Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 24/01/2022

Par : Hugo SALLEZ
Titre : Tarification d'un contrat d'assurance santé dans un contexte de nouvelle réforme et de pandémie.

Confidentialité : ☐ Non ☑ Oui (Durée : ☐ 1 an ☑ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : SeaBird Conseil
Signature :

Directeur de Mémoire en entreprise :
Nom : Arnaud CORLOUER
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

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

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

Résumé

Depuis 2020, deux évènements majeurs ont impactés le secteur de l'assurance santé. Le 100% santé, réforme réglementaire dont les assureurs avaient estimé les impacts pour adapter leurs tarifs et la pandémie de COVID qui elle, ne pouvait pas être anticipée.

Le but de ce mémoire est d'étudier l'impact de tels évènements sur un modèle de tarification et de faire une proposition sur la manière d'en tenir compte pour les années futures. Pour cela les éventuels effets qu'il y a pu avoir sur une réforme politique qui devait changer en profondeur certains domaines de l'assurance santé en France serons regardés. Les impacts d'une crise sanitaire telle que la pandémie de COVID-19 seront également analysés.

Après avoir présenté le système de santé en France, les différentes méthodes et techniques utilisées seront présentées d'un point de vue théorique. Elles seront mises en pratique afin d'obtenir un modèle qui sera ensuite appliqué. Enfin, les éventuelles différences avec les années précédentes et leur possibles raisons seront analysées.

Mots-clés : *Santé*, *COVID*, *Pandémie*, *Tarification*, *100% santé*, *Santé collective*, *Santé obligatoire*, *Machine learning*

Abstract

Since 2020, two major events have impacted health insurance sector. The "100% santé", a regulation reform, for which insurers had estimated the impacts in order to adapt their prices, and the COVID-19 pandemic that couldn't have been anticipated.

The aim of this dissertation is to study the impact of such events on a model of tarification and to make propositions on the way to consider it in the futur. For that, we'll look the potential effects that could have been seen on the tarification because of the regulation reform that should have changed deeply some fields of health insurance in France. We will also study the impacts of a sanitary crisis such as the COVID-19 pandemic.

After having presented the French health system, we'll present theorically the methods and statistical technics that we'll use. Theses technics we'll be used to have a model will be used to predict the pure premium. Finally, we'll look at the potential differences with the past years and the reasons of those differences.

Mots-clés: *Health*, *COVID*, *Pandemic*, *Tarification*, *100% santé*, *Collective Health*, *Mandatory health*, *Machine learning*

Note de Synthèse

Le secteur de l'assurance santé subit actuellement des turbulences non négligeables. Tous les acteurs se voient impactés, aussi bien les organismes de santé, les institutions ou encore les assurés. En effet, le contexte sanitaire dû à la pandémie de la COVID-19 (depuis janvier 2020) ainsi que l'adaptation de l'offre d'assurance maladie complémentaire à la réforme du 100% (depuis janvier 2019) santé vont changer en profondeur tout un secteur de l'assurance. Les organismes d'assurance maladie complémentaire devront donc dans le futur prendre en compte ces modifications afin de conserver un calcul de prime pure le plus juste possible.

Pour essayer de quantifier les changements liés à la pandémie ou à la réforme, nous allons calculer et comparer deux ratios $\frac{S}{P}$. Le premier sera calculé à partir des données issues de l'année 2019, et le second sera calculé à partir de celles de l'année 2020. Pour cela, nous disposons de données issues de différents assureurs. Afin d'obtenir des données homogènes, nous nous restreindrons aux données du portefeuille qui correspondent à des contrats collectifs et obligatoires.

Ce mémoire vise donc à soulever la question de la tarification des contrats d'assurance santé pour les années post-pandémie et avec une nouvelle réforme qui aura atteint son rythme de croisière. C'est une question importante puisque les seules années observables pour la réforme sont particulières à cause de la pandémie qui a grandement perturbé la consommation d'actes médicaux depuis janvier 2020.

Tarification

Après avoir présenté le système de santé français, nous allons donc calculer le coût moyen d'un individu pour un organisme d'assurance santé complémentaire, ceci correspond à la prime pure. La méthode retenue pour ce mémoire est l'approche coût-fréquence. Afin d'avoir des résultats qui seront facilement analysables, le calcul de la prime pure sera décomposé en calculant une prime pure par domaine médical. Une somme de chacune d'entre elles sera réalisée afin de pouvoir donner une prime pure globale. Ces domaines ont été récupérés à partir du nom des actes dispensés. Ils sont listés ici :

- Analyses
- Auxiliaires
- Chambres particulières
- Forfait hospitalisation
- Médecine générale
- Médecine spécialiste
- Monture optique
- Pharmacie
- Prothèses dentaires
- Radiologie
- Soins dentaires
- Soins hospitaliers
- Verres optiques

Pour chacun des domaines précédemment cités, la fréquence de consommation au sein du domaine ainsi que le coût moyen d'un acte dans ce domaine ont été modélisés séparément. Pour cela, trois

modélisations différentes ont été utilisées : des modèles linéaires généralisés, des forêts aléatoires et des XGBoost. Afin de choisir quel modèle retenir pour la fréquence ou la sévérité pour chacun des domaines, les données ont été séparées en deux bases, une base sur laquelle les modèles ont été entrainés et une base sur laquelle ils ont été testé. Le RMSE (root mean square error) a alors été calculé pour chacune des modélisations, et le modèle qui minimisait cette métrique a alors été retenu. Les résultats suivant ont été trouvés :

Table 1 : Récapitulatif des RMSE par domaine pour la sévérité

Domaine	GLM	Forêt aléatoire	XGBoost
Analyses	9,793	9,983	9,852
Auxiliaires	4,363	2,874	4,041
Chambre particulières	24,870	22,435	24,920
Forfait hospitalisation	4,674	3,723	4,079
Médecine générale	7,516	6,256	6,909
Médecine spécialiste	8,103	7,802	8,156
Monture optique	27,713	26,120	29,229
Pharmacie	2,910	2,870	2,888
Prothèses dentaire	193,182	199,519	189,163
Radiologie	13,882	13,853	13,691
Soins dentaire	73,179	64,145	71,351
Soins hospitaliers	34,083	34,700	35,336
Verre optique	62,178	52,858	62,559

Une fois que nous savions quel modèle utiliser pour chacun des domaines, nous étions en mesure de calculer des primes pures globales pour chaque profil d'assuré. Nous avons donc pu l'appliquer à l'ensemble de notre portefeuille afin de comparer les prédictions à la réalité. Cela a permit de juger de la qualité du modèle.

Études des effets de la pandémie et de la réforme

Comme nous l'avons évoqué précédemment, le but de ce mémoire est de soulever la question de la tarification pour les années post-pandémie. Pour cela, nous allons regarder les premiers résultats du 100% santé afin d'anticiper le coût de la réforme pour les assureurs. Cette réforme concerne les domaines optique, dentaire et audiologie. Les évolutions des remboursements en fonction des acteurs sont présentées dans les graphiques ci-dessous.

Nous pouvons donc voir que pour le domaine dentaire, les frais réels moyens ont augmenté de 2% alors que le reste à charge moyen a diminué de 17% . Pour le domaine audiologie, les frais réels n'ont pas sensiblement évolué mais le reste à charge a diminué de 14%. Enfin en optique, les frais réels ont diminué de 3.3% alors que le RAC a augmenté de 25%. Il semblerait que les premiers résultats de la réforme sont positifs pour les domaines dentaire et audiologie mais négatifs pour le domaine optique.

Après cette étude statistique, nous avons donc analysé les différences par le biais de la tarification que nous avons précédemment décrite. Pour les domaines de l'optique et du dentaire, les ratios charge

TABLE 2 : Récapitulatif des RMSE par domaine pour la fréquence

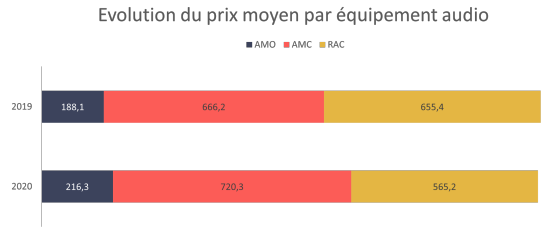
Domaine	GLM	Forêt Aléatoire	XGBoost
Analyses	9,640	9,963	9,646
Auxiliaires	14,647	15,392	14,766
Chambre particulières	2,314	2,423	2,328
Forfait hospitalisation	4,837	5,079	4,838
Médecine générale	3,146	3,245	3,133
Médecine spécialiste	3,101	3,015	2,933
Monture optique	0,368	0,380	0,366
Pharmacie	39,685	40,581	40,345
Prothèses dentaire	0,862	0,897	0,849
Radiologie	1,519	1,583	1,525
Soins dentaire	2,135	2,226	2,123
Soins hospitaliers	2,309	2,397	2,345
Verre optique	0,894	0,927	0,889

sinistre sur primes ont évolué à la baisse de 30% pour les montures optiques, de 10% pour les verres optiques et de 10% pour les prothèses dentaires. Puisque nous avons utilisé les même modèles entre les années 2019 et 2020, et que les différents adhérents changent peu entre les deux années, nous pouvons en conclure que la différence est presque complètement dû à l'évolution de la charge sinistre. Contrairement aux autres domaines, du fait de la mise en place du 100% santé nous pouvons nous attendre à une variation des montants moyens remboursés par l'AMC. C'est d'ailleurs ce que nous observons avec notamment une diminution du remboursement moyen de 22% pour les montures optiques entre 2019 et 2020.

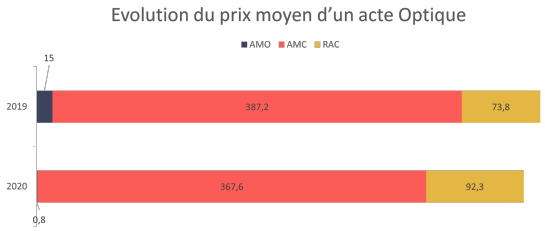
Nous avons donc également évalué les effets de la pandémie de COVID-19. D'un point de vue statistique sur l'ensemble du portefeuille de départ, on note que sur l'année 2020, le nombre total d'actes de santé a diminué de 4% par rapport à l'année 2019.Cette diminution est extrêmement marquée sur les mois de mars à mai 2020 comme nous pouvons l'observer sur le graphique ci-dessous, faisant écho au début du confinement en France et à la mise en place du chômage partiel pour certaines entreprises.

Ces premières conclusions ont été approfondies par l'étude des différences de primes pures calculées au préalable. Nous avons notamment pu observer que sur l'année entière, et avec l'étude des S/P, tous les domaines étudiés ont vu le S/P diminuer entre 2019 et 2020. Comme nous pouvons le voir dans la figure 3, cette diminution n'est pas homogène à travers les domaines. Certains ont été plus marqué que d'autres.

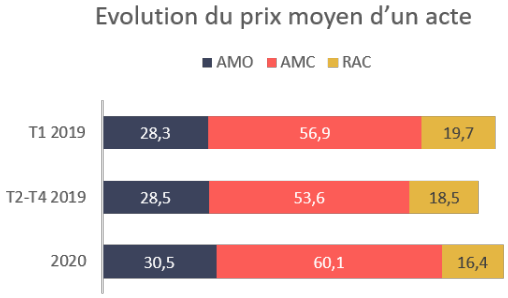
Ces diverses évolutions sont surtout dues à la modification de la consommation de santé du fait de la pandémie. Celle-ci a conduit le système de santé à s'organiser afin de pouvoir gérer la crise sanitaire : report des soins et interventions jugées « moins urgents », confinement et fermeture des commerces non essentiels dont les dentistes et opticiens ...



(a) Evolution des remboursements pour le domaine audiologie



(b) Evolution des remboursements pour le domaine optique



(c) Evolution des remboursements pour le domaine dentaire

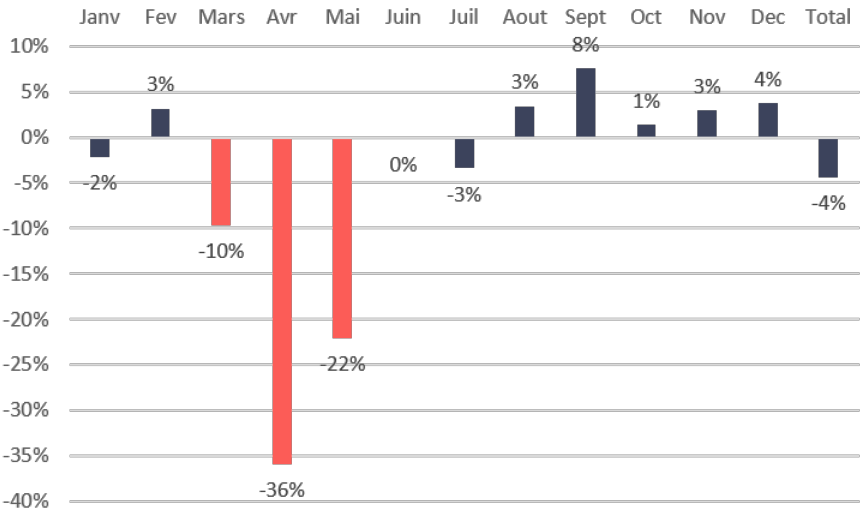


FIGURE 2 : Variation du nombre d'actes mensuels entre 2019 et 2020

Domaine	S/P 2019	S/P 2020	Evolution
Analyses	89,539	78,768	-12,029
Auxiliaires	112,015	87,130	-22,216
Chambre particulières	109,309	87,655	-19,811
Forfait hospitalisation	111,970	69,631	-37,813
Médecine générale	106,914	94,375	-11,728
Médecine spécialiste	96,470	86,631	-10,199
Monture optique	108,671	76,121	-29,953
Pharmacie	93,834	82,183	-12,416
Prothèses dentaire	107,124	94,054	-12,200
Radiologie	104,357	101,877	-2,376
Soins dentaire	101,019	87,956	-12,931
Soins hospitaliers	105,559	87,280	-17,316
Verre optique	108,670	98,279	-9,562
Total	103,903	84,578	-18,599

TABLE 3 : Évolution des S/P entre 2019 et 2020

Conclusion

Dans cette étude, nous nous sommes donc intéressés aux divers effets que peuvent avoir divers évènements, qu'ils soient prévisibles ou non sur la tarification des contrats d'assurance santé. La réforme aurait dû entrainer une hausse globale des remboursements des assureurs. Celle-ci a été anticipée par ces derniers qui ont donc augmenté leurs tarifs en conséquence. Cependant la pandémie et les diverses conséquences induites par celle-ci ont modifié de manière plus importante la consommation. La charge sinistre globale à donc grandement diminué, de près de 18% entre les années 2019 et 2020. Contrairement à une première intuition qui aurait pu être légitime, la pandémie a donc conduit à un gain technique de court terme pour les complémentaires santé.

Cependant tous les résultats présentés sont à nuancer dans la mesure où l'étude ne prend pas en compte les effets de long terme avec les pathologie plus graves (et donc plus coûteuses) qui n'ont pas pu être détectées du fait du report de certains soins. De plus, étant survenus à la même période, les effets de la pandémie et de la réforme sont difficilement différentiables et donc il est difficile de pouvoir attribuer une variation à l'un ou à l'autre.

Synthesis note

Health insurance sector is currently suffering turbulences that can't be ignore by insurers, state or insured. COVID-19 health crise and the adaptation of insurance offer in response of the political reform of the "100% santé" are going to change deeply a whole sector of insurance. Then health insurers will have to take into account those modifications in order to keep the better computation of the pure premium possible.

In order to quantify changes link to the pandemic or the reform, we are going to compute and compare two different ratios between the total sinister amount and the premium recolted by the insurer. The first ratio will be computed with the outcome data of the year 2019 and the second one will be computed with datas of the year 2020. For that we have some datas that are coming from several insurers. Because we want some homogenous datas, we'll only keep the datas that are refering mandatory and collective contracts.

This dissertation aims to raise the question of health insurance contact's tarification for the post-pandemic years with a new reform that will have its cruising speed. That is a very important question because the years that are observable for the reform are special beacause of the pandemic that has substantially disturb medical acts consumption since January 2020.

Tarification

After having presented the French health system, we will compute the mean cost of an individual for an health insurer, this is called the pure premium cost. For this study, the chosen method will be the "frequence - cost" approach. In order to have some easy analysable results, the pure premium computation will be split by computing a premium per medical field. Thoses premiums will be add to obtain the global pure premium. Those fields have been decided by looking at the name of medical acts. They are listed here :

- Tests
- Auxiliary
- Personal rooms
- Hospitalisation flat fee
- General medicine
- Specialist medicine
- Optical rim
- Pharmacy
- Dental prosthesis
- Radiology
- Dental treatment
- Hospital treatment
- Optical lens

For each of theses fields, we have modeled consumption frequence in the field and the mean cost of an act in the field separately. To do that, we used three different models : generalized linear model, random forest and XGBoost. To choose which model keep for the frequence or the cost of each field,

we have computed the RMSE (root mean square error) for every model. Then we kept the one that minimize the metric. Finally we had those tables :

Table 4: RMSE for the cost for the year 2019

Medical field	GLM	Random forest	XGBoost
Tests	9,793	9,852	9,983
Auxiliary	4,363	4,041	2,874
Personal rooms	24,870	24,920	22,435
Hospitalisation flat fee	4,674	4,079	3,723
General medicine	7,516	6,909	6,256
Specialist medicine	8,103	8,156	7,802
Optical rim	27,713	29,229	26,120
Pharmacy	2,910	2,888	2,870
Dental prosthesis	193,182	189,163	199,519
Radiology	13,882	13,691	13,853
Dental treatment	73,179	71,351	64,145
Hospital treatment	34,083	35,336	34,700
Optical lens	62,178	62,559	52,858

Table 5: RMSE for the frequence for the year 2019

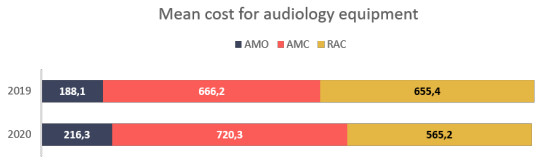
Medical field	GLM	Random forest	XGBoost
Tests	9,640	9,963	9,646
Auxiliary	14,647	15,392	14,766
Personal rooms	2,314	2,423	2,328
Hospitalisation flat fee	4,837	5,079	4,838
General medicine	3,146	3,245	3,133
Specialist medicine	3,101	3,015	2,933
Optical rim	0,368	0,380	0,366
Pharmacy	39,685	40,581	40,345
Dental prosthesis	0,862	0,897	0,849
Radiology	1,519	1,583	1,525
Dental treatment	2,135	2,226	2,123
Hospital treatment	2,309	2,397	2,345
Optical lens	0,894	0,927	0,889

Once we knew which model use for each of the fields, we were able to compute global pure premium for each insured profile. Then we were able to applied it to our whole portfolioin order to compare prediction with reality. It allowed us to judge about our model's quality.

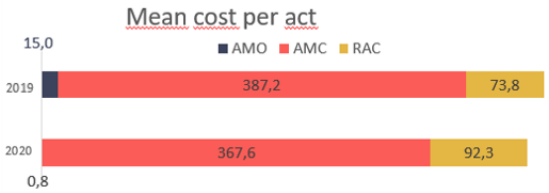
Study of the reform and pandemic effects

were able to compute global pure premium for each insured profile. Then we were able to apply the models to our entire portfolio to compare our predictions to the reality. It allowed us to judge the quality of our modelisation.

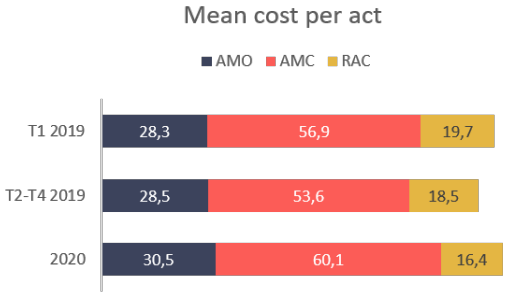
As we said before, the goal of this dissertation is to rise the question of the tarification for the coming years with the end of the pandemic. For that, we will look at the first results of the "100% santé" to anticipate its cost for insurers. This reform concern the fields of audiology, optic and dental. We obtained the following graphics for the reimbursment evolution according to the differents actors of the sector.



(a) Reimbursement evolution for the audiology field



(b) Reimbursement evolution for the optical field



(c) Reimbursement evolution for the dental field

We can see that for the dental field, the actual cost has increased by 2% while the mean required to pay for the insured has decresed by 17%. For the audiology field, the actaul cost haven't evolved but the mean required to pay has decreased by 14% percent. To finish, for the optical field the actual cost has decreased by 3.3% while the required to pay has increased by 25%. The we can say that the reform has some great first results for dental and audiology fields but not for the optical one.

After this stastical study we have analysed the differences with the tarifciation that we have previously presented. For the dental and optical field, the ratio of the accident amount over the premiums has decreased by 30% for the optical rim, and 10% for the optical lens and dental prothesis. As we used the same models for the year 2019 and the year 2020, and as the clients between the two years don't change a lot, the difference has to be mainly due to the evolution of the total accident amount. Contrary to the others medical fields, because of the begining of the "100% santé" reform, we can expect a change of the average amount reimbursed by the insurers. That is what we can see, especially

with the field of optical rim for which the decrease was about 22% of the average cost between 2019 and 2020.

We have also quantify the COVID-19 pandemic's effects. On a statistical point of view, we can see that for the year 2020 the total number of medical acts has decreased by 4% in regard of the year 2019. This reduction can is extremly strong for the march - may 2020 period as we can see on the following graphic.

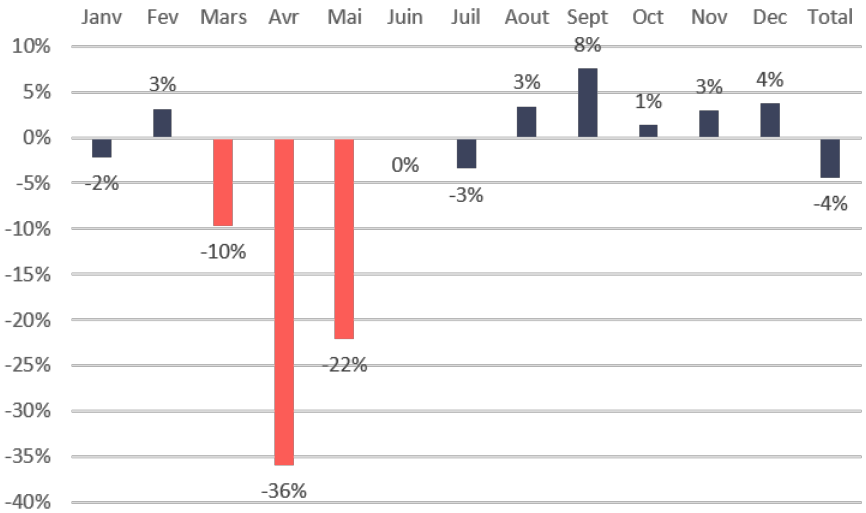


Figure 4: Medical acts evolution by month between 2019 and 2020

Those first conclusion have been deepen by the study of the differences of the pure premium that has been calculated before. We could have observed that on the entire year, and with the study of the $\frac{\text{Accident}}{\text{Premium}}$ ratios, all the fields studied have seen their ratios decreasebetween 2019 and 2020. As we can see on the table 6, this decrease isn't homogeneous among all the fields. Some have been more impacted than others

Domaine	S/P 2019	S/P 2020	Evolution
Tests	89,539	78,768	-12,029
Auxiliary	112,015	87,130	-22,216
Personal rooms	109,309	87,655	-19,811
Hospitalisation flat fee	111,970	69,631	-37,813
General medicine	106,914	94,375	-11,728
Specialist medicine	96,470	86,631	-10,199
Optical rim	108,671	76,121	-29,953
Pharmacy	93,834	82,183	-12,416
Dental prosthesis	107,124	94,054	-12,200
Radiology	104,357	101,877	-2,376
Dental treatment	101,019	87,956	-12,931
Hospital treatment	105,559	87,280	-17,316
Optical lens	108,670	98,279	-9,562
Global	103,903	84,578	-18,599

Table 6: Evolution of the ratios between 2019 and 2020

Those different evolutions are mainly due to the modification of medical act's consumption beacause of the pandemic. This one has led the health system to organize itself in order to be able to deal with the sanitary crisis : non-urgent treatment has been delayed, lockdown and closinf of the non-essential stores such as dentist and opticians...

