** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, CYLINDRE <= +4,00, SPHERE DE -6,00 A +6,00 |
| **2219381** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, CYLINDRE <= +4,00, SPHERE H.Z -6,00 A +6,00 |
| **2268385** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, CYLINDRE > +4,00, SPHERE DE -6,00 A +6,00 |
| **2206800** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, CYLINDRE > +4,00, SPHERE H.Z -6,00 A +6,00 |
| **2291088** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, SPHERE DE +6,25 A +10,00 |
| **2242457** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, SPHERE DE -6,00 A +6,00 |
| **2297441** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, SPHERE DE -6,25 A -10,00 |
| **2248320** | VERRE TEINTE SIMPLE FOYER, < 18 ANS, SPHERE H.Z DE -10,00 A +10,00 |

| | |
|---|---|
| **2226412** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, CYLINDRE <= +4,00, SPHERE -6,00 A +6,00 |
| **2254868** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, CYLINDRE <= +4,00, SPHERE HZ -6,00 A +6,00 |
| **2299523** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, CYLINDRE > +4,00, SPHERE H.Z -6,00 A +6,00 |
| **2252668** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, CYLINDRE >+4,00, SPHERE DE -6,00 A +6,00 |
| **2265330** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, SPHERE DE +6,25 A +10,00 |
| **2287916** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, SPHERE DE -6,00 A +6,00 |
| **2263459** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, SPHERE DE -6,25 A -10,00 |
| **2295896** | VERRE TEINTE SIMPLE FOYER, >= 18 ANS, SPHERE H.Z DE -10,00 A +10,00 |
| **Verres teintés multifocaux** | |
| **2264045** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE DE -4,00 A +4,00 |
| **2282221** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE DE -8,00 A +8,00 |
| **2202452** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE H.Z DE -4,00 A +4,00 |
| **2259660** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, < 18 ANS, SPHERE H.Z DE -8,00 A +8,00 |
| **2299180** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE DE -8,00 A +8,00 |
| **2291183** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE DE -4,00 A +4,00 |
| **2295198** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE H.Z DE -4,00 A +4,00 |
| **2252042** | VERRE TEINTE MULTIFOCAL OU PROGRESSIF, >= 18 ANS, SPHERE H.Z DE -8,00 A +8,00 |
| **Divers** | |
| **2204066** | VERRES ISEICONIQUES, < 18 ANS, SUR DEVIS |
| **2278219** | VERRES ISEICONIQUES, >= 18 ANS, SUR DEVIS |
| **2222408** | FILTRE CHROMATIQUE OU ULTRAVIOLET, < 18 ANS |
| **2269025** | FILTRE D'OCCLUSION PARTIELLE, < 6 ANS |
| **2256790** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE ELECTRONIQUE, LOUPE, ACHAT |
| **2289571** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE ELECTRONIQUE, LOUPE, LOC. HEBDO |
| **2238958** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE OPTIQUE, LOUPE |
| **2297926** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE OPTIQUE, VISION DE LOIN OU MIXTE |
| **2295815** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE OPTIQUE, VISION DE PRES |
| **2267397** | MATERIEL POUR AMBLYOPIE, < 20 ANS, AIDE VISUELLE OPTIQUE, VISION MICROSCOPIQUE |
| **2266676** | MATERIEL POUR AMBLYOPIE, GUIDE A ULTRASON POUR NOUVEAU-NES AVEUGLES |
| **2232855** | MATERIELS POUR AMBLYOPIE, SYSTEME DE REEDUCATION, LA BOITE DE 20, GRANDE TAILLE |
| **2202305** | MATERIELS POUR AMBLYOPIE, SYSTEME DE REEDUCATION, LA BOITE DE 20, PETITE TAILLE |
| **2287862** | PRISME INCORPORE < 18 ANS |
| **2247905** | PRISME INCORPORE >= 18 ANS |
| **2227920** | PRISME SOUPLE < 6 ANS |
| **2246716** | SUPPLEMENT POUR SPHERES > 20 DIOPTRIES, < 18 ANS |
| **2274109** | SUPPLEMENT POUR SPHERES > A 20 DIOPTRIES, >= 18 ANS |
| **2293957** | SYSTEME ANTIPTOSIS, < 18 ANS |
| **2200795** | SYSTEME ANTIPTOSIS, >= 18 ANS |
| **2251545** | LENTILLE DE CONTACT, FORFAIT ANNUEL, PAR OEIL APPAREILLE |

## TABLE DES GRAPHIQUES

## TABLE DES FIGURES

# TABLE DES TABLEAUX