Conclusion

In this study, we watched the different effects that can have some events, predictible or not on the tarification of health insurance contracts. The reform should have generated an global increase of reimbursement by the insurers. That increase has been anticipated by them : they increased their prices consequently. However, the pandemic and its consequences has change things in a more important way the consumption. The global amount of money reimbursed by insurers has deeply decreased by almost 18% between the years 2019 and 2020. Contrary to a first intuition that could have been reasonable, the pandemic has led to a short term technical gain for the health insurers.

Nevertheless, all the result presented has to be qualified in the way where the study doesn't look at the long term effects with more serious pathology (and by extension more expensive) that couldn't have been detected because of the treatment that has been delayed. Moreover, pandemic's effects and reform's effects are hardly separable so it's not easy to say if avariation is because of on or the other effect.

Remerciements

Je tiens tout d'abord à remercier SeaBird et l'ensemble de son personnel qui m'ont accueilli au sein de leur équipe actuariat pour mon stage de fin d'étude et qui m'a donné un contexte idéal à la bonne réalisation de ce mémoire.

Merci également à mon tuteur Arnaud CORLOUER, qui m'a accompagné tout au long de la réalisation de cette étude et qui m'a permit de profiter de ses compétences techniques ainsi que son expérience sans lesquels ce mémoire ne serait pas celui qu'il est aujourd'hui.

Merci aussi à Marina CHOU qui m'a elle aussi assisté, conseillé et aidé au quotidien pour ce mémoire.

Merci à Cassandra BELOT ma tutrice académique qui m'a apporté un regard extérieur et critique dans ce travail.

Table des matières

Introduction

Depuis 2020, le secteur de l'assurance santé a été marqué par deux évènements majeurs. Parmi eux, la réforme du 100% santé et la pandémie mondiale de la COVID-19 semblent marquer un tournant dans le monde de l'assurance santé. En effet, chaque acteur est et sera amené à faire face à de nombreux changements, aussi bien les assureurs dans leur gestion du risque, que les assurés dans leur façon de consommer. Nous avons tous entendu parler des différents confinements qui réduisaient les différentes sorties et contacts sociaux et par conséquent les consultations chez les professionnels de santé. Dans le même temps, la pandémie a touché des centaines de milliers de personnes en Francce et on peut donc s'attendre à une augmentation des dépenses de santé. Il apparait donc évident que la tarification pour les années futures sera impactée dans un sens ou dans l'autre.

Le but de ce mémoire est donc d'étudier l'impact de tels évènements sur un modèle de tarification et de faire une proposition sur la manière d'en tenir compte pour les années futures. Nous nous posons donc la question de l'évolution de la charge sinistre des assureurs depuis le début de la pandémie. Nous analyserons également les éventuels effets qu'ont pu avoir une réforme politique qui devait changer en profondeur certains actes de frais de santé en France : la réforme du 100% santé.

Dans la première partie de ce mémoire, le système de l'assurance santé en France sera présenté : acteurs, spécificités, et quelques réformes qui structurent et encadrent ce système. Les données et leurs principales caractéristiques seront également introduites dans ce chapitre avant de laisser la place à la présentation précise de ce qui sera fait au cours de la suite de l'étude. Cela permettra d'avoir le contexte pour la tarification et l'analyse qui seront menées par la suite. La seconde partie sera consacrée à l'élaboration du modèle de tarification, la présentation des différentes méthodes statistiques qui seront utilisées pour cela ainsi que leur application et les premiers résultats qui en découlent. La tarification ne sera pas fondée sur un produit existant mais consistera en une modélisation de la charge sinistre du portefeuille. Enfin, dans la troisième et dernière partie, les différents résultats obtenus seront comparés et les éventuels impacts des facteurs extérieurs au portefeuille seront analysés. Tout ceci afin de soulever la question pour les assureurs de la tarification en post-pandémie dans la mesure où le 100% santé sera en pleine application.

Chapitre 1

Contexte et cadre de l'étude

Ce premier chapitre a pour but de présenter le système de santé français, ses évolutions récentes ainsi que les enjeux de l'étude.

1.1 Le système de santé français

1.1.1 Acteurs et fonctionnement

En France, le système de remboursement des soins est organisé autour de deux acteurs majeurs, la sécurité sociale et les différents organismes complémentaires.

La Sécurité Sociale

La sécurité sociale en France regroupe un certain nombre d'institutions dont le but premier est de protéger la population française contre certains risques. Si la sécurité sociale telle que nous la connaissons a vu le jour au lendemain de la seconde guerre mondiale, l'idée de protection sociale en France date de la fin du 17ème siècle et d'un régime de retraite instauré par Jean-Baptiste Colbert.

En France, l'adhésion à la sécurité sociale est le premier étage de remboursement de l'assurance maladie. C'est un système à adhésion obligatoire. Tout français est affilié à l'un de ces régimes. En effet, celui-ci n'est pas unique et se décompose lui-même en trois grands blocs (*Les régimes* 2021) :

- Le régime général qui couvre les salariés du secteur privé et les travailleurs indépendants.
- Le régime agricole.
- Les autres régimes spéciaux qui couvrent par exemple certains salariés comme ceux de la SNCF ou encore les clercs de notaires.

Le nombre de régimes obligatoires est en réalité bien plus important que cela, notamment à cause des différents régimes spéciaux qui sont au nombre de 27. De plus, le régime général comprend également le régime local d'Alsace-Moselle, héritage de l'histoire de ces départements, qui étaient rattachés à l'empire Allemand et son régime Bismarckien jusqu'en 1918.

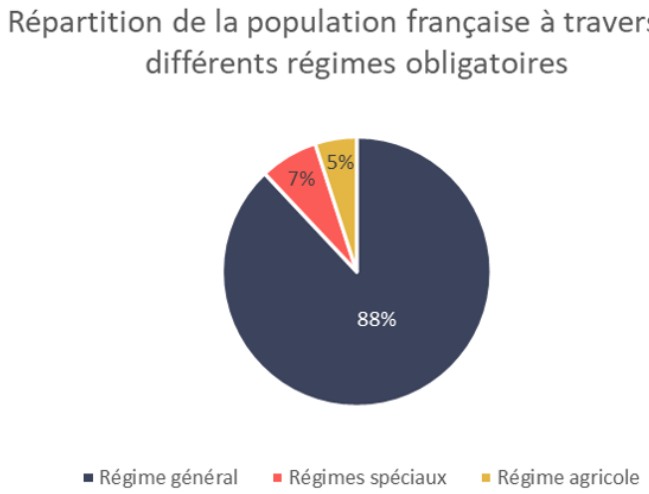


FIGURE 1.1 : Répartition des français dans les différents régimes obligatoires en 2020

La figure 1.1 montre que le régime général concerne près de 90% de la population française. Il s'agit donc du régime le plus répandu, les deux autres étant marginaux. Nous nous intéressons donc maintenant à la constitution de chacun des régimes de la sécurité sociale. Chacun des régimes de la sécurité sociale se décompose en 5 branches distinctes (*Les branches de la sécurité sociale* 2021) :

- La branche **famille**, qui aide les famille et les personnes vulnérables.
- La branche **maladie**, pour les soins, c'est la branche qui va principalement nous intéresser durant cette étude.
- La branche **accident du travail / maladies professionnelles**, qui gère les risques inhérent aux travailleurs.
- La branche **retraite**, qui gère la préparation de la retraite ainsi que ses prestations.
- La branche **recouvrement**, qui collecte les cotisation afin de financer les autres branches.

Les dépenses de la sécurité sociale sont très importantes. En effet, en 2019 elles s'élevaient à plus de 416 milliards d'euros dont 216 milliards pour la branche maladie de la sécurité sociale. La répartition des dépenses à travers les différentes branches se trouve sur la figure 1.2.

La sécurité sociale en France n'est que le premier étage d'un système de remboursement des soins. En effet, elle ne couvre pas l'intégralité des frais engendrés par le système de santé. C'est pourquoi, il existe également des systèmes de remboursement à adhésion facultative.

Les Organismes Complémentaires

Les régimes complémentaires ont pour but d'offrir une protection supplémentaire afin de réduire la part de dépenses en soins qui est à la charge de l'assuré (*Assurance maladie complémentaire* 2021). Ces organismes proposent des contenus différents et sont de natures différentes. Ceci implique une pluralité des organismes et donc des réglementations.

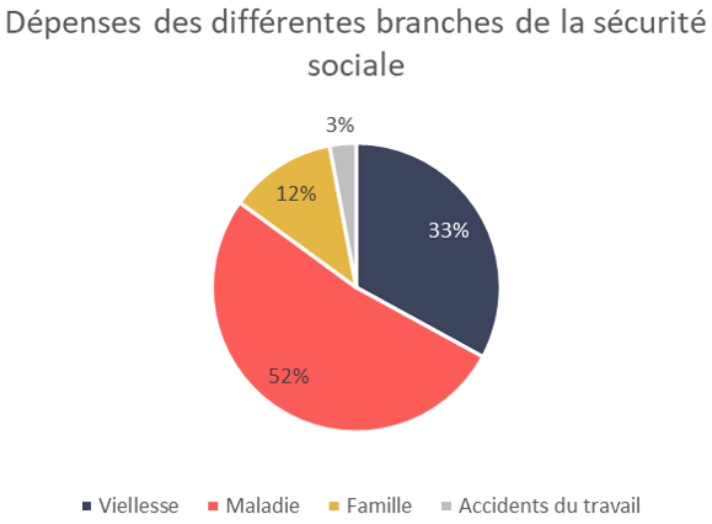


FIGURE 1.2 : Répartition des dépense de la sécurité sociale par branche en 2019

Les contrats peuvent être individuels ou collectifs. Dans le cas d'un contrat individuel, l'assuré choisi ses garanties et se voit proposer un tarif en fonction de ses caractéristiques (âge, lieu de résidence, présence d'ayant droit ...). Certains contrats peuvent être collectifs, auquel cas, le contrat est proposé par l'entreprise ou issu d'une convention collective. Dans cette configuration, le contrat sera tarifé en fonction des caractéristiques de l'entreprise comme l'âge moyen, la proportion de cadres ... L'employeur prend alors en charge une part des cotisations (en général 50%). L'adhésion à ces contrats est alors obligatoire. Cependant, il peut exister des options, qui sont elles facultatives et que l'employé peut décider de prendre. Cette présence d'option soulève alors pour l'assureur la question de l'anti-sélection. En effet, les individus qui sont plus susceptibles d'avoir recours à des soins moins bien remboursés par le contrat auront plus tendance à vouloir souscrire à ces options. C'est la notion d'anti-sélection qui entre alors en jeu.

Le secteur de l'assurance maladie complémentaire étant un secteur réglementé, seuls certains organismes sont autorisés à rentrer sur ce marché. Ils sont définis par la Loi Evin du 31 décembre 1989 comme :

- Les entreprises régies par le code des assurances.
- Les institutions de prévoyance.
- Les mutuelles relevant du code de la mutualité.

En résumé, le secteur comprend les société d'assurance (SA), les institutions de prévoyance (IP), les mutuelles. Les SA sont régies par le codes des assurances, les IP par le code de la sécurité sociale, et les mutuelles par le code de la mutualité.

Ces acteurs se partagent cependant le marché de l'assurance complémentaire de manière inégale, comme le montre la figure 1.3 (*Le marché de la santé et de la prévoyance progresse de 2,8 % en 2018* (2019)). Nous pouvons notamment voir que les sociétés d'assurance possèdent 45% des parts

de l'assurance complémentaire (collective et individuelle). De plus, 50% du marché concernent les assurances individuelles et 50% concernent le collectif tout acteurs confondus.

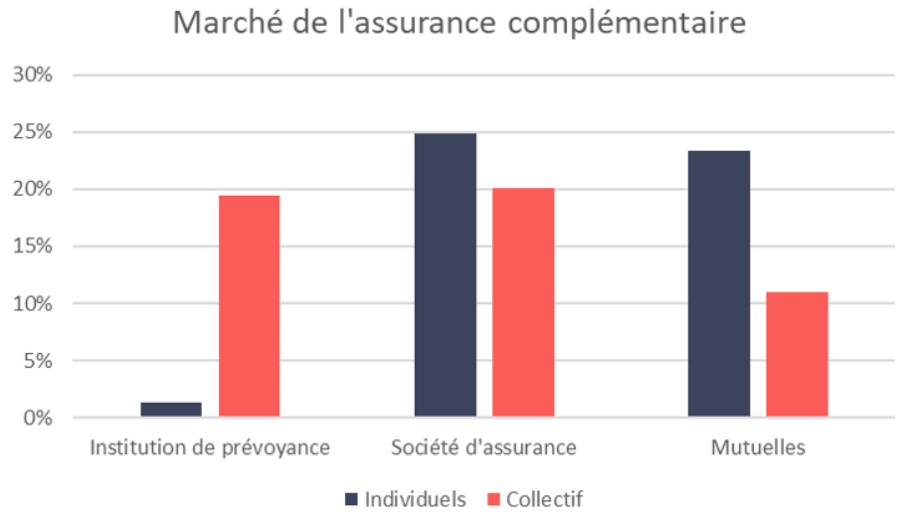


FIGURE 1.3 : Part de marché de l'assurance complémentaire en France (en terme de cotisations)

Méthode de remboursement

Cette partie décrit comment se décompose le remboursement des soins pour un particulier en France (LAZIC, 2020). Nous allons donc définir les termes « frais réels », « base de remboursement », « ticket modérateur », « reste à charge ». Ces termes sont omniprésents dans le domaine de la santé, il est donc important de les comprendre.

La première chose à définir est ce qui est appelé les frais réels (que notés parfois FR). Les frais réels représentent le coût des soins. Ces frais réels sont payés par le patient et le dispositif de tiers payant (sécurité sociale et éventuellement l'organisme complémentaire). Le tiers payant est un dispositif qui permet au patient de ne pas avoir à avancer la part qui sera prise en charge par l'assurance maladie au moment de la consultation. Cela permet de réduire la renonciation aux soins chez les ménages modestes pour lesquels un creux dans la trésorerie est un frein à la consommation.

Le coût d'une prestation de santé peut être réglementé (pharmacie, consultations, ...) ou libre (médecine douce, chambre particulière·) Lorsque le coût est réglementé, il existe une base de remboursement sécurité sociale (BRSS) qui sert au calcul du remboursement de la sécurité sociale. La BRSS est également parfois appelée tarif de convention. C'est ce qui sert à définir tous les remboursements qui vont être effectués. Les praticiens de secteurs 2 sont autorisés pour certains actes à facturer plus que le montant de la BRSS. On parle alors de dépassement d'honoraire (DH). Ils s'obtiennent simplement par la formule :

$$DH = FR - BRSS$$

Maintenant que nous avons défini les termes du côté des dépenses, nous présentons la nomenclature des remboursements. Le premier étage du remboursement est le remboursement de la sécurité sociale

(RSS). En effet celle-ci ne rembourse pas systématiquement 100% de la BRSS. Ces remboursements dépendent des actes effectués ainsi que du régime obligatoire auquel l'assuré est affilié. Quelques exemples de remboursements sont donnés dans la table 1.1.

TABLE 1.1 : Exemple de remboursements en pourcentage de la BRSS

Acte médical	**Régime général**	**Régime Alsace Moselle**
Consultation médecin généraliste	70%	90%
Séance de kinésithérapie prescrite par un médecin	60%	90%
Médicaments à service médical rendu modéré	30%	80%
Médicaments reconnus comme irremplaçables et particulièrement coûteux	100%	100%
Orthopédie	60%	90%
Frais d'hospitalisation	80%	100%

Le ticket modérateur (TM) est simplement la différence entre BRSS et RSS. Le remboursement complémentaire (RC) dépend du contrat de chaque patient et vient s'ajouter au remboursement de la sécurité sociale. On a alors le reste à charge (RAC) par la formule :

$$RAC = FR - (RSS + RC)$$

Le système de remboursement est alors résumé dans la figure 1.4. Cependant tous les actes ne sont pas forcément remboursés par la sécurité sociale, par exemple l'ostéopathie ne sera pas prise en charge pas la SS.

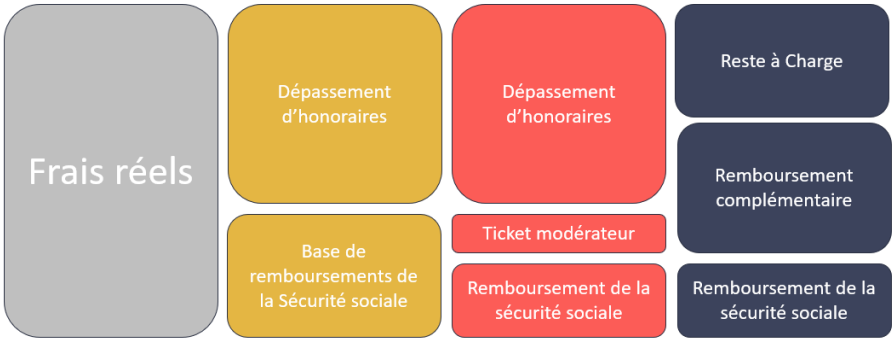


FIGURE 1.4 : Décomposition d'un remboursement

Exemple : Pour illustrer cette décomposition, l'exemple d'une garantie 200% BR y compris sécurité sociale pour l'achat et la pose d'une couronne dentaire qui est revenue à l'assuré à 350€ est considéré. Les frais réels sont ici de 350€. La BRSS pour la pose d'une couronne est de 107,50€ remboursés à 70% par la sécurité sociale. Aalors :

$$RSS = 70\% \times BRSS = \frac{70 \times 107,5}{100} = 75,25.$$

On en déduit alors le ticket modérateur :

$$TM = BRSS - RSS = 107,5 - 75,25 = 32,25.$$

L'assuré a souscrit à une garantie 200% BR y compris SS. Sa garantie sera donc de :

$$200\% \times BRSS = 2 \times 107,50 = 215.$$

Cette garantie est exprimé "y compris SS" ce qui veut dire que pour obtenir le remboursement de l'assurance complémentaire le montant remboursé par la SS y est soustrait.

$$RC = 215 - 75,25 = 139,75.$$

Maintenant, il est possible de calculer le reste à charge pour l'assuré :

$$RAC = FR - RSS - RC = 350 - 75,25 - 139,75 = 135.$$

L'exemple est résumé dans la table suivante :

BRSS	107,50€
RSS	70%
Frais réels	350 €
Remboursement sécurité sociale	75,25€
Remboursement assureur	139,25€
Reste à charge	135€

1.1.2 Le cadre réglementaire

Dans cette partie, nous présentons les principaux points de réglementation de l'assurance santé en France.

Contrats responsables et solidaires

Ces contrats sont définis par le décret n°2014-1374 du 18 novembre 2014. La notion de contrat solidaire signifie qu'il n'y a pas de sélection médicale à l'entrée du contrat. La notion de contrat responsable signifie que le contrat répond à un cahier des charges dont certains points sont décrits ci-dessous.

- prendre en charge l'intégralité du ticket modérateur pour les soins de ville.
- prendre en charge l'intégralité du forfait journalier hospitalier et ce, sans limite de durée.
- pour les médecins n'ayant pas adhéré à l'OPTAM, prendre en charge un dépassement d'honoraire d'un maximum de 100% de la BRSS. De plus la différence de remboursement OPTAM / non-OPTAM doit être d'au moins de 20%.

Le contrat responsable instaure également divers maxima et minima de remboursements dans plusieurs domaines comme l'optique par exemple.

Être un contrat responsable n'est pas quelque chose d'obligatoire, cependant dans les faits, presque tous les contrats le sont. En effet, les contrats responsables se voient attribuer des aides sociales et fiscales. Par exemple la taxe sur les produits d'assurance (TSA) se voit diminuer de 7% pour les contrats responsables, passant de 21,07% à 14,07%.

La « Gender directive »

La « gender directive » est une réglementation européenne qui interdit aux assureurs d'utiliser le sexe de l'assuré comme un facteur discriminant pour les primes ou les prestations d'un contrat. Autrement dit, toutes choses étant égales par ailleurs, un homme et une femme doivent payer la même prime pour les mêmes garanties. Cette directive européenne prend sa source le 13 décembre 2004 avec une première directive européenne qui instaure un principe d'égalité dans l'accès et la fourniture de biens et services entre les hommes et les femmes.

Cette égalité peut être cependant outrepassée pour les contrats d'assurance quand des données actuarielles et/ou statistiques le justifient. Les états peuvent donc autoriser cette discrimination selon le sexe s'ils le souhaitent. La France a prit le parti de laisser les assureurs libres d'intégrer ce facteur ou non dans leur tarification et notamment dans le domaine de l'automobile où les différences sont criantes.

Suite à une plainte d'une association de consommateurs Belge, le 1^{er} mars 2011, la cours de justice de l'Union Européenne publie un arrêt qui abroge l'existence de cette dérogation à partir du 21 décembre 2012. L'application de la directive de 2004 est maintenant stricte et sans dérogation, les tarifs doivent être similaires en fonction du sexe. En assurance automobile, les primes des jeunes conductrices ont ainsi pu augmenter jusqu'à 20%.

Cependant, cette interdiction ne concerne que la prime commerciale. Il est donc possible d'obtenir une prime pure en fonction du sexe et d'appliquer la proportion d'hommes et de femmes dans le portefeuille pour la prime commerciale (un exemple est disponible ci-dessous). De plus, cette interdiction ne concerne que la tarification et pas les provisions.

Exemple : La prime pure toute chose égale par ailleurs pour un homme est de 51 € et de 36€ pour une femme. Notre portefeuille est composé de 56% d'hommes et de 44% de femmes. La prime commerciale sera donc de : $51 \times 0,56 + 36 \times 0,44 = 44,4$€.

100% santé

La réforme du 100% santé est présentée dans cette section (*Réforme 100 % santé* 2021). Lors de la campagne présidentielle de 2017, le candidat Emmanuel Macron s'est engagé à faciliter l'accès aux soins et ce notamment dans les domaines de l'optique, du dentaire et de l'audition. En effet, des études montraient qu'une partie non négligeable de français renonçait à se soigner dans ces domaines par manque de moyens. De plus il s'agit de trois domaines où le reste à charge pour l'assuré était particulièrement élevé. Le RAC moyen sur l'ensemble des actes de santé est de 8% alors qu'il est de 23% en dentaire, 24% en optique et 56% en audiologie. Ceci est résumé dans la figure 1.5. L'objectif est donc de proposer un ensemble de soins dans ces trois domaines où le reste à charge pour l'assuré serait nul. Les remboursements de l'assurance maladie ainsi que la complémentaire santé devront alors couvrir sur ces actes l'entièreté des frais réels.

Cette obligation de reste à charge zéro sur certains soins ne s'applique cependant pas à tous les contrats d'assurance complémentaire. Elle s'appliquera uniquement aux contrats responsables.

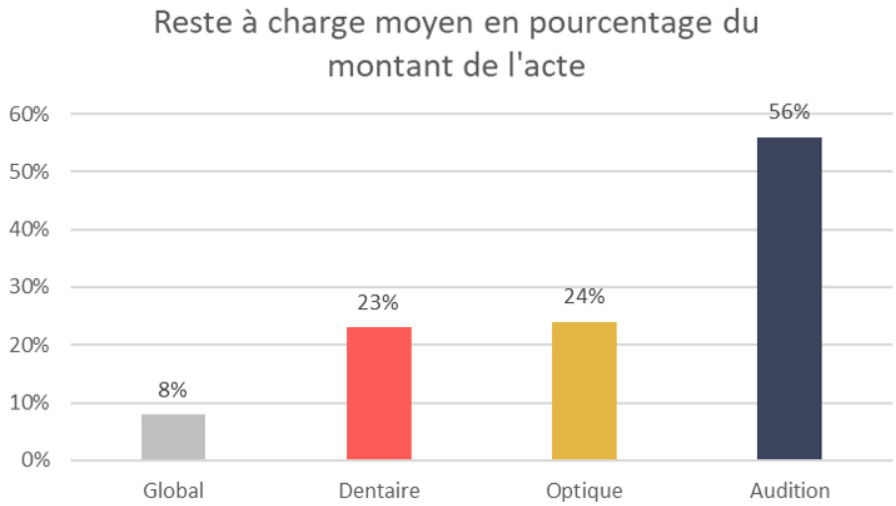


FIGURE 1.5 : Comparaison des restes à charge moyens dans les domaines du 100% avant l'application de la réforme

La mise en place de la réforme devait se faire à partir de 2019 selon un calendrier pré-établi dont la représentation est donnée en figure 1.6. Nous pouvons remarquer que la réforme avait déjà commencé a être mise en place en 2019 alors que cette année est considérée dans cette étude comme « normale ». Ce n'est pas un problème car les remboursements n'étaient pas affectés.

L'application sera effectuée pas le biais de trois leviers :

- **La mise en place de prix limite de vente** : les professionnels de santé se verront imposer un prix maximal à appliquer sur certains actes.
- **Une hausse de la base de remboursement de la sécurité sociale** qui augmentera naturellement la part prise en charge par l'assurance maladie.
- **Une hausse du remboursement de l'assurance complémentaire** : celle-ci devant ajuster les garanties de ses contrats pour que le reste à charge soit de zéro pour les soins concernés.

La réforme a donc été mise en place progressivement depuis 2019. Cette première année, les domaines de l'audition et du dentaire ont été marqués par l'instauration des prix limites de vente ainsi qu'une augmentation de la BRSS. En 2020, le domaine de l'optique a été marqué par la mise en place du reste à charge 0 et des prix limites de ventes pour les équipements concernés (les opticiens doivent présenter 17 montures pour adultes et 10 montures pour enfant de le cadre du dispositif 100% santé et des verres qui peuvent corriger toutes les déficiences visuelles). Cette même année pour le domaine dentaire les premiers actes à reste à charge 0 pour l'assuré (certaines prothèses dentaires) sont concernés par la réforme et ceux-ci seront complétés par les derniers actes en 2021 (les prothèses amovibles notamment). Pour les dispositifs auditifs c'est à partir de 2021 que seront remboursés à 100% les dispositifs de classe 1. En effet, en audiologie les dispositifs sont classés en deux catégories (classe 1 et classe 2) en fonction de leur caractéristiques techniques. Cette mise en place progressive est résumée à la figure 1.6.

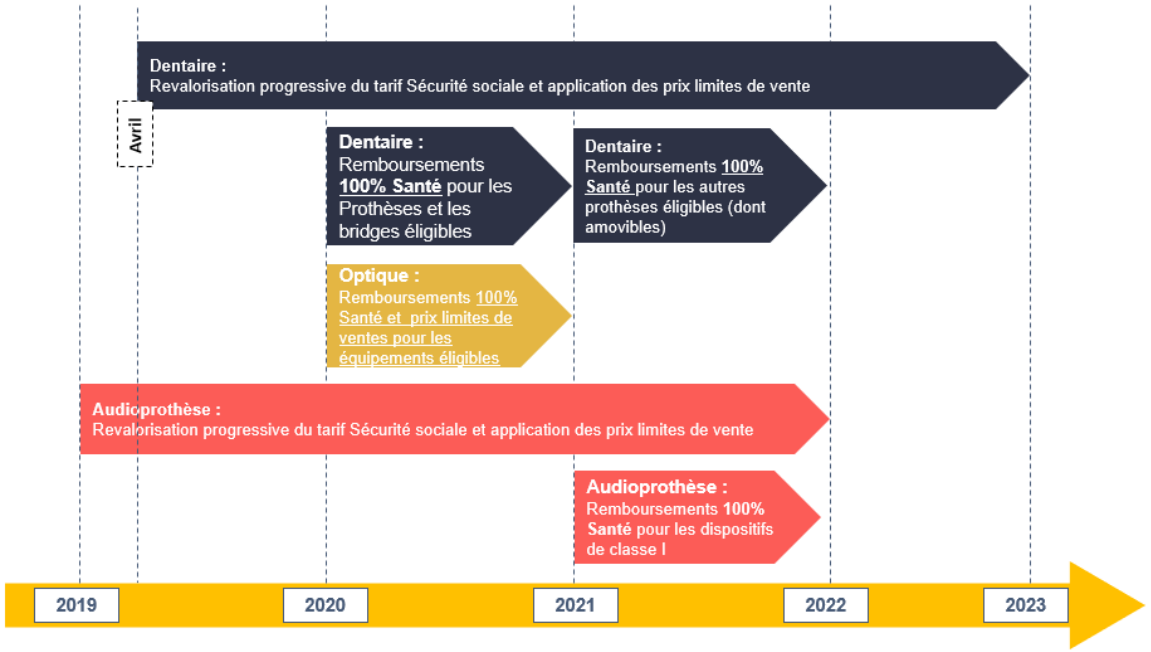


FIGURE 1.6 : Calendrier prévisionnel de la mise en place du 100% santé

Nous avons donc présenté les grandes spécificités du marché de l'assurance santé français, ainsi que certaines grandes parties du cadre réglementaire qui l'entourent. Il s'agit donc d'un système avec deux étages de remboursement des soins qui sont eux mêmes encadrés par une législation spécifique. Dans la partie suivante, les données ainsi que le portefeuille utilisés pour l'étude sont présentés.

Loi Evin

La loi Evin mutuelle prévoit un maintien des garanties santé des contrats collectifs pour les anciens salariés sous certaines conditions. Cette loi est en fait l'article 4 de la loi n° 89-1009 du 31 décembre 1989. Ce maintien de garantie s'applique aux anciens salariés qui ont du quitter l'entreprise pour une raison indépendante de leur volonté (invalidité, licenciement, retraite...) ou aux ayants droits d'un salarié décédé.

Cependant la cotisation de l'assuré n'est plus financée par l'employeur et celle-ci peut de plus augmenter. En effet, on a vu que l'employeur prenait en charge une partie des cotisations de ses salariés, cependant ce n'est pas le cas pour ses anciens employés. De plus, ne pouvant plus compter sur la mutualisation au sein de l'entreprise, l'assureur peut décider d'augmenter la cotisation des anciens employés. Cette augmentation est cependant encadrée dans la mesure où elle ne peut être supérieure à 50%. Un décret du 23 mars 2017 vient encadrer cette augmentation :

- pas d'augmentation la première année
- la seconde année, le tarif ne peut pas être supérieur de plus de 25% du tarif des salariés actifs
- la troisième année, le tarif ne peut pas être supérieur de plus de 50% du tarif des salariés actifs

De plus, les ayants droits ne peuvent plus profiter de cette couverture (sauf en cas de décès du salarié) et donc cela peut être un frein. Enfin, les garanties présentes à la fin de l'emploi de l'assuré sont fixes. Les besoins en matière de santé changent avec l'âge, les besoins de lunettes augmentent par exemple, et les garanties peuvent ne pas être adaptées au réel besoin de l'assuré.

Maintenant que le secteur de l'assurance santé français a été présenté, nous possédons donc le contexte de l'étude et présentons dans la partie suivante les données sur lesquelles elle sera réalisée.

1.2 Présentation des données

Les données utilisées pour l'étude sont décrites dans cette partie. C'est une partie très importante pour l'étude car c'est elle qui peut garantir une bonne qualité des résultats.

1.2.1 Construction et présentation de la base données

Périmètre de l'étude

Dans ce début de section, le périmètre de l'étude est présenté. Le portefeuille dont on dispose provient d'un gestionnaire pour compte de tiers qui a accepté de nous prêter les données. Par sa provenance, les données sont hétérogènes dans leur composition, en effet elles regroupent les sinistres de 5300 contrats santé collectifs obligatoires différents.

Les données étaient initialement regroupées dans plusieurs bases qui contenaient les sinistres de chaque trimestre (pour les années 2015 - 2020 et le premier trimestre de l'année 2021). Elles ont ensuite été agrégées pour ne disposer finalement que d'une base regroupant tous les actes médicaux : la base `Sinistres` qui liste les différents actes médicaux recensés. Nous nous restreindront cependant aux années 2019 et 2020 qui sont les seules années pour lesquelles où l'entièreté des données est en notre possession, les autres années étant limitées à une certaine période. Pour les actes médicaux de l'année 2019, on s'est restreint aux sinistres qui ont été réglés avant le premier avril 2020. En effet, les dernières données à disposition sont celles du premier trimestre de 2021. Il faut donc restreindre les données de l'année 2019 pour que les deux années soient comparables. Ceci implique forcément une perte de données mais celle-ci est nécessaire pour éviter un biais qui pourrait modifier le tarif calculé pour l'année 2019. Cette perte de donnée représente 56 735 lignes de la base sinistres, ce qui représente 0,85% des données à disposition pour l'année 2019.

Tables Effectifs et Sinistres

En plus de la base de données `Sinistres` déjà introduite, une base de données qui liste tous les assurés, les bénéficiaires, leur période d'exposition ainsi que leurs caractéristiques principales est à disposition.

La base de donnée `Effectifs` comportait initialement près de 800 000 lignes. Elle a alors été retraité afin d'enlever les lignes qui ne font pas sens telles que les erreurs de saisies ou les individus qui avaient une date de fin de couverture antérieure à celle de début (celles-ci représentant une partie négligeable

de la base totale). Ensuite, certains bénéficiaires apparaissaient sur plusieurs lignes. Nous avons alors réalisé un regroupement en fonction de leur identifiant. Enfin, on s'est restreint aux données qui provenaient des contrats collectifs et obligatoires, afin de posséder des données les plus complètes et homogènes possibles (par exemple on ne possède pas le code NAF pour des données individuelles). L'exposition mensuelle de chacun des individus est alors calculée pour les mois de 2019 et 2020. On obtient ensuite pour chaque individu et chaque mois de la période un nombre entre 0 et 1. Une fois ces retraitements effectués, la base comporte 248 492 bénéficiaires distincts ainsi que leurs expositions sur 2019 et 2020.

Maintenant les principales variables présentes dans la base de données `Effectifs` sont présentées. Comme évoqué précédemment, ces données proviennent d'un gestionnaire pour compte de tiers. En effet, certains assureurs qui ont un portefeuille de contrat conséquent délèguent, moyennant une commission, la gestion de ce portefeuille à une autre entité et donc les sinistres qui y sont rattachés. C'est ce qui explique la présence dans la base de données de diverses variables d'identification, une pour le porteur de risque et une autre pour le délégataire de gestion.

Chacun des contrats possède donc :

- Les dates de début et fin de couverture du contrat.
- Le nom du souscripteur qui peut être un particulier ou une entreprise, le portefeuille contenant des données collectives et des données individuelles.
- Des informations sur l'assuré, telles que son sexe, son année de naissance, un numéro d'identification, elles sont nommées ASSURE_SEXE, ASSURE_DATENAISSANCE, ID_ASSURE.
- Les mêmes informations concernant le bénéficiaire du contrats, en effet certains assurés ont également des ayants droits. On possède également dans ce cas le type du bénéficiaire : Assuré, Conjoint, Enfant .
- Le code postal rattaché au contrat.

La partie suivante présentera la base `Sinistres`. Celle-ci comporte des actes médicaux des années 2015 à 2021. Cependant les années 2015 à 2018 et 2021 sont incomplètes et ne seront donc pas utilisées pour l'étude. De plus, la base contient les montants de chaque acte ainsi que le bénéficiaire qui a profité de l'acte. Elle comportait initialement plus de 29 millions de lignes avec 62 variables. Cette base de données est composée comme celle des effectifs de diverses variables qui permettent l'identification du contrat : numéro de contrat chez le délégataire de gestion, chez le porteur du risque, le nom et la description de la police associée ainsi que des identifiants concernant l'assuré et le bénéficiaire. Pour la suite,nous ferons la distinction entre assurance maladie obligatoire (AMO) et assurance maladie complémentaire (AMC). Pour conserver uniquement les sinistres liés aux contrats collectifs et obligatoires, seules les lignes correspondant à des bénéficiaires présents dans la base `Effectifs` retravaillée ont été conservées.

Chaque ligne de la base correspond à un acte précis. Par exemple un assuré qui va en pharmacie pour plusieurs médicaments mais le même jour apparaitra sur plusieurs lignes puisque chacun de ces médicaments ne sera pas forcément remboursé de la même manière.

L'autre partie des variables présentes concerne les actes médicaux en eux même :

- LETTRECLE et COMPLETCLE qui sont des diminutifs du type d'acte effectué.
- une variable DOMAINE qui définit le domaine de l'acte médical, il peut être optique, hospitalisation, médecine générale etc.
- Le libellé qui permet de savoir clairement de quel acte il s'agit nommé LIBELLE.
- Une variable nommée MONTANT_DEPENSE qui correspond aux frais réels présenté précédemment.
- Le SIREN de l'entreprise (quand le contrat est souscrit à titre collectif)
- La BRSS associé à l'acte pour laquelle la variable se nomme BASE_AMO. Ainsi que le taux de remboursement associé qui sera nommé TAUX_AMO. Il a déja été vu qu'en multipliant ces deux quantités on trouvait le RSS qui est également présent dans la base et qui se nomme REMB_AMO.
- Le type de remboursement, en fonction de si le remboursement est de base ou issu d'une surcomplémentaire.
- Le remboursement de la complémentaire santé, nommé REMBOURSEMENT_AMC.
- Les dates des actes ainsi que les dates de remboursements. On pourra donc calculer des facteurs de développement si on en rencontre le besoin par la suite.

Les données seront regroupées domaine par domaine afin de réaliser une tarification par domaine et ainsi pouvoir capter les effets de la COVID-19 sur chacun d'entre eux.

Variables ajoutées et jointure

A partir de ces informations, d'autres variables ont pu être créées pour permettre une meilleure manipulation de ces données. Pour commencer, le code postal donnera le numéro de département et il sera ainsi possible de regrouper les contrats en fonction de leur région ce qui nous donnera deux nouvelles variables. De plus, l'identifiant du bénéficiaire sera enrichit en y ajoutant le type du bénéficiaire (certains bénéficiaires pouvant avoir un contrat d'ayant droit et un contrat d'assuré, cette variable sera utile pour les différentes jointures qui seront réalisées par la suite). Enfin, à partir de la date de naissance, les variables d'âges et de groupe d'âge ont pu être créées. Les âges calculés sont les ages pendant l'année, et non les ages à la date de consommation. Enfin les catégories d'âges ont été choisie de manière classique : une catégorie moins de 18 ans, une catégorie 18 - 25 puis une catégorie par tranche de 10 ans jusqu'à la dernière 75 ans et plus. Il est possible de retrouver la part de chaque groupe d'âge dans le portefeuille dans la figure 1.7. Les plus de 65 ans ne représentent que 1% du portefeuille, ce qui est normal puisqu'on regarde uniquement des données collectives.

Une variable exposition a été également créée à partir des dates de couverture. Cette variable compte le nombre de jour dans chaque mois pour lesquels l'assuré était couvert par le contrat. Ce nombre est alors divisé par le nombre de jour total dans le mois, ceci donne alors un nombre entre 0 et 1. Ce nombre nous servira a ajuster le nombre d'acte consommé par individu pour les différents modèles qui seront réalisés. Seuls les assurés étant couverts sur au minimum 30 jours dans l'année seront conservés afin de ne pas introduire un biais. Le calcul des expositions a été réalisé afin d'obtenir des poids pour l'apprentissage des modèles.

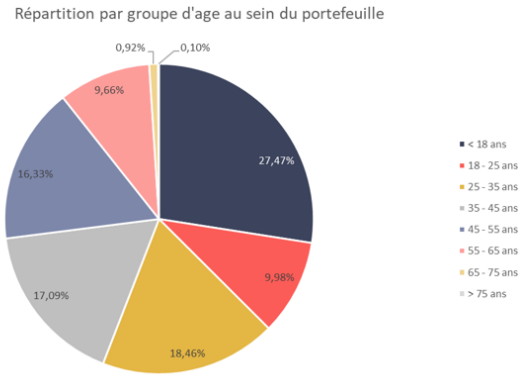


FIGURE 1.7 : Répartition par groupe d'age des bénéficiaires dans les effectifs

Initialement, il manquait une variable afin d'avoir le code NAF de l'entreprise. Les codes NAF et les SIRET associés pour les différents clients ont alors été récupérés. Les codes NAF ont été ajoutés par le biais d'une jointure sur le SIREN. Une fois que nous avions le NAF pour chacun des individus pour lesquels c'était possible, seule la lettre de celui-ci a été conservée afin d'avoir une quantité d'information adéquate. Le code NAF nous permettra de connaitre le secteur d'activité des assurés.

Une autre variable d'importance qui manquait était le niveau de garantie du contrat choisi par l'assuré. Cependant, nous possédons des bénéficiaires de plusieurs assureurs (92 assureurs différents) et plusieurs contrats qui ne sont donc pas similaires. Nous voulons donc synthétiser l'information contenue par les 5300 contrats différents que l'on possède afin de classer ceux-ci en trois catégories :

- Garantie Basse
- Garantie Médium
- Garantie Haute

Une tarification domaine par domaine va être réalisée, nous devrons donc définir le niveau de garantie pour chacun des domaines. En effet, certains contrats peuvent bien rembourser les soins liés au domaine de l'optique mais ne pas bien rembourser ceux du dentaire. Pour cette catégorisation, une variable GARANTIEYCSS qui donne le niveau de garantie en pourcentage du tarif de convention en prenant compte le remboursement de la sécurité sociale a été créée. Pour la création de cette variable, seules les lignes qui comportaient un RAC strictement positif ont été conservées afin d'être sûr que la garantie de l'AMC était complètement utilisée. On a ensuite calculé la garantie par la formule :

$$GarantieYCSS = 100 \times \frac{Remboursement\_amc + Remb\_amo}{base\_amo}$$

Pour déterminer quelles garanties sont basses, moyennes ou hautes, il faut avant tout déterminer des seuils de garantie. Par exemple, pour un acte donné si la garantie y compris sécurité sociale est en dessous de 150% elle sera dite "basse", entre 150% et 200% ce sera une garantie "moyenne" et au delà de 200% ce sera une garantie "haute". Pour déterminer ces seuils pour chacun des actes, on a utilisé une méthode de classification ascendante hiérarchique (CAH). La CAH est un algorithme de classification

qui fonctionne de la manière suivante : les individus sont considérés un par un puis on réunis les deux qui sont les plus proches au sens d'une règle de calcul (que l'on appellera indice d'agrégation). L'opération est alors répétée en considérant les individus groupés comme un seul individu. L'algorithme continue d'agréger jusqu'à avoir un seul groupe. L'indice d'agrégation utilisé lors de notre CAH était l'indice de Ward qui mesure la perte d'inertie interclasse en réunissant les deux groupes. Ce n'est cependant pas le seul indice possible.

La classification peut alors être représentée sous forme de dendrogramme (un exemple très simple de dendrogramme est donné en figure 1.8, il ne provient pas de nos données). Plusieurs techniques existent pour choisir le nombre de classes idéales, cependant on a utilisé une fonction sur `R` (fonction « best.cutree » du package `JLutils`) pour cela. La coupure idéale donnait alors trois classes.

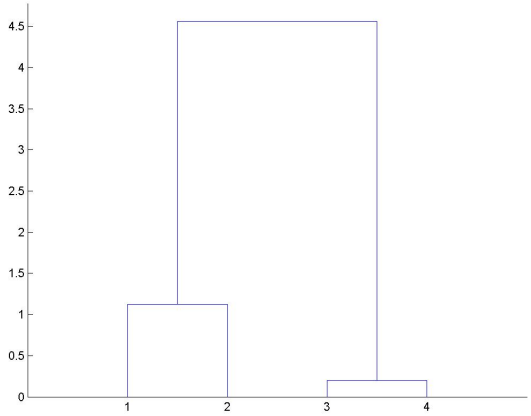


FIGURE 1.8 : Exemple de dendrogramme

Comme nous cherchons a donner un niveau de garantie pour chacune des polices, nous avons sélectionné le ou les actes les plus représentatifs de certains domaines eux mêmes représentatifs du niveau de remboursement du contrat. Ils ont été sélectionnés soit parce que c'étaient les actes qui apparaissaient le plus parmi les actes avec un RAC strictement positif ou ils permettaient de bien discriminer les différentes polices. Avec la CAH nous avons alors eu pour chacun des actes étudiés des valeurs seuils selon lesquelles on pouvait classer les polices dans 3 catégories distinctes. Dans la figure 1.2 sont indiqués les actes étudiés ainsi que les valeurs seuils associées.

Il est notamment visible que tous les domaines ne sont pas représentés dans ce tableau. Nous nous sommes contenté de catégoriser des domaines représentatifs de la qualité du contrat. Le niveau de garantie a été déterminé pour les actes dont le niveau de remboursement peut être différent d'un contrat à un autre. Par exemple pour la pharmacie dont le remboursement est de 100% au titre des contrats responsables il n'y a pas d'intérêt à définir un niveau de garantie alors que pour une prothèse dentaire il existe des différences significatives (différences réduites avec l'entrée en vigueur du 100% santé). Ce travail est réalisé sur des regroupements de contrats similaires. Une fois que chaque regroupement de contrat possédait un niveau de garantie pour les actes sélectionnés, une jointure a pu être réalisée pour donner un niveau de garantie sur les actes où il n'y avait pas de RAC. Certains

		Remboursement Y compris SS (en % BRSS)		
Domaine	**Acte**	**Bas**	**Medium**	**Haut**
Dentaire	**Prothèses fixe céramique**	< 344	344 – 461	> 461
Médecine Générale	**Consultation spécialiste**	< 116	116 – 157	> 157
Optique	**Verre Optique (en euro)**	< 93	93 – 190	> 190
Optique	**Monture Optique (en euro)**	< 63	63 – 114	> 114

TABLE 1.2 : Actes retenus et seuils associés

contrats n'ayant jamais de RAC dans un domaine, il n'était pas possible de les classer dans une des trois catégories initialement prévues, on a alors créé une dernière catégorie "Non Renseignée".

La CSP de chaque assuré a également été reconstituée en fonction du nom donné pour la description du contrat. Chaque police a été classée dans une des trois catégories suivante :

- Cadre
- Non Cadre
- Ensemble du personnel

Comme évoqué précédemment, les données collectives et individuelles ont été séparées. Cette distinction s'est effectuée sur la base de la règle de décision suivante pour ne garder que les polices collectives :

- Le prénom du souscripteur n'apparait pas dans les données.
- La description de la police ne comporte pas la mention ≪ TNS ≫ (travailleurs non salariés).
- La description de la police ne comporte pas la mention ≪ GM ≫ (gérant majoritaire).
- La description de la police ne comporte pas la mention ≪ individuel ≫.
- La description de la police ne comporte pas la mention ≪ Madelin ≫.
- La description de la police ne comporte pas la mention ≪ retraite ≫.
- La description de la police ne comporte pas la mention ≪ Evin ≫.

Une dernière variable a été créée indiquant la présence ou non d'une option pour chacun des bénéficiaires. Enfin, les données ont été séparées en deux bases différentes selon la période de couverture : une pour l'année 2019 et une autre pour l'année 2020.

Une fois toutes les variables créées la jointure entre la base de données `Effectifs` et `Sinistres` a pu être réalisée. Celle-ci a été faite sur les variables ID_BENEF et TYPE_BENEFICIAIRE. On possède enfin une base de données sur laquelle on va pouvoir travailler et créer des modèles.

Contrôles et traitements des données

L'idée de cette partie est de s'assurer que les données qui sont en notre possession sont cohérentes et sans erreurs. Pour cela, les points suivants vont être vérifiés :

- L'absence de doublons.
- La cohérence des dates de naissances.
- La cohérence des dates de couvertures.
- La cohérence de tous les montants, certains pouvaient être négatifs avant le retraitement.
- Le respect de l'anonymat des données.

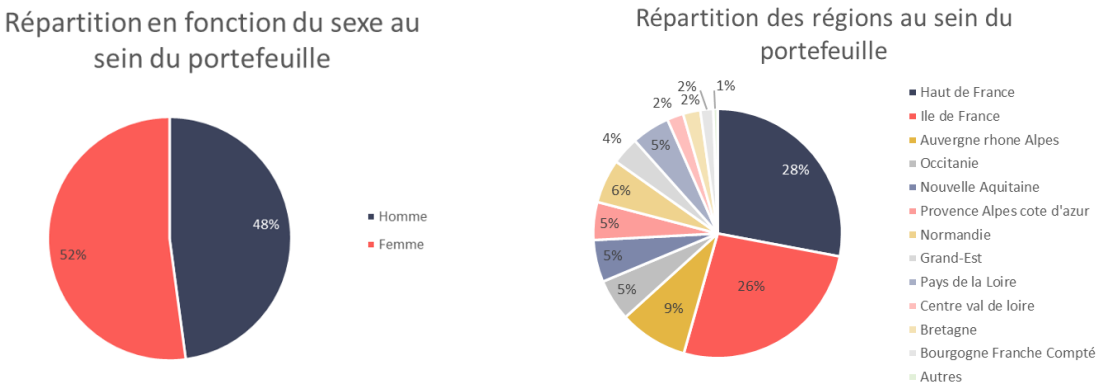
L'absence de doublons est un point facilement vérifiable grâce à la fonction `distinct` du package `dplyr` de R. En ce qui concerne les dates de naissance, il faut s'assurer qu'elles sont antérieures à 2020, et postérieures à 1904 (année de naissance de la doyenne française connue). Pour la cohérence des dates de couverture, nous avons simplement regardé la différence en jours des dates de fin de couverture et celles de début. On a alors écarté les lignes correspondants aux valeurs négatives. En ce qui concerne les montants, les lignes présentant des erreurs étant assez rares elles ont été corrigées au cas par cas. Finalement, pour l'anonymat des données, nous nous sommes assuré qu'aucun nom n'apparaissait dans la base de données.

1.2.2 Description des données

Caractéristiques du portefeuille

Le portefeuille total est composé de près de 248 492 bénéficiaires de caractéristiques différentes. La répartition des différents âges et régions est donnée dans la figure 1.9. Nous avons cependant un certain nombre de bénéficiaires qui ne sont pas assurés sur toute l'année. En appliquant un coefficient à ces personnes, (une personne qui est assurée 3 mois comptera pour 25%), le nombre de bénéficiaires devient alors environ 180 000. La proportion d'hommes et de femmes est similaire.

En ce qui concerne les régions, une forte concentration des bénéficiaires dans la région Hauts-de-France (28%) est remarquable, celle-ci est due à la présence géographique historique du gestionnaire pour compte de tiers qui nous a fourni les données. De plus l'Île de France est également beaucoup représentée avec 26% des bénéficiaires. Le graphique de la répartition complète est donné en figure 1.9.



(a) Répartition Homme/Femme dans le portefeuille

(b) Répartition géographique du nombre de bénéficiaire du portefeuille

FIGURE 1.9 : Premières caractéristiques du portefeuille

Dans la figure 1.10 la proportion d'ayants droits par rapport aux assurés est présente. Dans cette figure, le type de bénéficiaire « Parent » a volontairement été ignoré puisqu'il représentait une trop petite proportion pour être visible sur le graphique. Cependant, nous verrons par la suite que cette modalité pourra être intéressante.

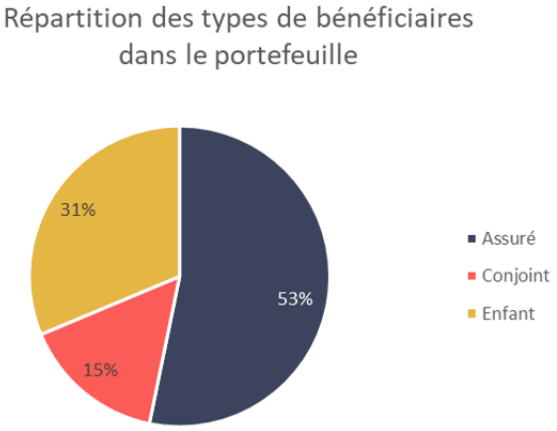


FIGURE 1.10 : Répartition des types de bénéficiaire dans le portefeuille

La répartition des âges au sein du portefeuille a été donné dans la figure 1.7. Cependant, la répartition à travers les différentes classes d'âges est similaire. De plus, les plus de 65 ans représentent moins de 1% des bénéficiaires et les moins de 18 ans sont majoritaires avec 27% du portefeuille. Ceci s'explique par le fait que cette catégorie d'âge concerne plus de personnes.

La répartition des expositions en fonction de l'âge et du type de bénéficiaire est disponible en figure 1.11 :

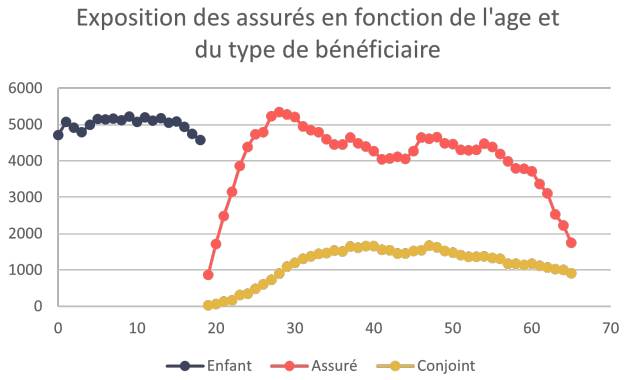


FIGURE 1.11 : Exposition en fonction de l'âge et du type de bénéficiaire

Description de la sinistralité

Certains actes et domaines médicaux n'ont pas été modélisés dans le cadre de cette étude. Il s'agit des domaines suivants :

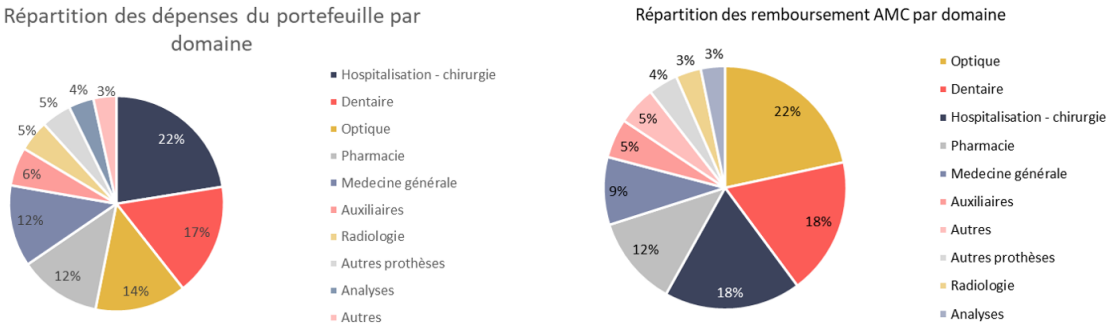
- Assistance
- Autres prothèses
- Cures
- Décès - obsèques
- Frais de transport
- Maternité
- Médecine douce
- Divers

Le choix de ne pas les modéliser a été fait puisque ces domaines n'étaient pas représentés de manière suffisante dans le portefeuille et les résultats obtenus auraient été peu robustes et très volatiles. De plus, le but de cette étude n'est pas de créer un produit d'assurance qui pourrait être commercialisé : son but premier est de réaliser une étude des déformations des domaines de santé du fait de la pandémie et de la réforme du 100% santé.

Nous avons vu que les actes sont classifiés en différents domaines, il peut être intéressant de voir quels domaines sont les plus représentés dans les frais généraux et dans les remboursements des assurances maladies complémentaires.

Nous pouvons donc observer plusieurs choses. La première qui confirme une intuition évidente, tous les domaines ne représentent pas la même proportion des dépenses totales ou des dépenses au titre de l'assurance maladie complémentaire.

Une seconde chose notable est que cette répartition des frais généraux n'est pas entièrement conservée dans la répartition des remboursements AMC. L'optique, le dentaire et l'hospitalisation restent les trois domaines les plus lourds pour l'AMC (près de 60% des remboursements), mais dans un ordre différent. En effet, l'optique représentait 14% des frais totaux de notre portefeuille mais représente 22% des remboursements de l'AMC. Ceci est expliqué par le fait que tous les soins ne sont pas remboursés de la



(a) Frais réels par domaine (b) Remboursements AMC par domaine

FIGURE 1.12 : Caractéristiques des différents domaines

même manière. De plus, l'optique est un domaine où le RAC est important (c'est d'ailleurs en partie pour cela que le 100% santé a été mit en place).

Il peut être intéressant de voir si le type de bénéficiaire semble influer sur la valeur des montants dépensés et/ou des remboursements de l'AMC. Pour cela, on regarde le montant moyen dépensé et le montant remboursé par l'AMC en fonction du type de bénéficiaire. Les chiffres seront retrouvés sur la figure 1.13. Les bénéficiaires enfants, conjoints, et assuré principal, semblent similaires dans leurs dépenses. Cependant, les remboursements au titre de l'assurance maladie complémentaire pour le type de bénéficiaire « parents » sont beaucoup plus importantes. Il faudra donc pénaliser (dans la tarification) grandement ces ayant-droits. Cependant ce type de bénéficiaires représente une très faible part du portefeuille, (0.002% des bénéficiaires du portefeuille). Il est important de souligner ce point car il explique parfaitement le phénomène de « l'anti-sélection ». De plus, le faible nombre de bénéficiaire « Parent » implique une certaine prudence quant aux résultats présentés dans le graphique. Ceux-ci peuvent varier de manière conséquente si l'on possédait plus de données.

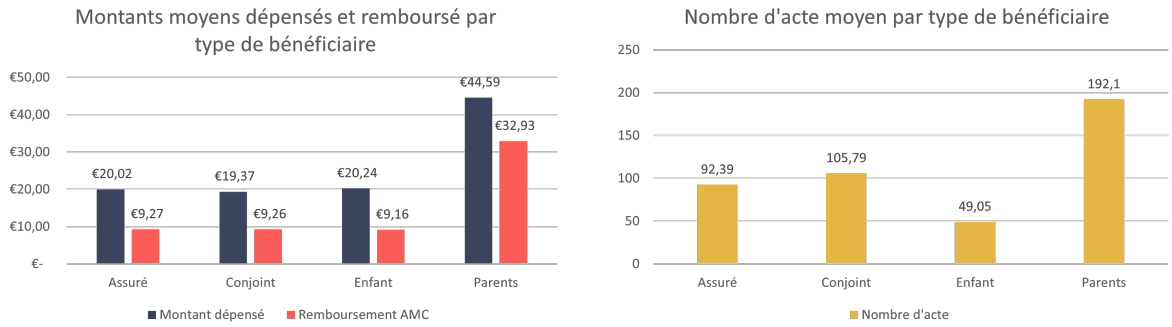


FIGURE 1.13 : Dépenses et nombre d'actes consommés des différents types de bénéficiaire

Court focus sur l'anti-sélection : L'anti-sélection ou encore sélection adverse, est une représentation de l'asymétrie d'information en assurance. En effet, c'est l'idée selon laquelle l'assuré se connait mieux lui-même que l'assureur. Cette notion est centrale en assurance santé, d'autant plus depuis la mise en place des contrats solidaires, qui interdisent la sélection à l'entrée. Les assurés qui savent qu'ils auront besoin d'une certaine catégorie de soins auront tendance à souscrire de meilleures garanties dans ces domaines, quitte à payer une prime supérieure.

Maintenant que la base de données a été présentée, l'étude qui sera réalisée ainsi que les motivations qui ont poussées à la réaliser sont présentés en détail.

1.3 Motivation de l'étude

Dans cette partie une présentation succincte de l'étude et de ses motivations est réalisée. Il sera également décrit les principales méthodes qui seront mises en place.

1.3.1 Tarification

Présentation théorique des méthodes utilisées

Des méthodes statistiques seront utilisées pour permettre la tarification, elles seront présentées d'un point de vue théorique avant d'être utilisées. De plus, certains résultats mathématiques seront démontrés quand ce sera nécessaire. Il s'agit des méthodes suivantes :

- la méthode coût-fréquence
- les modèles linéaires généralisés
- la méthode de la forêt aléatoire
- la méthode du XGBoost

Le modèle de référence

L'objectif principal de ce mémoire est de comparer une tarification sur une année courante (2019) et sur une année « exceptionnelle » (2020) afin de préparer une tarification future. En effet, l'année 2019 n'a pas été impactée de manière significative par la réforme du 100% santé et la pandémie de COVID-19 s'est déclarée début 2020. Ce modèle nous servira alors de référence et de point de comparaison pour la suite. L'étude cherche à montrer si se poser la question de la tarification de l'année 2022 est pertinent dans la mesure où nous pouvons supposer que la pandémie sera terminée mais pas le 100% santé. Dans la pratique, une tarification de contrat se fait à partir de plusieurs années de données afin d'éviter les fluctuations mineures qui peuvent arriver mais impacter la tarification. Ce n'est pas fait ici puisque seules les données postérieures à 2019 sont disponibles et complètes sont à notre disposition.

La tarification sera réalisée via une approche coût-fréquence. Cette méthode sera présentée en détail dans le chapitre 2. Pour résumer cette méthode, elle consiste à modéliser de manière indépendante le nombre de sinistres et le coût moyen de ceux-ci. Le tarif moyen est alors obtenu en multipliant les deux termes. C'est une méthode classique en assurance non-vie.

Le nombre de sinistres sera alors modélisé de plusieurs manières via différentes méthodes statistiques. La qualité de chacune des méthodes sera calculée via une certaine métrique. La méthode qui minimisera l'erreur sera alors retenue pour la tarification. Le coût sera modélisé de la même façon. Les techniques utilisées seront des modèles linéaires généralisés avec différentes lois ou fonctions de lien, et des forêts

aléatoires, méthode de Machine Learning qui est une extension de la méthode CART (Classification And Regression Tree). En ce qui concerne l'évaluation de la qualité du modèle, l'idée est de comparer les écarts entre les prédictions du modèle et les résultats observés dans la pratique. La métrique RMSE (Root Mean Square Error) sera utilisée. Ainsi nous possèderons une valeur pour chacun des modèles et le modèle qui aura la plus petite valeur sera donc le modèle le plus adapté à la tarification et sera donc le modèle retenu. L'intérêt de faire cette multiple modélisation est de s'assurer de choisir la meilleure modélisation possible avec les données dont on dispose.

Afin de tester la qualité de chacun des modèles, la base de données sera découpée aléatoirement en deux échantillons : une base dite « d'apprentissage » et une base dite de « test ». La première comportera 85% des données de la base de départ et la base de test comportera le reste des données. Cette découpe permettra de comparer la qualité du modèle sur des données dont le contexte est similaire. Ainsi, pour la base de test, nous aurons à la fois les prédictions du modèle ainsi que la réalisation réelle, nous pourrons alors calculer les écarts. De plus comme les modèles n'auront pas été calibrés sur les données de cette base, les résultats ne seront pas biaisés.

Une fois ce modèle fait, on possèdera une tarification de référence pour pouvoir étudier les impacts de la pandémie. Les impacts seront donc étudiés dans la suite. Comme les données de l'année 2020 sont également en notre possession, il sera possible d'appliquer le modèle afin de regarder si celui-ci peut s'appliquer à cette année.

1.3.2 Analyse des impacts

La tarification de référence a volontairement écarté les données de l'année 2020. Le but de cette partie est de quantifier les éventuelles variations entre une année non troublée et une année qui l'est. Si les dernières années avaient un contexte extérieur comparable il aurait été possible de s'arrêter au premier modèle et l'appliquer aux nouveaux bénéficiaires pour en déduire leurs primes sans se poser la question de l'adaptabilité du modèle. Cependant, ces années ont été marquées par divers facteurs qui ont changé en profondeur notre manière de vivre. Il se peut donc que la consommation de soin de santé soit modifiée et donc les cotisations le seraient aussi en conséquence. C'est pourquoi ce modèle sera testé sur les données de l'année 2020 (l'année 2021 ne sera pas regardée car celle-ci n'est pas terminée). Il sera alors possible de quantifier les variations des dépenses de santé entre 2019 et 2020.

De la COVID-19

Comme annoncé, les premiers impacts qui seront évalués seront ceux de la COVID-19. La pandémie sera alors présentée en détail. Il est en effet important de comprendre les retombées de cette pandémie sans réel précédent. Notre société a en réalité connu de nombreuse épidémies mais rien de réellement comparable à la COVID-19 comme le montre le tableau 1.3 des épidémies des 100 dernières années.

La sévérité pandémique est un indicateur de l'intensité de la pandémie sur une échelle de 1 à 5 développé par les CDC (Centers for Disease Control and Prevention) aux États-Unis. De plus, notre société n'a jamais été aussi mondialisée qu'elle peut l'être aujourd'hui et donc beaucoup plus internationnale qu'elle a pu l'être par le passé. Ces analyses seront effectuées en détails par la suite.

TABLE 1.3 : Récapitulatifs des épidémies depuis 1918

Épidémie	Dates	Nombre de cas	Nombre de Morts	Sévérité pandémique
COVID 19	2019 - ?	204 million +	4.3 millions +	5
Grippe H1N1	2009 -2010	0.7- 1.4 milliard	151 000 - 575 000	1
Grippe de Hong kong	1968 - 1969	0.25 - 1 milliard	1 - 4 millions	2
Grippe asiatique	1957 - 1958	0.25 - 1 milliard	1 - 4 millions	2
Grippe Espagnole	1918 - 1920	500 millions	20 - 100 millions	5

Des statistiques sur les dépenses de santé seront réalisées afin de voir si les années touchées par la COVID-19 ont vu le comportement des français changer. Pour rappel, la consommation de soins et biens médicaux augmentent chaque année d'environ 2% par rapport à l'année précédente. Les évolutions des dernières années sont données dans la figure 1.14. Si les dépenses ont plus augmenté que cela en 2020, alors les cotisations payées risquent d'être insuffisantes et les assureurs auront réalisé une perte technique. A l'inverse si l'évolution est inférieure ils réaliseront un gain technique.

D'un côté, nous pouvons nous attendre à ce que les dépenses augmentent du fait des nombreux cas de COVID, cependant d'autres domaines tels que le dentaire ou l'optique ont été obligé de fermer pendant le confinement et donc ont du moins facturer, une étude sera réalisée afin de savoir si ces effets se neutralisent ou si l'un des deux l'emporte.

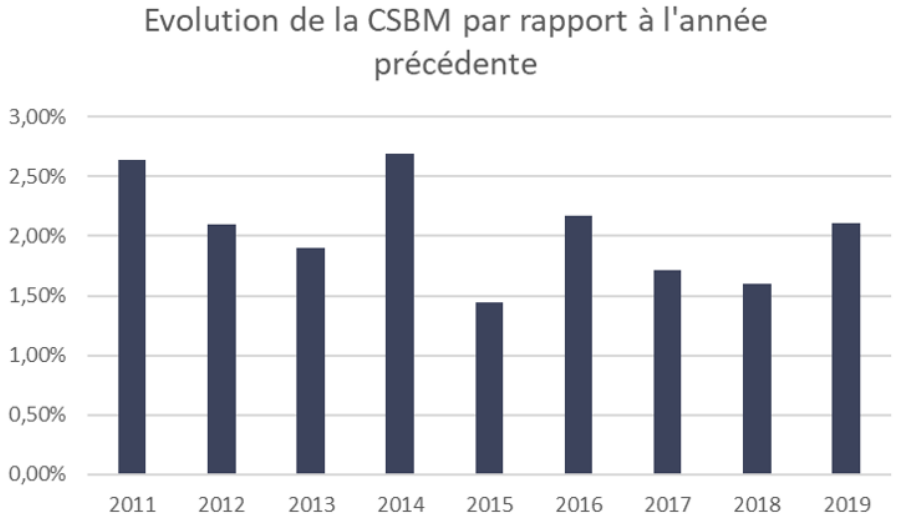


FIGURE 1.14 : Évolution de la consommation de soins et de biens médicaux depuis 2011

C'est pourquoi pour estimer l'impact de la pandémie sur la tarification, le modèle sera testé sur les données de l'année 2020. Les éventuelles différences de résultat seront donc comparées et donc nous pourrons en déduire si la pandémie a eu un effet positif ou négatif sur le résultat technique en santé des assureurs. Pour cela, les ratios entre la charge sinistre annuelle observée et les primes récoltées sur l'ensemble du portefeuille si nous avions appliqué le modèle calibré sur les données de l'années non-troublée seront étudiés.

Du 100% santé

Le 100% santé, qui a pour but de fournir des soins pour certains actes avec un reste à charge nul, implique forcément une hausse des remboursements des assurances complémentaires. Cependant si les complémentaires santé voient leur dépenses augmenter, elles vont devoir s'adapter. La première solution serait de ne pas changer leur garanties, mais dans ce cas là, leurs contrats ne seraient plus responsables et elles perdront alors les avantages sociaux et fiscaux associés. L'autre possibilité serait de prendre en compte ces modifications des remboursements dans leur tarifs proposés.

La réforme ayant commencé à être mise en place en 2019, certains résultats et conclusions peuvent déjà être obtenues sur l'efficacité de cette réforme. D'après le portefeuille, il apparait qu'entre 2019 et 2020 cette réforme a permis de diminuer le reste à charge des assuré dans les domaines de l'audiologie et du dentaire mais que ce dernier a augmenté en optique (malgré une baisse du prix global). La figure 1.15 montre l'évolution du prix moyen par acte entre 2019 et 2020. Les figures similaires pour les autres domaines sont données en annexe.

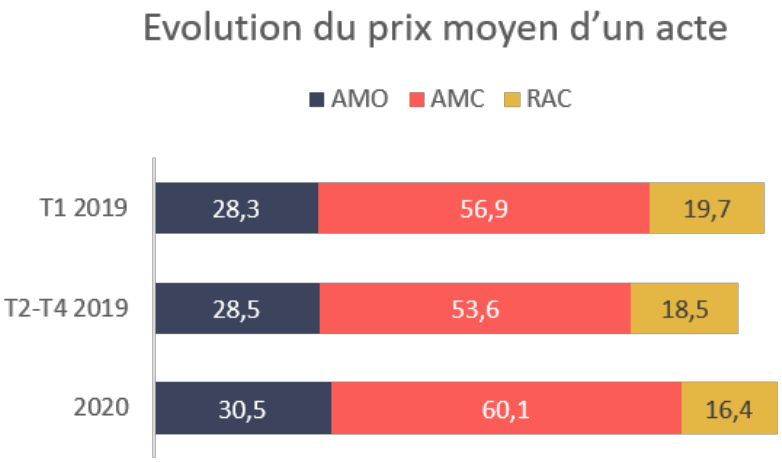


FIGURE 1.15 : Évolution du montant de remboursement de chaque acteur sur les actes dentaires depuis 2019

Ainsi les impacts de la réforme seront étudiés sur la tarification en prenant des hypothèses afin de se positionner dans un monde où la pandémie n'était pas arrivée. En effet cette réforme devait se mettre en place en 2019 mais les impacts pour les assurés devaient entrer en vigueur à partir de 2020 (voir la figure 1.6). Or depuis début 2020 la pandémie à monopolisé l'actualité et les médias. Ainsi la réforme est passée inaperçue et la population n'a pas été assez sensibilisée à celle-ci.

Maintenant que le contexte de l'étude a été présenté, la tarification du modèle de référence va être réalisée après la présentation des méthodes statistiques utilisées.

Chapitre 2

Tarification de référence

Dans ce chapitre, un modèle est réalisé afin d'obtenir une tarification de référence sur les données issues de l'année 2019. Celle-ci permettra de comparer avec l'année 2020 et justifier l'interrogation portant sur la tarification des années futures. Les méthodes seront présentées d'un point de vue théorique avant d'être appliquées.

2.1 Théorie des méthodes utilisées

2.1.1 Fréquence - sévérité

Pour beaucoup de tarification d'assurance non-vie, les assureurs ont recours à la méthode dite « Fréquence - Sévérité » qui est aussi parfois appelée « Coût - fréquence » (Lazic (2020)). Par la suite les deux terminologies seront utilisées sans distinction. C'est cette méthode qui constituera la base de la modélisation effectuée. Ce sont ces deux quantités qui seront les variables à expliquer dans les modèles de prédiction.

Présentation de la méthode

Nous cherchons à modéliser la charge totale des sinistres d'un portefeuille. En considérant qu'un assureur possède n contrats et que le coût de chaque contrat est donné par une variable aléatoire $X_i, \forall i \in \{1, ...n\}$. La charge totale de sinistres est alors donnée par la variable aléatoire S définie comme :

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

On parle alors de modèle individuel. Cependant dans les faits, le nombre de sinistres est aléatoire.La charge sinistre est alors modélisée par ce qu'on appelle le modèle collectif, n est alors remplacé par un processus de comptage au temps t qui est noté N_t. La charge sinistre totale à la date t est alors obtenue de la manière suivante :

$$S_t = \sum_{i=1}^{N_t} X_i.$$

Nous devons alors introduire des hypothèses pour pouvoir modéliser le modèle collectif. Le processus de comptage est supposé indépendant des coûts et ceux-ci sont généralement supposés indépendant et identiquement distribués ($i.i.d$) selon la même loi qu'une certaine variable aléatoire réelle X. Nous supposerons alors que $\mathbb{E}[X_i] = \mu$ et $\mathbb{V}(X_i) = \sigma^2$.

La base d'une tarification et de calculer la prime pure. Celle-ci se définie la manière suivante :

$$\pi = \mathbb{E}[S_t].$$

Dans la suite, il sera noté N à la place de N_t pour simplifier les notations, et la charge sinistre sera modélisée sur une année. On a alors :

$$\pi = \mathbb{E}[\sum_{i=1}^{N} X_i].$$

Cette égalité peut se ré-écrire en passant par l'espérance conditionnelle :

$$\pi = \mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i|N]].$$

N est mesurable par rapport à sa propre filtration on a donc $\mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i|N]] = \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i|N]]$. De plus l'espérance conditionnelle est linéaire nous pouvons donc écrire :

$$\pi = \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i|N]].$$

Or il a été supposé que N était indépendant des X_i donc on a forcément l'égalité suivante :

$$\mathbb{E}[X_i|N] = \mathbb{E}[X_i]].$$

. L'égalité devient alors :

$$\pi = \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i]].$$

Or comme les X_i sont supposé $i.i.d$, ils ont tous la même espérance μ et donc :

$$\pi = \mathbb{E}[\sum_{i=1}^{N} \mu] = \mu \times \mathbb{E}[N].$$

Ce qui se réécrit finalement :

$$\pi = \mathbb{E}[X] \times \mathbb{E}[N].$$

Il est bien visible ici qu'il faut modéliser le nombre moyen de sinistres d'un côté et le coût moyen d'un sinistre de l'autre. Une fois ceci fait, nous auront donc la prime pure. Comme le coût et la fréquence sont modélisés séparément, il est possible de considérer certaines variables pour l'un et des variables différentes pour l'autre. Il peut également être intéressant de connaitre la variance de notre charge sinistre. En effet, une variance forte signifie que il y aura beaucoup d'écart à la moyenne et donc un résultat technique très variable, ce qui est peu souhaitable pour les assureurs. Le résultat technique est définit comme la différence entre la somme des primes encaissées et la somme des sinistres remboursés. De plus, on a que :

- X et N sont définis sur le même espace de probabilité,
- Il est légitime de supposer que la variance de Y est finie dans la mesure où un risque de variance infinie peut être compliqué à assurer.

La variance de S est calculée en utilisant le théorème de la variance totale :

$$\mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}[\mathbb{E}(S|N)]. \tag{2.1}$$

Nous commençons par calculer le terme $\mathbb{E}[\mathbb{V}(S|N)]$.
Puisque N est supposé indépendant des X_i nous avons :

$$\mathbb{V}(S|N) = \mathbb{V}(\sum_{i=1}^{N} X_i|N) = \mathbb{V}(\sum_{i=1}^{N} X_i).$$

De plus, les X_i sont supposés indépendant les uns des autres, leur covariance est donc nulle et donc :

$$\mathbb{V}(S|N) = \sum_{i=1}^{N} \mathbb{V}(X_i).$$

Comme les X_i sont identiquement distribués et de variance σ^2 nous obtenons :

$$\mathbb{V}(S|N) = N \times \sigma^2.$$

Cependant gardons à l'esprit que N est une variable aléatoire et donc que le travail n'est pas complètement fini. Nous allons maintenant calculer $\mathbb{E}[S|N]$.

$$\mathbb{E}[S|N] = \mathbb{E}[\mathbb{E}[S|N]] = \mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i|N]].$$

Or, il a été vu dans la démonstration de l'espérance de la charge sinistres que cette dernière quantité était égale à $N \times \mu$. En reprenant l'équation 2.1, et en y appliquant la variance et l'espérance aux deux calculs que nous venons d'effectuer l'expression devient :

$$\mathbb{V}(S) = \mathbb{E}[N \times \sigma^2] + \mathbb{V}(N \times \mu).$$

Ce qui nous donne finalement puisque l'espérance est linéaire et la variance quadratique :

$$\mathbb{V}(S) = \sigma^2 \times \mathbb{E}[N] + \mu^2 \times \mathbb{V}(N).$$

En conclusion, une connaissance parfaite des lois de X et N, procure une connaissance parfaite de la loi de S. La prime pure à faire payer aux assurés ainsi que la volatilité associé au risque supporté seront alors connues.

2.1.2 Modèle linéaires généralisés

Dans cette section les modèles linéaire généralisés sont présentés. Toute cette section est grandement inspirée des notes de cours de Sophie DONNET (Donnet (2019)) pour son cours de modèles linéaires.

Modèles linéaires

Dans cette première sous section, il est présenté les modèles linéaires, leur principe, les hypothèses, leur fonctionnement ainsi que leur limites. L'idée des modèles linéaires est d'expliquer une variable réponse quantitative y en fonction de p variables explicatives $x^j, \forall j \in [\![1, p]\!]$. Il est également supposé que la relation entre la variable réponse et les variables explicatives est linéaire. Donc, pour tout individu $i, i \in [\![1, n]\!]$, et en introduisant un terme d'erreur ϵ_i :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i.$$

Ce qui se réécrit de manière matricielle :

$$Y = X\beta + \epsilon. \tag{2.2}$$

Avec :

$$X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{ et } \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Dans l'équation 2.2, le vecteur inconnu que nous cherchons donc à estimer est le vecteur β, X et Y étant connus et ϵ étant un vecteur aléatoire sur lequel il est fait les hypothèses ci-dessous :

- $\mathbb{E}[\epsilon] = 0_{\mathbb{R}^n}$, [P1]
- La variance du vecteur est constante : $\mathbb{V}(\epsilon) = \sigma^2 I_n$, [P2]
- Les termes d'erreur ϵ_i sont tous indépendants, [P3]
- (Dans le cas d'un modèle linéaire gaussien) ϵ suit une loi normale, [P4]

Le vecteur β est alors estimé soit par moindres carrés soit par maximum de vraisemblance. Une estimation de ce vecteur β peut être donné par la formule :

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

qui est trouvé en résolvant le problème d'optimisation :

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||^2.$$

Ces quatre postulats sont donc à vérifier pour pouvoir utiliser un modèle linéaire gaussien. L'utilisation des GLM permet de relâcher les hypothèses et de s'affranchir des postulats.

Modèles linéaires Généralisés

Maintenant que les modèles linéaires ont été présentés leur généralisation est à présent décrite : les GLM. Pour cela, il sera nécessaire d'introduire les notions de famille exponentielle et de fonction de lien. En effet, un GLM est la donnée d'une fonction de lien et d'une densité de probabilité. L'intérêt des GLM est leur relative simplicité et leur grande qualité d'adaptation. En effet, les modèles linéaires impliquent que la variable réponse puisse prendre ses valeurs dans $\mathbb{R}$ tout entier, ce qui peut être complètement faux si nous cherchons à modéliser le nombre d'acte de santé consommés dans une année.

La matrice X est redonnée ici ainsi que différentes écritures qui pourront être utiles par la suite.

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^{p+1} \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^{p+1} \end{pmatrix} = \begin{pmatrix} X^1 & \cdots & X^{p+1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Les X^i représentent chacune des variables explicatives (sauf la première colonne qui correspond à l'intercept) et que les x_i représentent chacun des individus.

On appelle fonction de lien une fonction g telle que :

$$g(\mathbb{E}[Y_i]) = x_i\beta$$

Et en notant E l'espace de $\mathbb{E}[Y]$.

$$g : E \to \mathbb{R}$$

La fonction de lien permet d'introduire une relation qui n'est plus linéaire en entre les variables explicatives et la variable réponse, c'est pour cela qu'il s'agit d'une généralisation des modèles linéaires.

La famille exponentielle est maintenant définie. C'est une famille de lois de probabilité qui regroupe une grande partie des lois usuelles. Une loi de probabilité appartient à la famille exponentielle si :

$$f(x) = \exp\left\{\frac{x\theta - b(\theta)}{\gamma(\phi)} + c(x,\phi)\right\}.$$

Avec :

- f la fonction de densité de la loi étudiée
- c une fonction dérivable
- b une fonction 3 fois dérivable et telle que b' sa première dérivée est inversible
- θ et ϕ sont deux paramètres appelés respectivement naturel et de dispersion.

A titre d'exemple, la démonstration que la loi de Poisson appartient à la famille exponentielle est disponible en annexe 3.3.3.

Pour réaliser un GLM il faut donc deux éléments, une loi de probabilité pour Y et une fonction de lien. La loi de probabilité doit appartenir à la famille exponentielle. Celle-ci est choisie en fonction de la nature du sujet d'étude à modéliser. Ainsi pour un modèle qui classifie les individus dans deux catégories, il est possible d'utiliser une loi de Bernouilli, pour un modèle qui prédit un nombre d'occurrence, il est possible d'utiliser une loi de Poisson.

Le choix de la fonction de lien est souvent plus délicat. En effet, n'importe quelle bijection du domaine de $\mathbb{E}[Y]$ dans $\mathbb{R}$ peut convenir. Cependant, c'est cette fonction qui fait le lien entre l'espérance et les variables explicatives, donc son choix est très important. Généralement, la fonction dite canonique est retenue, c'est à dire la fonction telle que :

$$g(x) = b'^{-1}(x), \ \forall x.$$

Cette fonction est dite canonique car dans ce cas précis : $\theta = x\beta$. Ce choix n'est cependant pas obligatoire, des modèles avec d'autres fonctions de lien peuvent être réalisés pour voir si celles-ci s'adaptent mieux aux données.

Dans le tableau suivant il est présenté quelques lois usuelles ainsi que la fonction de lien canonique associée.

TABLE 2.1 : Lois de probabilité usuelles et les fonction de lien canonique associées

Loi de densité	Fonction de densité/masse	Fonction de lien canonique
Poisson(λ)	$f(k) = e^{\lambda}\frac{\lambda^k}{k!}$	$\log(x)$
Binomiale(n, p)	$f(k) = \binom{k}{n}p^k(1-p)^{n-k}$	$\log(\frac{x}{1-x})$
Gamma(k, θ)	$f(x) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k}$	$-\frac{1}{x}$
Binomiale Négative(n, p)	$f(k) = \binom{n}{k+n-1}p^n(1-p)^k$	$\log(x)$

La principale limite des GLM, est qu'il faut supposer une loi à priori pour leur réalisation et cela peut instaurer une erreur si cette étape n'est pas correctement réalisée ou que les données ne suivent pas une loi « classique ».

Dans ce mémoire, les GLM seront mit en concurrence avec des méthodes d'apprentissage statistique dites de "Machine Learning". Ces méthodes sont présentées dans la partie suivante.

2.1.3 Machine Learning

Le machine learning est présenté dans cette section. Le machine learning se démarque des statistiques traditionnelles dans la mesure où le modèle est uniquement construit à partir des données et pas à partir d'hypothèses de lois de probabilité. Arthur Samuel le définit comme «un domaine d'étude dans lequel les ordinateurs prennent des décisions sans avoir été programmé explicitement ». Il existe différents algorithmes de machine learning tel que les forêt aléatoires ou encore le XGBoost. Ces deux algorithmes sont présentés par la suite. Les données seront également transformées par la méthode One Hot Encoding (OHE) nécessaire pour la bonne réalisation des algorithmes précédemment cités.

One hot Encoding

Le one hot encoding est une méthode de binarisation des variables catégorielles. En effet, certains algorithmes de machine learning ne fonctionnent pas avec des variables qui ne sont pas numériques. Pour binariser les variables, la fonction `onehot` du package `onehot` (Graves (2017)) sera utilisée. Cette fonction va créer une variable par modalité de chacune des variables catégorielles et associer 1 ou 0 en fonction de si l'individu possède ou non cette caractéristique. Enfin, nous avons prit soin de supprimer une des catégories de chaque variable afin que le modèle soit identifiable.

Exemple de OHE : pour cet exemple à la couleur des yeux de 5 individus est considérée. Ils peuvent être soit bleus, soit marrons, soit verts. L'individu qui avait les yeux bleus a un 1 dans la colonne Couleur_des_yeux.bleus et un 0 pour l'autre colonne. C'est l'inverse pour ceux qui ont des yeux marrons. Enfin, pour l'individu qui a les yeux verts, il a deux 0, il appartient donc obligatoirement à la dernière modalité possible.

Données avant OHE

Individu	Couleur_des_yeux
1	Bleu
2	Marron
3	Vert
4	Marron
5	Marron

Données après OHE

Individu	Couleur_des_yeux.bleu	Couleur_des_yeux.Marron
1	1	0
2	0	1
3	0	0
4	0	1
5	0	1

Figure 2.1 : Exemple simple de One Hot Encoding

Arbre CART

Les arbres CART sont à la fois un modèle de régression et de classification. Ils permettent donc à la fois de modéliser des variables réponses quantitatives et qualitatives. Leur principe général est de partitionner l'espace des données de manière récursives afin de déterminer une règle de décision.

Pour l'explication suivante il est supposé que nous avons p variables explicatives qui sont $X^1, \cdots, X^p$ qui expliquent une variable quantitative notée de manière classique Y. L'algorithme répète des étapes qui seront détaillées ci-dessous tant qu'un certain critère d'arrêt ne sera pas atteint. Supposons maintenant être à l'étape M et donc que l'espace ait été divisé en M régions. CART procède alors de la manière suivante :

- CART choisit une variable explicative X_j
- CART choisit une valeur seuil s (ou un ensemble non-vide dans le cas d'une variable qualitative)
- CART divise alors la région choisie en deux parties (pour les variables qualitatives on remplacera « $\leq$ » par « $\in$ » et « $\geq$ » par« $\notin$ ») : $R(j,s)^+ = \{i | x_i^j \leq s\}$ et $R(j,s)^- = \{i | x_i^j \geq s\}$

La variable X_j et le seuil s ne sont cependant pas choisis aléatoirement, ils sont choisis de manière à résoudre le problème d'optimisation suivant :

$$\min \left[\sum_{x_i \in R(j,s)^+} (y_i - m_1)^2 + \sum_{x_i \in R(j,s)^-} (y_i - m_2)^2 \right].$$

Où m_1 et m_2 sont donnés comme les valeurs moyennes des variables réponses pour chacune des régions. De plus, x_i correspond à la valeur de la variable explicative j pour l'individu i. Ces étapes se répètent alors jusqu'à validation d'un certain critère d'arrêt. Cependant certains arbres peuvent être donc très profonds et conduire à un sur-apprentissage des données. Pour contourner ce problème, les arbres sont « élagués », c'est à dire qu'ils sont coupés selon une certaines règle de décision.

Une fois que l'algorithme s'arrête nous possédons alors une règle de décision pour classer un individu (dans le cas d'une variable réponse qualitative).

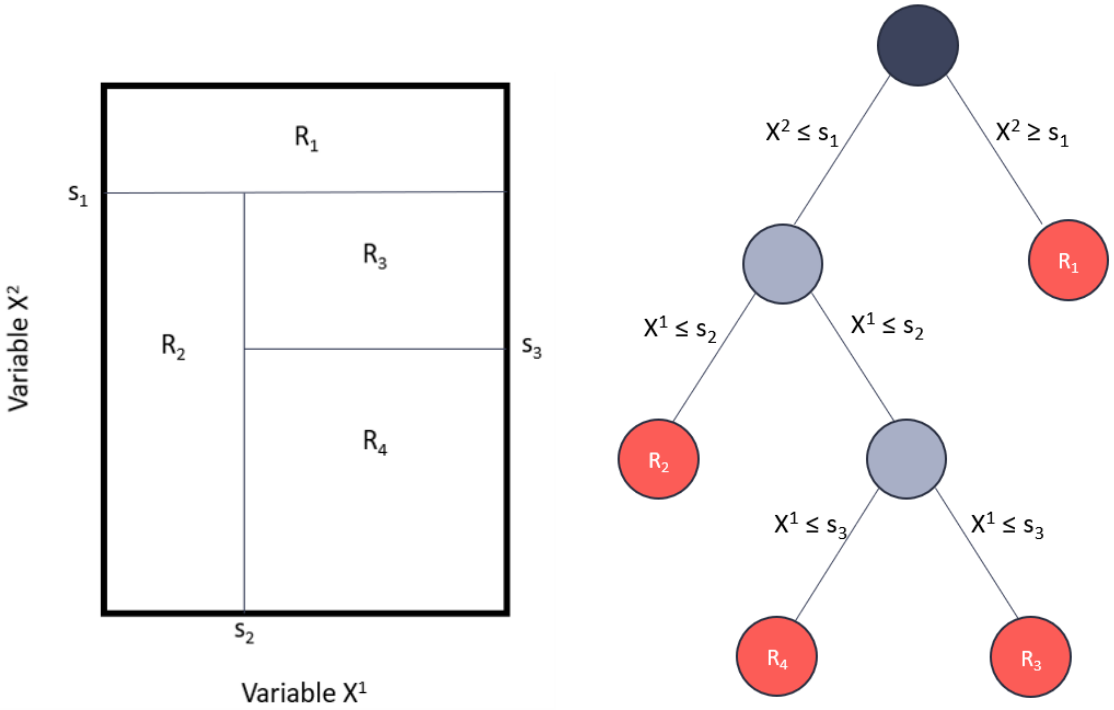
Nous possédons donc notre arbre de décision. En notant z un nouvel individu alors la prédiction associée sera notée $\hat{y}(z)$.

Exemple d'arbre : La découpe d'un domaine est alors réalisée grâce à l'algorithme CART et la représentation graphique en est donnée sous forme d'arbre dans la figure 2.2. Ici, le domaine est tout d'abord divisé en 2 régions selon la frontière $X^2 = s_1$. Puis la région $\{X^2 \leq s_1\}$ est divisée selon la frontière $X^1 = s_2$. Enfin la région $\{X^1 \geq s_2\}$ est divisée selon la frontière $X^1 = s_3$. Ainsi, la région d'un individu z qui a pour caractéristiques $X_1 = \frac{s_1+s_3}{2}$ et $X_2 = s_1 - 1$, sera la région R_2.

Deux visions différentes de l'algorithme CART ont été représentée sur le domaine. Sur la seconde vision, le cercles représentent les nœuds, en rouge sont représentés les nœuds terminaux.

Forêt aléatoire

Il a été vu comment construire un arbre de décision grâce à l'algorithme CART, il est donc maintenant présenté comment construire une forêt aléatoire. L'idée de la forêt aléatoire est de construire plusieurs arbres de décision afin de contourner les problèmes de sur-apprentissage potentiellement introduit par la présence d'un unique arbre. Pour cela, un certain nombre d'arbres indépendants sont construit. Pour cela, B tirages bootstrap des données d'origines seront réalisés et pour chacun de ces nouveaux jeux de données l'arbre CART associé sera construit. De plus l'ensemble des variables explicatives ne sera pas considéré mais seulement une partie d'entre elles qui seront choisies aléatoirement (en général un tiers du nombre de variables totales est sélectionné). Nous avons donc maintenant B arbres différents, notés $T_b, \forall b \in [\![1, B]\!]$ et $\hat{y}_b(z)$ la prédiction associé à l'arbre b. La construction d'une forêt aléatoire est synthétisée dans la figure 2.3.



(a) Découpage de l'ensemble en régions (b) Arbre construit

FIGURE 2.2 : Visualisation d'un arbre CART

La méthode se termine en prenant la moyenne des prédictions de chacun des arbres pour un nouvel individu :

$$\hat{y}(z) = \frac{1}{B}\sum_{i=1}^{B} \hat{y}_b(z).$$

La principale limite des forêts aléatoire est le temps d'exécution sur `R`, face à des données volumineuses la réalisation d'une forêt aléatoire peut être extrêmement chronophage.

Focus sur le bootstrap : Les techniques de bootstrap sont des techniques de réplication du jeu de donnée selon le principe du ré-échantillonnage. Tout d'abord un nombre d'échantillons à réaliser sera défini, il sera noté comme précédemment B. Admettons que nous avons n individus dans nos données, n tirages aléatoire avec remise dans nos données seront réalisés afin d'obtenir un echantillon bootstrap. Cette étape est alors répétée B fois et nous avons donc nos différents échantillons.

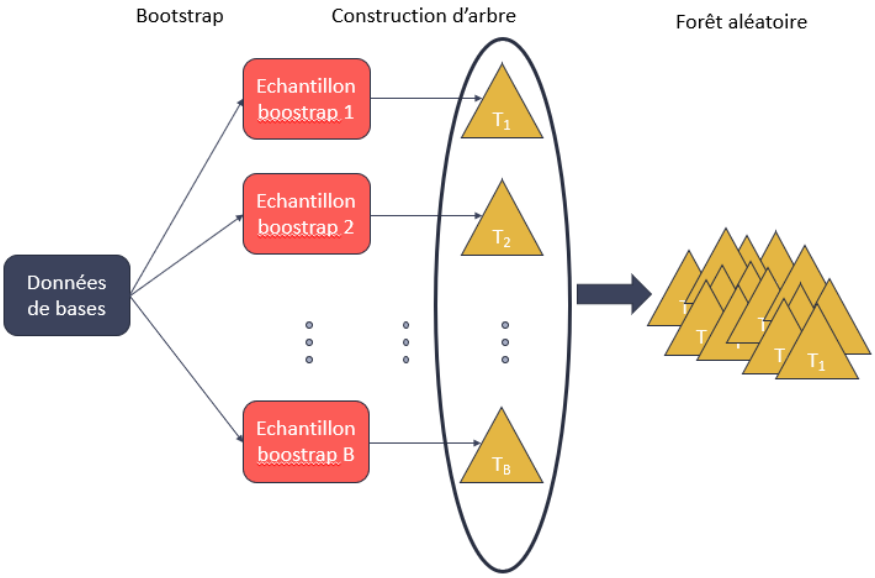


FIGURE 2.3 : Schéma de la création d'une forêt aléatoire.

XGBoost

Enfin l'algorithme XGBoost (eXtreme Gradient Boosting) est présenté de manière succincte. Il s'agit d'un algorithme de machine learning très efficace pour la prédiction de variables qualitatives ou quantitatives. Il s'agit d'un algorithme d'apprentissage supervisé qui combine différentes méthodes de prédiction afin d'en obtenir une meilleure. C'est un algorithme qui agit de manière séquentielle, il construit un modèle qu'il évalue puis ajuste en fonction le poids de chaque individu pour être le plus précis possible. Cet algorithme est réputé plus précis que l'algorithme de forêt aléatoire. Il est souvent le gagnant des compétitions *Kaggle*.

Tout comme la forêt aléatoire, le XGBoost se base sur l'agrégation de plusieurs « weak-learners »pour la composition d'un « strong-learner ». Plusieurs petits algorithmes seront utilisés afin d'en avoir un gros. C'est le même principe que la forêt aléatoire qui est une forêt composée de plusieurs arbres. La principale différence entre les deux méthodes est que pour le XGBoost, les weak-learners dépendent les uns des autres, alors que les arbres qui composent une forêt aléatoire doivent être indépendants (*Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost* (2020)).

La limite principale de l'algorithme XGBoost est son aspect mystérieux. Il est parfois qualifié de « boîte noire ». L'utilisateur en connait les entrées et les sorties mais n'a pas ou peu connaissance du fonctionnement interne (*Boîte noire (système)* 2021, de CASTEX, 2019).

2.2 Application des méthodes

Maintenant que ces méthodes ont été présentées d'un point de vue théorique, nous allons pouvoir les appliquer afin d'obtenir un modèle : le modèle de référence.

2.2.1 Sélection des variables

Avant de créer les modèles, il faut sélectionner les variables qui composeront le modèle. Cela permet d'avoir le modèle le plus adapté possible, tout en ne tombant pas dans le sur-apprentissage. Pour cela, on va faire une sélection des modèles forward-backward. Cette méthode consiste à partir du modèle qui est composé d'aucune variable explicative (il est également possible de partir du modèle composé de toutes les variables explicatives à notre disposition). Puis nous ajoutons ou nous retireons des covariables avant de comparer le critère d'Akkaike. Celui-ci est donné par la formule :

$$AIC = 2k - 2\ln(L).$$

Le nombre k dans la formule est le nombre de paramètres alors que le L est le maximum de la vraisemblance du modèle. L'idée est de minimiser ce critère pour avoir le meilleur modèle possible. En effet, plus le maximum de la vraisemblance sera élevé, meilleur sera notre modèle et plus faible sera l'AIC. Cependant un modèle avec une grande quantité de paramètres sera également pénalisé et pourra donc ne pas être retenu. On retient donc la combinaison de variables qui minimise le critère d'Akkaike.

On a donc regardé ce critère sur chacun des domaines pour les GLM et les variables ont pu donc être sélectionnées.
On commence par rappeler chacune des variables à notre disposition ainsi que leur signification.

- `type_beneficiaire`, qui définit s'il s'agit de l'assuré principal, de son enfant ou de son conjoint
- `benef_sexe`, qui définit s'il s'agit d'un homme ou d'une femme
- `age_2019`, qui est l'age du bénéficiaire en 2019
- `region`, la région d'habitation du bénéficiaire
- `naf`, le secteur d'activité du contrat auquel est rattaché le bénéficiaire
- `GarantieMOPT`, le niveau de garantie pour l'acte monture optique
- `GarantieVOPT`, le niveau de garantie pour l'acte verre optique
- `GarantiePFC`, le niveau de garantie pour l'acte prothèses fixe céramique
- `GarantieMG`, le niveau de garantie pour l'acte consultation de spécialiste
- `Taille`, l'effectif de l'entreprise souscriptrice
- `Présence_option`, la présence ou non d'une option pour l'assuré
- `CSP`, la CSP de l'assuré, Cadre, non-cadre ou ensemble du personnel

Dans le tableau 2.2 les variables qui seront retenues par domaine pour la modélisation du coût sont listées. Il est fait de même pour les modèles de fréquence, les variables retenues sont données dans le tableau 2.3. Les noms de variables définissent les lignes alors que les domaines sont en colonnes.

On liste ici les différents domaines étudiés dans cette étude ainsi que les abréviations utilisées pour parler de chacun d'entre eux :

- Analyses (Ana)
- Auxiliaires (Aux)
- Chambre particulières (CP)
- Forfait hospitalisation (FH)
- Médecine générale (MG)
- Médecine spécialiste (MS)
- Monture optique (MO)
- Pharmacie (Pha)
- Prothèses dentaire (PrD)
- Radiologie (Rad)
- Soins dentaire (SD)
- Soins hospitaliers (SH)
- Verres optique (VO)

Un point vert signifie que la variable dans la ligne sera conservée pour la colonne considérée. A l'inverse un point rouge signifie qu'elle ne sera pas incorporée au modèle.

TABLE 2.2 : Variables retenues par domaine pour la sévérité

	Ana	Aux	CP	FH	MG	MS	MO	Pha	PrD	Rad	SD	SH	VO
type_beneficiaire	●	●	●	●	●	●	●	●	●	●	●	●	●
benef_sexe	●	●	●	●	●	●	●	●	●	●	●	●	●
age_2019	●	●	●	●	●	●	●	●	●	●	●	●	●
Region	●	●	●	●	●	●	●	●	●	●	●	●	●
naf	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieMOPT	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieVOPT	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantiePFC	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieMG	●	●	●	●	●	●	●	●	●	●	●	●	●
Taille2019	●	●	●	●	●	●	●	●	●	●	●	●	●
Presence_option	●	●	●	●	●	●	●	●	●	●	●	●	●
CSP	●	●	●	●	●	●	●	●	●	●	●	●	●

TABLE 2.3 : Variables retenues pour la fréquence

	Ana	Aux	CP	FH	MG	MS	MO	Pha	PrD	Rad	SD	SH	VO
type_beneficiaire	●	●	●	●	●	●	●	●	●	●	●	●	●
benef_sexe	●	●	●	●	●	●	●	●	●	●	●	●	●
age_2019	●	●	●	●	●	●	●	●	●	●	●	●	●
Region	●	●	●	●	●	●	●	●	●	●	●	●	●
naf	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieMOPT	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieVOPT	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantiePFC	●	●	●	●	●	●	●	●	●	●	●	●	●
GarantieMG	●	●	●	●	●	●	●	●	●	●	●	●	●
Taille2019	●	●	●	●	●	●	●	●	●	●	●	●	●
Presence_option	●	●	●	●	●	●	●	●	●	●	●	●	●
CSP	●	●	●	●	●	●	●	●	●	●	●	●	●

Nous pouvons voir que comme attendu, en fonction des domaines, les mêmes variables ne sont pas conservées. De plus, certaines variables peuvent être conservées pour la fréquence et pas pour le coût et vice-versa. De plus, cette sélection de variables ne sera utilisée que pour les modèles GLM, puisqu'en effet, les forêts aléatoire ainsi que les XGBoost possèdent déjà une sélection de variable incorporée dans leur algorithme. Maintenant que nous savons quelles variables utiliser, nous allons pouvoir modéliser le coût et la fréquence pour les GLM.

2.2.2 Modélisation par GLM

Une fois les variables sélectionnées pour chacun des domaines, la prime pure sur l'année 2019 sera premièrement calculée par GLM. Comme il a été précisé plus tôt, nous utiliserons par la suite des méthodes de machine learning. De plus, le coût moyen et la fréquence seront modélisés séparément.

Coût

En ce qui concerne le coût, trois modélisations seront réalisées pour chacun des domaines. Une première en utilisant un GLM, une seconde en utilisant une forêt aléatoire et une dernière en utilisant un XGBoost. Pour le GLM, la loi sera choisie à l'aide de l'étude des QQ-plots. Il existe des tests statistiques qui permettent de confirmer ou infirmer l'adéquation à une certaine loi, comme les tests de Kolmogorov-Smirnov ou Anderson-Darling, cependant la présence de nombreuses observations conduit souvent ces tests à rejeter l'adéquation de la loi (Delignette-Muller et Dutang (2015)). C'est pourquoi les méthodes graphiques seront préférées ici.

Il a fallu choisir des lois d'adéquation parmi les lois usuelles, en général pour la modélisation du coût, les lois log-normale et gamma sont étudiées, nous nous sommes donc restreint à l'étude de l'adéquation de nos données à ces lois. Afin d'estimer les paramètres, le package `R` fitdistrplus et sa fonction fitdist ont été utilisés avec la méthode du maximum de vraisemblance. Par exemple pour le domaine prothèses dentaires les trois graphiques de la figure 2.5, montrent que la loi gamma semble la mieux adaptée à ces données. Ce sont des lois utilisées généralement pour la modélisation des coûts. De plus, il arrive que pour la modélisation de sinistres la loi de Weibull soit également considérée. Celle-ci n'a pas été regardée puisqu'elle n'appartient pas à la famille exponentielle et donc n'aurait pas pu être utilisée pour le GLM. Les graphiques concernant les autres domaines sont donnés en annexe 3.3.3. Cependant, les lois choisies pour chacun des domaines sont données dans le tableau 2.4. En ce qui concerne la fonction de lien, pour les modélisation faites avec la loi gamma choisit la fonction de lien canonique a été choisie : la fonction inverse. Pour celles réalisées avec la loi log-normale le GLM sera réalisé en considérant le logarithme de la variable d'intérêt (le remboursement de l'AMC), une loi Gaussienne et sa fonction de lien canonique : la fonction identité. Les répartitions des coûts de sinistres sont données dans la figure 2.4.

Pour rappel, les lois gamma et log-normales sont définies sur $\mathbb{R}^+$. Une variable aléatoire X suit une loi log normale si $Y = \ln(X)$ suit une loi normale. De plus, la fonction de densité de la loi gamma de paramètres $k > 0$ et $\theta > 0$ est :

$$f(x) = \frac{1}{\Gamma(k)\theta^k} \times x^{k-1} \exp\left\{-\frac{x}{\theta}\right\}.$$

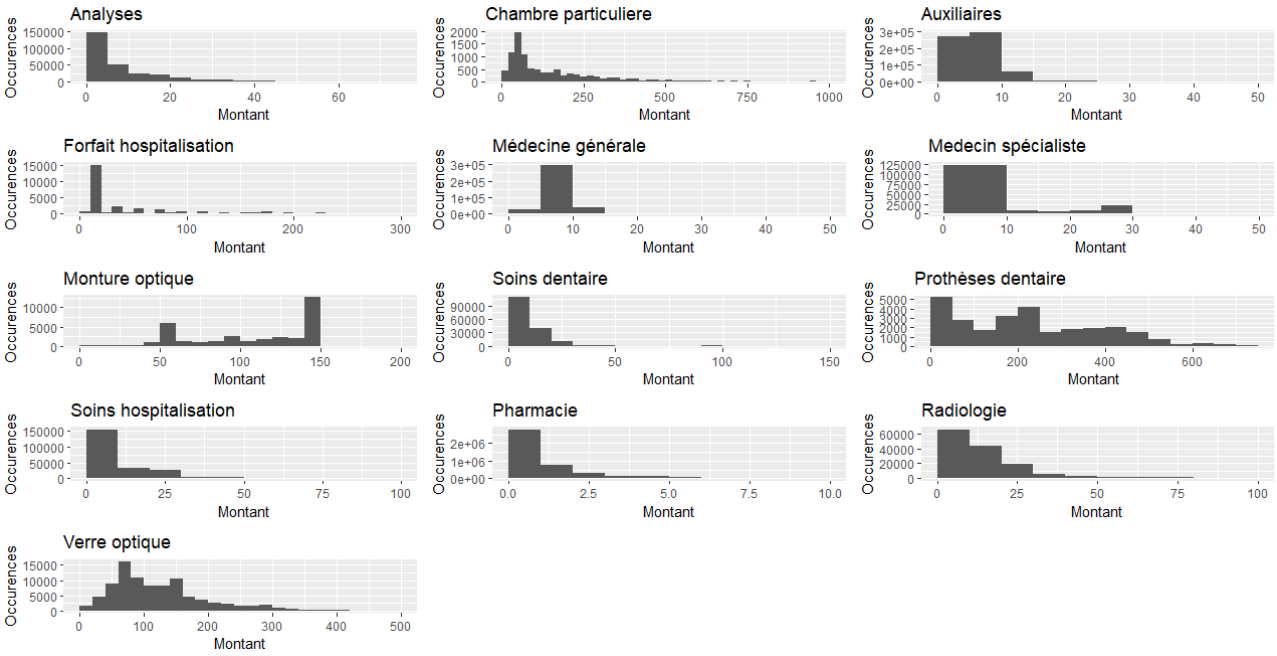
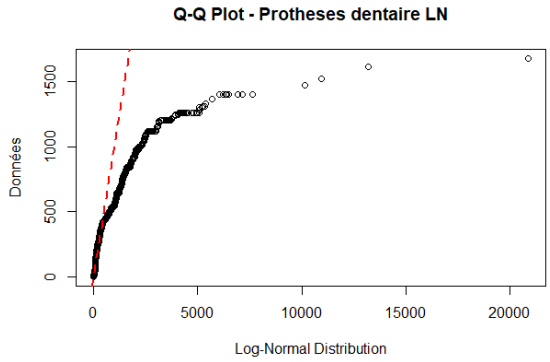
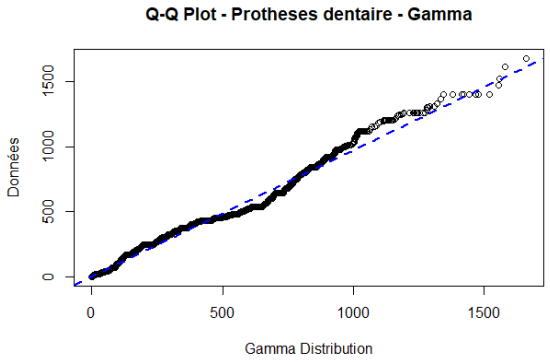


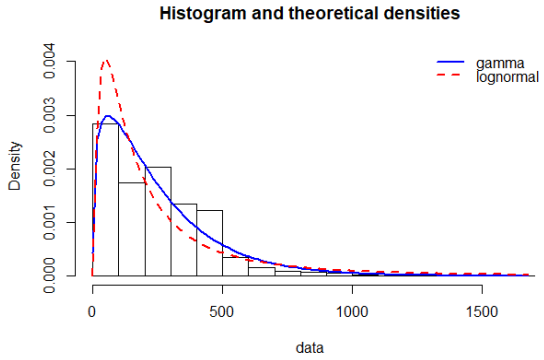
FIGURE 2.4 : Répartition des dépenses par domaine



(a) QQplot des données avec une loi log-normale



(b) QQplot des données avec une loi gamma



(c) Histogramme

FIGURE 2.5 : Choix de la loi d'adéquation pour le domaine prothèses dentaire

TABLE 2.4 : Lois retenues pour la modélisation du coût de chaque domaine

Domaine	Loi retenue pour le GLM
Analyses	Loi Gamma
Auxiliaire	Loi Gamma
Chambre particulière	Loi Log-Normale
Forfait hospitalier	Loi Log-Normale
Médecine générale	Loi Log-Normale
Médecine spécialiste	Loi Gamma
Monture optique	Loi Gamma
Pharmacie	Loi Log-Normale
Prothèses dentaire	Loi Gamma
Radiologie	Loi Gamma
Soins dentaire	Loi Log-Normale
Soins hospitaliers	Loi Log-Normale
Verre Optique	Loi Gamma

Enfin, pour que la modélisation soit la plus proche de la vérité et la plus fine possible, on a majoré le remboursement des actes à un certain seuil. Ce seuil a été choisit en prenant la valeur x_0 qui définissait le quantile à 99,5% des données de chaque domaine. Ainsi, les données supérieures à x_0 ont été ramenées à ce seuil : si on appelle la série originale Y nous avons pris, $Y_{attritionnels} = \min(x_0, Y)$). Cependant, pour quand même prendre ces valeurs en compte et donc ne pas sous-tarifer, à chacune des valeurs prédites par le modèle la moyenne des dépassements de seuils sera ajoutée.

Fréquence

Tout comme pour le coût, nous allons commencer par sélectionner les lois à utiliser pour le GLM. Une fois ceci fait, la modélisation sera réalisée par le biais d'un GLM, d'une forêt aléatoire, et finalement d'un XGboost. Enfin, le modèle retenu sera choisi par comparaison d'une certaine métrique. Le choix du modèle le mieux adapté sera fait pour chaque domaine. Ainsi le domaine « Analyse » pourra être modélisé par forêt aléatoire quand le modèle optique pourrait être modélisé par le GLM.

La fréquence des sinistres est logiquement à valeurs dans $\mathbb{N}$, les entiers naturels. Le choix de la loi pour modéliser est donc limité. En assurance non-vie, usuellement les lois de Poisson ou la loi binomiale négative sont utilisées. En pratique, la loi binomiale négative est utilisée en cas de surdispersion, quand (pour une variable aléatoire N) nous observons :

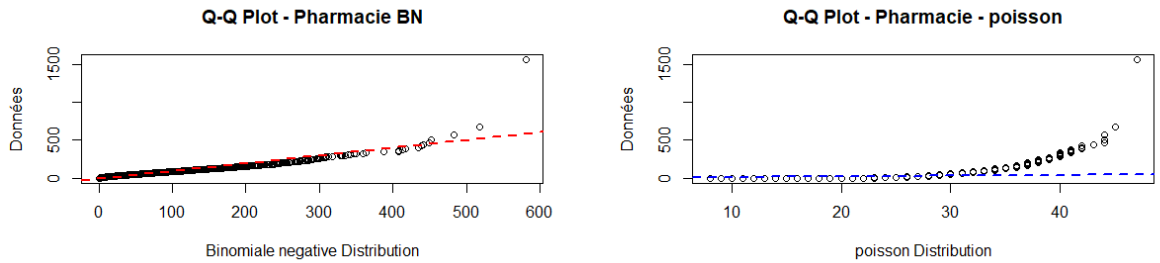
$$\mathbb{V}(N) > \mathbb{E}[N].$$

La loi de Poisson est en générale utilisée en cas d'équidispersion, c'est-à-dire pour une variable aléatoire N :

$$\mathbb{V}(N) = \mathbb{E}[N].$$

Pour sélectionner la loi pour chaque domaine il serait possible de simplement comparer l'espérance et la variance de chacune des répartitions du nombre de sinistres. Ce premier critère conduit alors à choisir la loi binomiale négative pour chacun des domaines. De plus le critère d'information d'Akaike a été calculé par le package FITDISTRPLUS et celui-ci conduisait également à choisir pour chacun des

domaines la loi Binomiale Négative. Enfin, les qqplots de chaque domaine ont également été étudiés et conduisaient à presque toujours choisir la loi Binomiale négative. Pour le domaine « Monture Optique », l'analyse graphique laissait penser que la loi de Poisson pouvait être bien adaptée également c'est cette dernière qui a donc été sélectionnée. L'espérance et la variance ont donc été comparé et ceci confirmait cette intuition. Il est donné en exemple avec la figure 2.6 les qqplots du domaine pharmacie.



(a) Loi Binomiale Négative (b) Loi de Poisson

FIGURE 2.6 : Choix de la loi d'adéquation de la fréquence pour le domaine pharmacie

La fonction de lien canonique sera également utilisée pour le modèle de fréquence. La fonction de lien canonique pour une modélisation avec une loi de probabilité binomiale négative est le lien logarithmique.

Pour rappel, une variable aléatoire N qui suit une loi binomiale négative de paramètres p et n est définie sur $\mathbb{N}$ et a pour distribution :

$$\forall k \in \mathbb{N},\ \mathbb{P}(N = k) = \binom{k+n-1}{n} p^n (1-p)^k.$$

2.2.3 Modélisation par Machine learning

Forêt aléatoire

Une fois les GLM réalisés, des méthodes de machine learning ont également été mises en place afin de challenger les GLM et d'avoir une modélisation la plus précise possible. La première méthode à avoir été mise en place était la forêt aléatoire. Cette méthode a été choisie car son principe est simple à comprendre mais elle reste efficace. Cette méthode a été présentée précédemment, on ne reviendra donc pas dessus ici. Les différentes forêts aléatoires ont été construite via le package `R` RANGER et la fonction du même nom.

Cette fonction `R` demande à l'utilisateur plusieurs arguments. Le premier et celui qui est le plus important, est le nombre de variables à sélectionner pour la création de chaque arbre. Comme évoqué dans la présentation de la méthode, en général un tiers des variables explicatives à notre disposition sont retenues. Après le one-hot encoding, 48 variables dans notre base de données étaient à disposition et donc chaque arbre a été construit en prenant 16 variables aléatoirement.

Ensuite, le deuxième paramètre important pour la création de la forêt, est le nombre d'arbre qui la compose. Ce nombre dépendra de la taille des données ainsi que de la puissance de calcul à disposition. Ici, le choix de créer chaque forêt avec 250 arbres a été fait.

Extreme Gradient Boosting

Une fois les forêt aléatoires réalisées, des XGBoost ont également été mis en place. Cette méthode de « gradient boosting » est moins répandue que les forêts aléatoires mais souvent plus efficace c'est pourquoi il a été choisi d'utiliser cette méthode de modélisation. Les XGBoost réalisés pour cette étude ont été réalisés à l'aide du package `R` XGBOOST et fonction du même nom.

Cet algorithme est caractérisé par un nombre important d'hyperparamètres qui lui octroient une grande polyvalence. Une hyperparamètre est un paramètre qui est défini avant l'exécution de l'algorithme. Parmi eux, la profondeur maximale des arbres, le pas de chaque itération et le nombre de weak-learners à utiliser.

Plusieurs combinaisons d'hyperparamètres ont été testés. Il a été alors gardé celles qui semblaient minimiser le RMSE tout en conservant un temps de calcul raisonnable. On a donc conservé les paramètres suivants :

- La profondeur maximale des arbres a été définie à 500
- Le taux d'apprentissage (qui représente le pas de chaque itération pour l'optimisation de la minimisation du RMSE) a été défini à 0,3
- 10% des variables ont été conservé à chaque weak-learner
- le nombre de weak-learner a été défini à 30

Une fois que chacune des méthodes a été réalisée, elles doivent être comparé afin de définir laquelle est la plus adaptée.

2.2.4 Comparaison des méthodes

Afin de pouvoir comparer les différentes méthodes de modélisation, les données ont été séparées en deux bases, une partie pour l'apprentissage des modèles et la seconde pour tester la qualité du modèle. C'est cette seconde étape qui nous intéresse dans cette section. Lors de la création du modèle, 15% des données ont donc été mises de coté. Le modèle est alors utilisé pour prédire le coût moyen et le nombre d'acte prévisionnel selon les caractéristiques de l'individu. On compare alors les prédictions avec la réalisation réelle (que l'on possède puisqu'elles ont été également mises de côté au préalable). La qualité du modèle est alors évaluée grâce à une ou plusieurs métrique. Cependant il faut faire attention à ce dernier point puisque différentes métriques peuvent conduire à des choix différents. C'est pourquoi pour cette étude, un seul indicateur sera considéré.

Il faut donc définir une métrique qui permettra d'évaluer le modèle. La métrique choisie dans le cadre de cette étude est le Root Mean Square Error. Elle est définie de la manière suivante :

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

Ici n représente le nombre de lignes de la base de test, y_i représente la quantité réellement observée pour la ligne i quand $\hat{y}_i$ représente la quantité prédite par le modèle pour cette même ligne.

Cette métrique est très standard quand on veut évaluer la qualité d'un modèle. Cependant, elle n'est pas très adaptée quand on veut comparer deux modèles sur deux quantités qui n'ont pas le même ordre de grandeur. En effet, pour comparer notre modèle pour le domaine « Pharmacie » et notre modèle pour le domaine « Hospitalisation », cette métrique ne sera pas très adaptée. Les dépenses dans ces deux domaines étant très différentes, il y aura forcément un RMSE pour le coût bien supérieur pour le domaine « Hospitalisation ». Pour compenser cela, il est possible de normaliser le RMSE en le divisant par la moyenne ou l'étendue par exemple. Cela ne sera cependant pas fait ici puisqu'il n'y a pas besoin de comparer les domaines entres eux.

2.3 Conclusion

2.3.1 Choix du modèle

Maintenant que les modèles ont été réalisés et que la métrique d'évaluation a été définie, il va falloir faire un choix sur le modèle qui sera utilisé pour chacun des domaines de santé étudiés.

Récapitulatifs des modèles

Chacun des modèles essayés a été donc évalué pour chacun des domaines avec les métriques définies dans la section précédente. Les valeurs de métriques ainsi calculées ont été regroupées dans différents tableaux, un tableau pour les métriques pour le coût et un tableau pour les métriques pour la fréquence.

Dans la figure 2.5 les valeurs pour le coût et la métrique RMSE sont récapitulées. Ces tableaux listent donc la racine carrée de la moyenne des écarts entre les prédictions du modèle et la réalité au carré. Il ne s'agit pas des coûts et des fréquences par domaines.

Dans la figure 2.6 on fait de même pour la fréquence.

Pour certains domaines les différentes méthodes de modélisation peuvent être proches ou très proches au sens de cette métrique. Cependant, la méthode qui possède la plus faible métrique par domaine sera retenue systématiquement.

Dans les deux tableaux récapitulatifs, la méthode qui minimisait le RMSE à été écrite en vert, c'est celle-ci qui a été choisie pour la modélisation.

TABLE 2.5 : Récapitulatif des RMSE par domaine pour la sévérité

Domaine	GLM	Forêt aléatoire	XGBoost
Analyses	9,793	9,983	9,852
Auxiliaires	4,363	2,874	4,041
Chambre particulières	24,870	22,435	24,920
Forfait hospitalisation	4,674	3,723	4,079
Médecine générale	7,516	6,256	6,909
Médecine spécialiste	8,103	7,802	8,156
Monture optique	27,713	26,120	29,229
Pharmacie	2,910	2,870	2,888
Prothèses dentaire	193,182	199,519	189,163
Radiologie	13,882	13,853	13,691
Soins dentaire	73,179	64,145	71,351
Soins hospitaliers	34,083	34,700	35,336
Verre optique	62,178	52,858	62,559

TABLE 2.6 : Récapitulatif des RMSE par domaine pour la fréquence

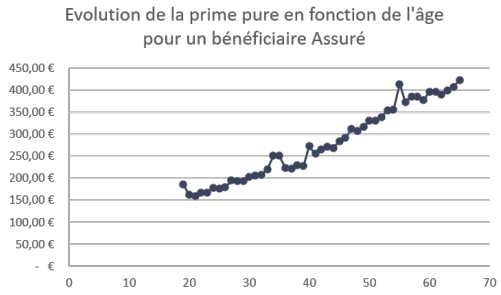
Domaine	GLM	Forêt Aléatoire	XGBoost
Analyses	9,640	9,963	9,646
Auxiliaires	14,647	15,392	14,766
Chambre particulières	2,314	2,423	2,328
Forfait hospitalisation	4,837	5,079	4,838
Médecine générale	3,146	3,245	3,133
Médecine spécialiste	3,101	3,015	2,933
Monture optique	0,368	0,380	0,366
Pharmacie	39,685	40,581	40,345
Prothèses dentaire	0,862	0,897	0,849
Radiologie	1,519	1,583	1,525
Soins dentaire	2,135	2,226	2,123
Soins hospitaliers	2,309	2,397	2,345
Verre optique	0,894	0,927	0,889

De plus, il semble important de savoir si notre modèle est de bonne qualité. En effet, en pratique, la prime pure est calculée, il est souhaitable que la réalisation soit très proche de ce qui a été calculé, sinon l'assureur réalisera un gain ou une perte technique. Pour estimer la qualité du modèle, le modèle sera donc appliqué aux données sur lesquelles il a été calibré. Ensuite la somme de toutes les primes encaissées à la somme de tous les remboursements des assureurs du portefeuille sur l'année 2019 seront

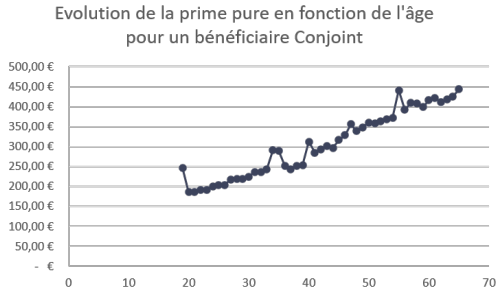
comparés. Cette comparaison sera réalisée à l'aide du ratio $\frac{S}{P}$ où S représente la charge sinistre et P les primes encaissées. Comme la prime pure a été modélisée sur ces mêmes données, nous nous attendons à avoir un résultat proche de 100%. Si ce n'est pas le cas, c'est que le modèle n'est pas de bonne qualité et il faudra alors le calibrer d'une autre manière. Ici, on a :

$$\frac{\text{Sinistres}}{\text{Primes}} = \frac{55\ 946\ 608}{53\ 845\ 042} = 103,903\%.$$

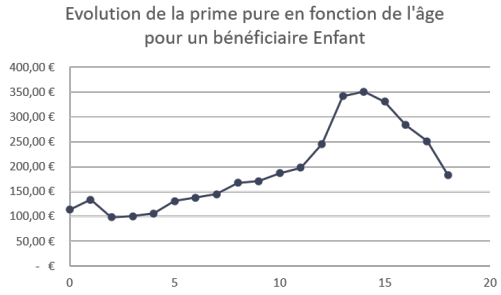
Ceci laisse supposer que le modèle est d'une bonne qualité. Un autre point de contrôle sera réalisé pour s'assurer de la bonne qualité du modèle : on va regarder l'évolution de la prime pure en fonction de l'âge toute autre chose étant égale par ailleurs. Si le modèle est cohérent, l'évolution doit être globalement croissante avec l'âge. Nous avons donc calculé à partir du modèle créé des primes pures pour des hommes en ne faisant varier uniquement l'âge. Les autres caractéristiques étaient : un homme d'Île-de-France qui ne possède pas d'option, qui a souscrit à une garantie moyenne dans chacun des domaine et dont l'assuré principal fait partie de l'ensemble du personnel dans une entreprise de 500 personne. Nous distinguons trois graphiques, le premier pour un type de bénéficiaire « Enfant » pour lequel l'âge variera de 0 à 18 ans et les deux suivants pour les types de bénéficiaires « Assuré » et « Conjoint » pour lesquels les âges varieront de 18 à 65 ans.



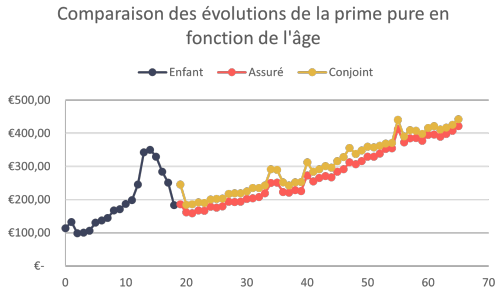
(a) Evolution de la prime pure en fonction de l'âge pour les assuré principaux



(b) Evolution de la prime pure en fonction de l'âge pour les conjoints



(c) Evolution de la prime pure en fonction de l'âge pour les enfants



(d) Comparaison des différentes primes pures par catégorie d'âge

FIGURE 2.7 : Evolution des primes pure en fonction de l'âge et du type de bénéficiaires

Il est donc observable que l'évolution du tarif en fonction de l'âge se sépare en deux périodes distinctes. Entre la naissance et 14 ans la prime pure est croissante avec l'âge puis celle-ci décroit jusqu'à la majorité. Enfin, jusqu'aux 65 ans des assurés (âge maximum étudié ici), la prime pure semble augmenter en fonction de l'âge. De plus, sur le graphique 2.7d nous pouvons observer que pour

un âge donné, un conjoint coûte plus cher qu'un assuré principal. Ceci est cohérent avec ce qu'il se passe réellement, nous pouvons donc dire que le modèle semble être de bonne qualité. Il sera donc conservé pour la suite.

Agrégation

Maintenant des modèles qui permettrons de donner la prime pure par domaine d'un nouvel individu sont créés. Cependant, il n'apparaît pas envisageable de donner à un nouvel assuré une prime pure annuelle par domaine. Il faut donc agréger toutes celles qui ont été trouvées.

Comme définit en préambule pour trouver la prime pure π d'un domaine, la fréquence n donnée par le modèle retenue est multipliée par le coût donné par le modèle sévérité c. on a alors :

$$\pi = n \times c.$$

Ceci est fait pour chacun des domaines et 13 primes pures différentes notées π sont alors trouvées (on ajoutera en indice le nom du domaine concerné). Pour calculer la prime pure globale du nouvel individu les primes pures de chaque domaines sont alors trouvées. Ceci se justifie puisque en notant S la variable aléatoire modélisant la charge sinistre, on a :

$$S_{Totale} = S_{Ana} + S_{Aux} + S_{CP} + S_{FH} + S_{MG} + S_{MS} + S_{MO} + S_{Pha} + S_{PrD} + S_{Rad} + S_{SD} + S_{SH} + S_{VO}.$$

En passant à l'espérance, en utilisant sa linéarité et en se rappelant que $\pi_{Ana} = \mathbb{E}[S_{Ana}]$ on peut alors en déduire la formule suivante :

$$\Pi_{Totale} = \pi_{Ana} + \pi_{Aux} + \pi_{CP} + \pi_{FH} + \pi_{MG} + \pi_{MS} + \pi_{MO} + \pi_{Pha} + \pi_{PrD} + \pi_{Rad} + \pi_{SD} + \pi_{SH} + \pi_{VO}.$$

En appliquant notre modèle aux données de 2019, une prime pure est alors trouvée par assuré. Dans le tableau 2.7 sont donné les primes moyennes trouvées par domaines puis au global. Il a été fait de même pour les coûts et les fréquences.

Nous sommes maintenant capable de donner une prime pure pour tout nouvel individu qui voudrait souscrire à un contrat. Cependant le but de cette étude est d'évaluer les différents impacts qu'ont pu avoir la COVID-19 ou encore le 100% santé. C'est ce qui va être vu dans le chapitre 3.

2.3.2 Limites de la modélisation

Afin de pouvoir apporter un jugement critique sur l'étude, il est important d'en définir les différentes limites, plus ou moins importantes, de celles-ci. C'est ce qui est réalisé dans cette section.

L'approche Coût - fréquence : Pour pouvoir obtenir une prime pure dans cette étude, une approche coût-fréquence a été réalisée. Bien que classique en tarification santé, cette approche repose sur des hypothèses mathématiques fortes. Premièrement, le processus de comptage qui donne la fréquence est supposé indépendant du coût des sinistres. Cependant, cela n'a pas été vérifié ici et il serait même

TABLE 2.7 : Prime, fréquences et coût moyen par domaine

Domaine	Cout moyen (€)	Fréquence moyenne	Prime moyenne (€)
Analyses	8,45	1,66	**14,00**
Auxiliaires	5,59	3,31	**16,87**
Chambre particulières	57,13	0,18	**10,74**
Forfait hospitalisation	18,53	0,44	**8,06**
Médecine générale	9,57	2,04	**19,51**
Médecine spécialiste	8,82	1,54	**13,07**
Monture optique	103,62	0,18	**18,83**
Pharmacie	1,68	26,94	**45,43**
Prothèses dentaire	232,35	0,15	**36,09**
Radiologie	14,62	0,74	**10,84**
Soins dentaire	31,24	0,92	**27,65**
Soins hospitaliers	15,79	1,19	**19,19**
Verre optique	116,90	0,46	**56,64**
Total	48,02	3,06	**296,92**

possible d'ajouter que cette intuition ne semble pas évidente. En effet, prenons par exemple le domaine « Analyses ». La succession d'analyses différentes pourrait conduire à réaliser des analyses de plus en plus complexes et donc coûteuses. Dans ce cas, l'indépendance ne serait donc clairement pas vérifiée. De plus, pour cette approche les coûts sont supposés $i.i.d$ ce qui n'a pas non plus été vérifié.

La notion de temporalité : Dans cette étude, nous nous sommes restreint pour la tarification à l'année 2019. Ceci implique une faible profondeur des données. De ce fait, il est possible de ne pas prendre en compte certains effets allant dans un sens où dans l'autre. C'est pourquoi il aurait été préférable de posséder des données sur plusieurs années complètes pour réaliser une tarification de référence plus robuste. De plus, la tarification a été réalisée annuellement. Il aurait pu être intéressant de regarder ce qu'il se passe mois après mois. Certains mois étant possiblement plus couteux pour les assureurs. Enfin avec ceci, il aurait également été possible de comparer précisément les effets des différents confinements qui ont marqué l'année 2020.

Les garanties : Enfin, une tarification classique en santé repose sur deux grandes familles d'éléments. Les caractéristiques de l'assuré ainsi que les garanties dont il dispose pour chaque domaine. Dans cette étude, une simplification a été faite en ne prenant que trois niveaux possibles de garanties pour certains domaines uniquement. C'est une importante simplification qui a donc été faite ici.

Une tarification de référence sur l'année 2019 a donc été réalisée. Il est possible de donner une prime pure (et donc a peu de chose près) un tarif annuel pour un nouvel assuré. La suite de l'étude consiste à faire de même avec l'année 2020 et a étudier les variations (si elles existent) dues à l'épidémie de la COVID-19 et au 100% santé. Ceci clôture ce chapitre technique mais nécessaire pour l'analyse et la réflexion portée dans le chapitre 3.

Chapitre 3

Analyse des facteurs extérieurs

Maintenant que la tarification de référence a été réalisée, les effets de la pandémie ainsi que du 100% santé vont pouvoir être étudiés. Pour cela, la pandémie sera présentée en détail avant de regarder les impacts que celle-ci a eu sur la consommation d'actes de santé. Enfin le modèle trouvé au chapitre 2 sera appliqué aux données de 2020 afin de voir si celui-ci s'adapte à une période de pandémie ou a une période post-pandémique. En d'autres termes, il sera observé si la pandémie a eu un impact sur la tarification future des contrats d'assurance santé. Ensuite, il sera fait de même pour la réforme du 100% santé.

3.1 Pandémie de COVID-19

Le premier impact qui sera analysé est donc la pandémie de COVID-19 qui touche le monde entier depuis janvier 2020. Il est légitime de se poser la question de son impact sur les contrats d'assurance et d'en évaluer l'impact afin d'adapter l'offre de prestation pour les années futures.

3.1.1 Présentation de la pandémie

Pour commencer ce chapitre, la pandémie est présentée afin de comprendre l'impact de celle-ci sur notre société moderne et de rappeler le contexte de l'étude. Ainsi il sera possible de comprendre pourquoi il est justifié de se demander les éventuels impacts que peuvent avoir un tel évènement sur chacun des secteurs de notre vie, y compris l'assurance santé et donc éventuellement adapter le tarif présenté pour les années futures.

Chiffres et particularité

La pandémie dite de la COVID-19 est une pandémie mondiale d'une maladie infectieuse provoquée par le coronavirus SARS-Cov-2. La maladie se déclare en Chine pendant le mois de novembre 2019 et se répand rapidement à travers le monde. L'OMS prononce l'état d'urgence de santé publique de portée internationale à la fin du mois de janvier 2019. L'entièreté du globe est alors touché par cette maladie.

Les symptômes sont très variés mais sont le plus souvent de la fièvre, de la toux, de la fatigue et une éventuelle perte du goût et de l'odorat. Cependant le virus est marqué par une proportion importante de malades asymptomatiques et de cas plus graves qui impliquent une prise en charge médicale importante entraînant même une surcharge des services de réanimations (**noauthor˙pandemie˙2021**).

Pour tenter d'endiguer la pandémie le gouvernement français a instauré un confinement qui imposait de rester chez soi et la fermeture des commerces non-essentiels, ceci incluant les opticiens et les dentistes qui nous intéresseront particulièrement. Comme la pandémie n'est pas encore terminée, tous les chiffres sont susceptibles de changer.

TABLE 3.1 : Chiffres COVID-19 au 31/05/2021

Localisation	Nombre de cas cumulé	Décès cumulés
Monde	167 millions	3.5 millions
Europe	32 millions	720 000
France	5.6 millions	110 000

D'autres caractéristiques importantes d'une épidémie sont le taux de reproduction ainsi que la létalité. En ce qui concerne le taux de reproduction, on distingue le taux de reproduction initial (noté R_0) et le taux de reproduction effectif (noté R_{eff}). Le R_0 permet de comparer les différentes épidémies puisqu'il ne varie pas avec le temps. Le R_{eff} correspond au nombre de contaminations induites par une personne malade au temps t. Le R_0 était pour la COVID de trois (une personne infectée transmet en moyenne la maladie à trois personnes) ce qui est synonyme d'une grande contagion. De plus la létalité du virus est estimée à 3%. Le seul précédent historique à ces niveaux là est la grippe espagnole de 1918.

Précédents historiques

Pour terminer cette introduction, les dernières pandémies mondiales en date sont maintenant présentées succinctement. Le XX^{e} siècle a été marqué par plusieurs crises pandémiques. Les principales étaient les grippes dites asiatiques et de Hong Kong respectivement en 1957 et 1968. Ces dernières sont similaires en tout point : le taux de reproduction initial, nombre de cas estimé entre 250 millions et 1 milliard et nombre de décès entre 1 et 4 millions. On pourrait également citer la grippe H1N1 de 2009 ou le SARS de 2002 (entre 700 millions et 1,4 milliards de cas pour moins de 600 000 décès). Ces pandémies sont donc assez différentes de celle de la COVID-19 par leur taux de mortalité relativement faible (*Grippe asiatique* 2021, *1957-1958 Pandemic (H2N2 virus)* 2019).

Enfin, un court paragraphe sur la dernière pandémie comparable à celle de la COVID 19 : la grippe espagnole. Celle-ci s'est déclarée en 1918 en Chine avant de se répandre partout dans le monde du fait de la première guerre mondiale. Elle est appelée « espagnole » car seuls les médias espagnols parlaient de cette nouvelle maladie. En effet, l'Espagne était alors neutre durant la guerre et donc pas touchée par la censure de l'armée. Cette pandémie est caractérisée par une virulence particulière du virus puisqu'elle a été active sur une durée relativement courte. Elle a touché environ 500 millions de personnes (ce qui représente à l'époque un tiers de la population mondiale de l'époque) et tuée entre 20 et 100 millions de personnes.

Il a donc été vu que le seul précédent historique réellement comparable en terme de dangerosité est la grippe espagnole. Cependant les différences restent énormes, une mortalité bien plus forte à l'époque, une société beaucoup moins mondialisée, un système de remboursement des frais de santé qui n'avait rien à voir. Nous pouvons donc dire que la pandémie de COVID-19 est sans réel précédent historique. C'est pourquoi on ne pourra pas comparer nos résultats avec ce qu'il s'est passé par le passé. Ainsi, cette étude soulève d'importantes questions.

3.1.2 Impacts sur la consommation de soins

Tout l'enjeu de la gestion de la pandémie est de réduire au maximum le nombre de morts du fait de cette pandémie. Pour cela, l'état veut réduire le nombre de personnes en réanimation et donc le nombre de cas. En effet, nous avons pu observer que les services de réanimation ont été surchargés pendant la pandémie. Ceci devrait induire une augmentation de la consommation de soins en hospitalisation et en médecine courante et donc de remboursements de l'AMC. Le tarif théorique de l'assurance maladie devrait donc augmenter en temps de pandémie. Cependant, dans le même temps, afin de contenir la pandémie et réduire le nombre de cas, on a vu que le gouvernement français (entre autre) a décidé d'instaurer un confinement et d'interdire l'ouverture des commerces dits « non-essentiels ». Cela incluait sur la période de mars à mai 2020 les opticiens et les dentistes. Ainsi, à ce niveau précis le montant remboursé sur ces domaines a sûrement diminué. De plus, les gens étant confinés, ils ont donc surement eu moins recours à la médecine en général (il y avait moins de contamination de COVID mais également des autres maladies). Nous commençons par présenter les évolutions du nombres d'actes consommés entre 2019 et 2020 sur certains domaines.

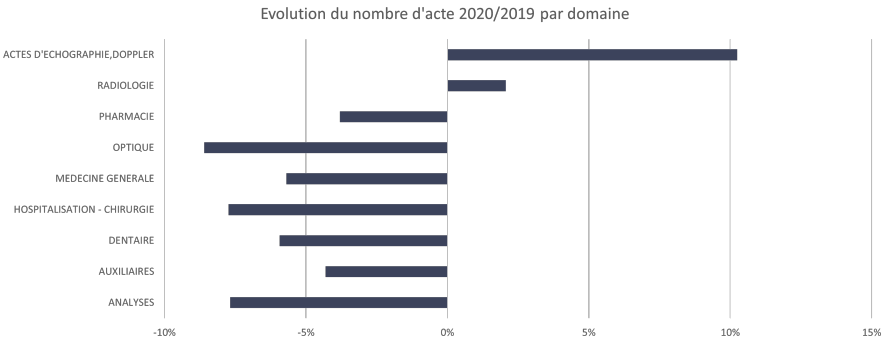


FIGURE 3.1 : Variation du nombre d'actes par domaine entre 2019 et 2020 en pourcentage

Il est notamment observable qu'en dehors du domaine « Radiologie » et de l'acte « Actes d'échographie, Doppler » tous les domaines médicaux ont vu le nombre d'actes consommés diminuer. Ceci est expliqué par le fait que certaines conséquences de la COVID peuvent être détectées par le biais de ces actes médicaux. Dans les sections suivantes, les effets de la pandémie seront étudiés en fonction de la géographie et en fonction de l'âge des assurés.

Par région

La pandémie n'a pas affecté chaque région de la même manière. Nombre de cas, de décès, tension de réanimation ou encore taux d'incidence, autant de paramètres qui peuvent varier d'une région à une autre. C'est pourquoi il peut être intéressant de regarder si la pandémie a eu un impact différent sur

notre système de santé en fonction des régions. Nous nous intéressons donc à l'évolution du nombre d'actes consommés dans chaque région en fonction du nombre de bénéficiaires de celle-ci. C'est ce qui est représenté dans la figure 3.2.

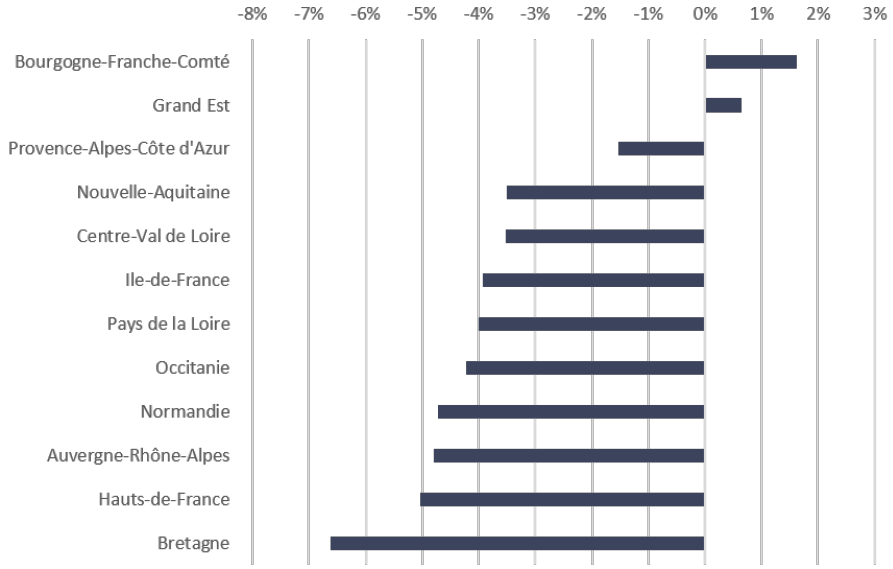


FIGURE 3.2 : Variation du nombre d'actes par bénéficiaire entre 2019 et 2020 en pourcentage

Les résultats de ce graphique sont cependant dépendants du nombre de bénéficiaires présents dans chacune des régions. Il a en effet été indiqué que toutes les régions n'étaient pas représentées de la même manière au sein du portefeuille. Il faut donc prendre en compte ce paramètre lors de la lecture de celui-ci, une sous-représentation de la région pouvant entrainer une volatilité des résultats.

Ces chiffres proviennent de la base de données initiale utilisée pour cette étude. Le nombre d'actes consommés n'a pas évolué de la même manière d'une région à l'autre. Nous pouvons notamment voir que les région Bourgogne-Franche-Comté et Grand-Est ont vu la consommation du nombre d'acte augmenter en 2020 alors qu'elles ont diminué partout ailleurs. Ceci s'explique par le fait que pendant les premières semaines de la pandémie, ce sont les régions qui ont été les plus touchées par la COVID-19. Cependant, le reste de la France a subit une baisse de la consommation d'actes de santé. Ceci représente 93% des bénéficiaires du portefeuille. Pour expliquer cette baisse globale, nous allons regarder l'évolution globale de la consommation d'actes mois par mois. C'est ce qui est visible en figure 3.3. Pendant les mois de confinement (représentés en rouge sur le graphique), le nombre d'actes a fortement diminué par rapport à l'année précédente. Alors que le reste de l'année est resté presque stable. Nous pouvons donc en conclure que c'est le confinement des mois de mars, avril et mai 2020 qui sont à l'origine de la baisse globale du nombre d'actes consommés.

Par age

De la même façon que pour les régions, la pandémie n'a pas affecté chaque génération de la même manière. En effet, nous savons que par exemple, l'âge est un facteur de risque face aux symptômes du virus. Dans la figure 3.4a, nous pouvons voir que chacune des catégories d'âges a été impactée à la

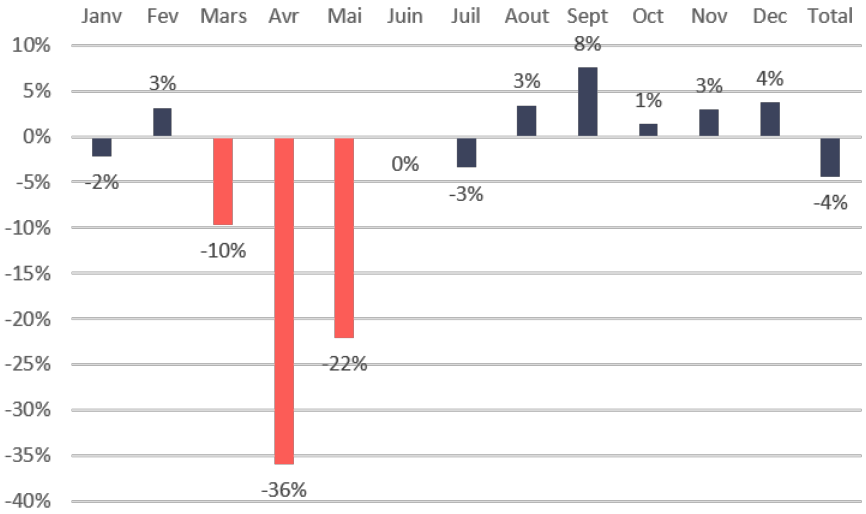
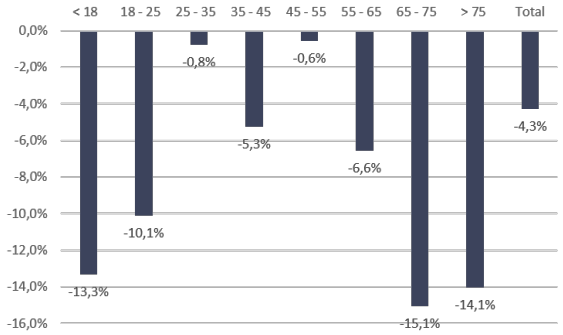


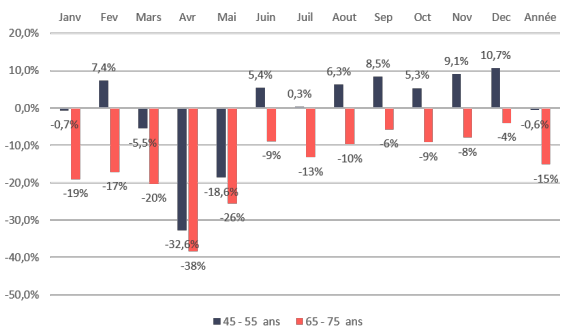
FIGURE 3.3 : Variation du nombre d'actes mensuels entre 2019 et 2020

baisse pour la consommation d'actes médicaux, cependant chacune d'elle l'a été à un niveau différent. Par exemple, la partie la plus âgée de la population (les plus de 65 ans) ont réduit leur consommation de près de 15% en 2020 par rapport à 2019. A l'inverse les 25 - 35 ans et 45 - 55 ans n'ont presque pas changé leurs habitudes entre les deux années étudiées. Cela peut s'expliquer par une appréhension de la part des populations les plus âgées par rapport au virus et donc une réticence à sortir et donc à consommer des actes médicaux divers.

Comme nous pouvons le voir sur la figure 3.4b et l'étude des évolutions mensuelles entre 2019 et 2020 sur les catégories d'âges 45 - 55 ans et 65 - 75 ans, cette différence est dûe principalement à un rattrapage de la consommation d'actes pour la première catégorie dans les mois qui ont suivis le confinement de mars à mai 2020. Il est cependant à noter que la catégorie d'âge « 65- 75 ans » est assez peu représentée dans le portefeuille et cela peut induire une faible robustesse des résultats. Cette remarque est également valable pour la figure 3.4a.



(a) Évolution des consommations d'actes médicaux par catégorie d'âge



(b) Évolution de la consommation d'actes médicaux par mois

FIGURE 3.4 : Evolution entre 2019 et 2020 pour les catégories d'âge

Les variations de consommation du nombre d'acte entre 2019 et 2020 ont été regardées, nous pouvons donc avoir une première idée des impacts de la pandémies de COVID-19. Dans la suite de l'étude, une tarification sera réalisée à partir des données de consommation de santé sur l'année 2020 et les résultats de cette tarification seront comparés avec ceux trouvés dans le chapitre 2.

3.1.3 Application du modèle

Dans la pratique, les assureurs ne refont pas la tarification des contrats d'assurance santé tous les ans. Ils estiment l'augmentation des dépenses médicales d'une année sur l'autre (les dernières augmentations sont disponibles dans la figure 1.14) et ajustent leurs tarifs en fonction de cette prévision. Cette estimation prend en compte ce qui est prévisible : l'augmentation du coût des actes, les éventuelles réformes qui vont rentrer en vigueur. Ainsi avec la réforme du 100% santé, l'augmentation des primes pour les assureurs devait être de l'ordre de 3%. En utilisant le modèle trouvé au chapitre 2, il aurait donc suffit de multiplier la prime pure par un facteur (qui sera noté f_{2020}) qui prend en compte les évolutions anticipées des dépenses de santé pour avoir une prime cohérente pour l'année 2020. Cependant les assureurs ne peuvent pas anticiper l'arrivée d'une pandémie mondiale qui va grandement changer la consommation.

Toutes les primes considérées sont des primes pures théoriques. Elles sont obtenues grâce au modèle réalisé dans la précédente section et ne sont pas présentes dans la base de données initiale. Par conséquent, elles ne prennent en compte aucun chargement, aucune taxe et aucune marge de risque.

La charge sinistre pour l'année 2020 est notée S_{2020}, les primes récoltées en 2020 sont notées P. Comme précédemment, pour que le modèle soit ajusté on veut que :

$$\frac{S_{2020}}{P} = 100\%. \tag{3.1}$$

Cependant, comme précisé précédemment, le modèle pour l'année 2020 n'est pas réalisé, nous nous contentons de multiplier les primes obtenues avec le modèle précédemment créé par f_{2020}. Pour cette année, l'augmentation estimée était de l'ordre de 3%. En notant P_{2019} les primes obtenues en appliquant le modèle créé au chapitre 2, les primes obtenues pour l'année 2020 sont trouvées en faisant : $P = f_{2020} \times P_{2019} = 1,03 \times P_{2019}$. L'équation 3.1 devient alors :

$$\frac{S_{2020}}{P_{2019} \times 1,03} = 100\% \iff \frac{S_{2020}}{P_{2019}} = 103\%. \tag{3.2}$$

P_{2019} sera obtenu en appliquant les modèles issus des données de 2019 aux bénéficiaires présents dans le portefeuille en 2020, les primes trouvées seront valables pour des périodes de couverture d'un an, elles seront donc multipliées par l'exposition. Enfin, la charge sinistre sera obtenue en sommant tous les remboursements de l'année 2020. Le ratio $\frac{S}{P}$ sera alors analysé et si le modèle est adapté on devra avoir une quantité proche de 103%.

$$\frac{S_{2020}}{P_{2019}} = \frac{47\ 027\ 571}{55\ 602\ 552} = 84,578\%.$$

Cette quantité est très éloignée du 103% recherché. Cela peut être dû à deux chose distinctes, la quantité de prime perçue est trop élevé ou la charge sinistre est trop faible. La charge sinistre 2020

est observée, c'est donc la quantité de primes que l'on obtient avec le modèle qui est trop importante. Il y a donc eu moins de sinistres en 2020 que ce qui avait été estimé.

Comme réalisé précédemment, afin de comparer les deux années, nous allons regarder le ratio charge sinistre sur primes récoltées. Pour cela, un $\frac{S}{P}$ sera calculé pour chacun des domaines les variations d'une année sur l'autre pourront êtres observées.

Présentation

Comme précédemment, les ratios de la charge sinistre divisé par les primes collectées dans chaque domaine seront étudiés afin d'observer les variations de ces ratios entre les applications du modèle sur les données de l'année 2019 et sur celles de l'année 2020. Un ratio supérieur à 100% indiquera que la charge sinistre est supérieure à la quantité de prime récoltée. Un ratio inférieur à 100% indiquera une quantité de primes supérieure à la charge sinistre.

L'étude se fait domaine par domaine puisqu'il est parfaitement possible que la consommation ait varié pour l'optique mais pas pour la médecine générale par exemple. Ceci permettra de voir les évolutions de consommation pour l'année 2020. Il est nécessaire de garder à l'esprit que les dépenses de santé augmentent d'année en année, le S/P cible pour l'année 2020 avec ce modèle est donc légèrement supérieur à 100% dans chacun des domaines.

Résultats

Cette section débute par donner les évolutions des ratios $\frac{S}{P}$ dans le tableau 3.2. Dans ce tableau tous les chiffres exprimés sont des pourcentages.

Tout d'abord si 2020 avait été une année classique, nous nous attentions à ce que les évolutions des $\frac{S}{P}$ soient positives puisque les dépenses de santé augmentent d'une année sur l'autre. Ici chaque domaine est négatif, nous pouvons donc déjà en conclure que le modèle issu des données 2019 n'est adapté pour aucun des domaines étudiés. Les évolutions domaine par domaine sont donc étudiées maintenant.

Tout d'abord, au global, le $\frac{S}{P}$ a diminué de 18,6%. Ceci est principalement dû à la diminution des dépenses de santé plus qu'à une augmentation des primes. En effet, les primes ont augmenté de 3,26% alors que les dépenses ont diminué de 15,94%.

Ensuite trois domaines ont subit des diminutions de S/P de plus de 20% : le domaine auxiliaires, le domaine forfait hospitalisation, et celui des montures optiques. Ceci s'explique par le fait que dans ces trois domaines il y a eu une baisse de consommation du nombre d'actes du fait de la pandémie. Pendant celle-ci, les hôpitaux ont repoussé les opérations non urgentes afin d'être en mesure d'accueillir les patients atteint de COVID. De même pour le domaine auxiliaires qui est composé principalement d'actes de kinésithérapie. Pour les montures optique, les opticiens étaient fermés pendant les mois du premier confinement, ce qui explique en partie la baisse de la charge sinistre totale. Cependant, ceci s'explique également par le fait que depuis le 1^{er} janvier 2020, la réforme du 100% santé s'est étendue

TABLE 3.2 : Ratios S/P pour les années 2019 et 2020 en %

Domaine	S/P 2019	S/P 2020	Evolution
Analyses	89,539	78,768	-12,029
Auxiliaires	112,015	87,130	-22,216
Chambre particulières	109,309	87,655	-19,811
Forfait hospitalisation	111,970	69,631	-37,813
Médecine générale	106,914	94,375	-11,728
Médecine spécialiste	96,470	86,631	-10,199
Monture optique	108,671	76,121	-29,953
Pharmacie	93,834	82,183	-12,416
Prothèses dentaire	107,124	94,054	-12,200
Radiologie	104,357	101,877	-2,376
Soins dentaire	101,019	87,956	-12,931
Soins hospitaliers	105,559	87,280	-17,316
Verre optique	108,670	98,279	-9,562
Total	103,903	84,578	-18,599

au domaine de l'optique mettant ainsi en place des prix limites de vente et imposant aux assureurs de ne rembourser au plus que 100€ par monture (contre 150€ auparavant).

Un seul autre domaine se démarque des autres, il s'agit du domaine radiologie pour lequel le $\frac{S}{P}$ a diminué de 2,4%. Ceci s'explique par le fait que ce domaine regroupe divers actes parmi lesquels un acte nommé « Actes d'échographie, Doppler », en effet c'est grâce à cet acte que l'on peut détecter certains effets de la COVID comme des thromboses, de plus les complications pulmonaires suite à la COVID sont détectés par d'autres actes de radiologie. En effet, en regardant l'évolution de la consommation de cet acte précis, nous pouvons observe qu'il a augmenté de plus de 10% entre 2019 et 2020. Cette augmentation compense la diminution du reste des actes du domaine et conduit donc à une diminution moindre du $\frac{S}{P}$.

3.2 Réforme du 100% santé

Nous allons maintenant regarder les premiers résultats de la réforme du 100% santé. Pour cela, l'ensemble des données de la base initiale ont été utilisées. Dans cette étude les montants en euros sur l'ensemble des actes dentaire sont étudiés (en considérant que la répartition des actes est uniforme). S'intéresser aux montants permet de regarder les effets de la réforme sans prendre en compte les effets de la pandémie, celle-ci ayant surtout affectée le nombre d'actes consommés.

3.2.1 Domaines et premiers résultats

La réforme a été succinctement présentée dans le chapitre 1. Elle concerne les domaines du dentaire, de l'audiologie et de l'optique. Cependant les résultats de la réforme dans chacun de ces trois domaines

ne sont pas similaires. Dans la suite de cette section, les premiers résultats domaine par domaine seront présentés (Corlouer, 2021).

Dentaire

Dans la figure 1.15, les évolutions de remboursement en fonction des différents acteurs étaient observables. Le graphique est redonné ci-dessous. Entre le premier trimestre de 2019 et l'année 2020, le montant moyen des actes a augmenté de 2% ()figure 3.5). En parallèle de cela, le reste à charge pour l'assuré a diminué de presque 17%. Ceci semble indiquer que le 100% santé a bien répondu à son objectif premier : offrir des actes dentaires avec un reste à charge nul pour l'assuré.

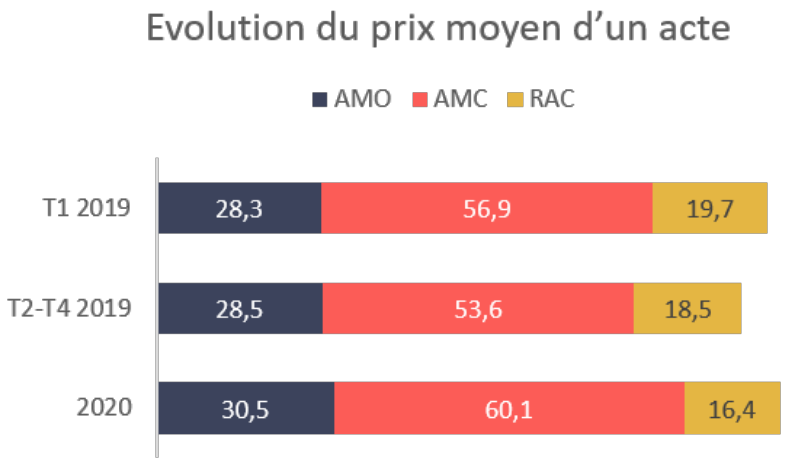
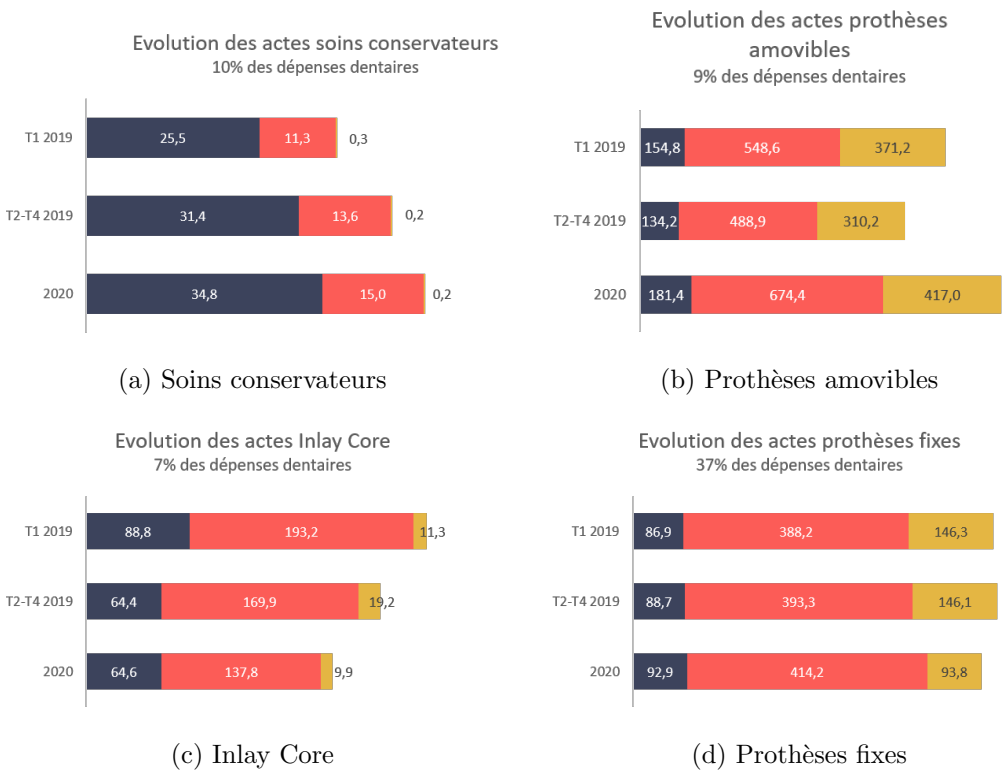


Figure 3.5 : Évolution du montant de remboursement de chaque acteur sur les actes dentaires depuis 2019

Il pourrait également être intéressant de différencier cette même étude en fonction des actes dentaires. En effet, la réforme ne s'applique pas aux soins dentaires courants mais uniquement aux divers prothèses dentaires.En figure 3.6 les mêmes graphiques sont donnés, en différenciant par grande catégorie d'actes.

En regardant les évolutions en fonction des différents types d'actes du domaine dentaire, nous pouvons voir que l'augmentation des frais réels du domaine s'explique grandement par l'augmentation des frais réels pour les soins conservateurs (figure 3.6a). Cette augmentation est importante, environ 35%. Cependant les assurés ne la subissent pas puisque ces soins sont remboursés à 100% par les organismes d'assurance maladie. Ce sont eux qui supportent donc la hausse des prix sur ce type d'actes. A l'inverse, pour les inlay core et les prothèses fixes (figures 3.6c et 3.6d), la réforme et les prix limites de vente instaurés ont conduit à une diminution des frais réels de 28% et de 3%. La réforme a de plus atteint son objectif pour ces actes dans la mesure où le reste à charge moyen pour l'assuré a diminué respectivement de 13% et de 36% pour ces actes.

Enfin cette analyses par types d'actes se conclue par les prothèses amovibles (figure 3.6b). Les frais réels ont augmenté de 18% et le reste à charge moyen pour l'assuré a lui aussi augmenté (de 12%).



(a) Soins conservateurs

(b) Prothèses amovibles

(c) Inlay Core

(d) Prothèses fixes

FIGURE 3.6 : Évolution par grande catégories d'actes du domaine dentaire

Ceci s'explique par le fait que la réforme est rentrée en vigueur pour ce type d'acte le 1^{er} janvier 2021. Notre étude ne portant que sur les actes intervenus pendant les années 2019 et 2020, la réforme n'était pas encore en vigueur pour les prothèses amovibles.

Optique

Dans la figure 10a l'évolution des montants de remboursement des différents acteurs, est également visible, cette fois pour le domaine « Optique ». On remet le graphique ici pour une meilleure visibilité du lecteur.

Sur ce graphique, les frais réels sont étudiés, ils ont diminué de 3.3% alors que dans le même temps, le reste à charge de l'assuré a augmenté de 25%. L'assurance maladie obligatoire a pour sa part, divisé par 15 sa prise en charge moyenne sur le domaine Optique. Dans ce domaine, la conclusion est que la réforme n'a pas atteint son objectif principal puisque le reste à charge moyen pour l'assuré a augmenté alors que la part de l'AMC et de l'AMO a diminué. Les opticiens ont obligation de présenter un équipement avec reste à charge 0 mais il semblerait que ça ne soit pas fait ou pas choisi régulièrement.

Audiologie

Cette section se termine par l'étude du domaine audiologie. Comme précédemment, le graphique de l'évolution des remboursements est redonné (disponible à la figure 10b) ici pour le confort du lecteur.

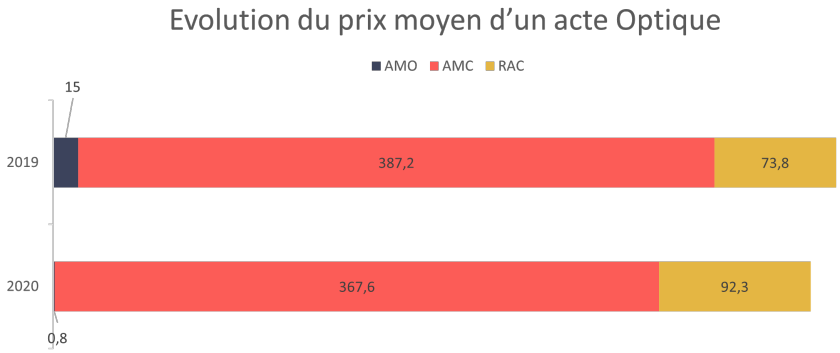


FIGURE 3.7 : Évolution du montant de remboursement de chaque acteur sur les actes optiques depuis 2019

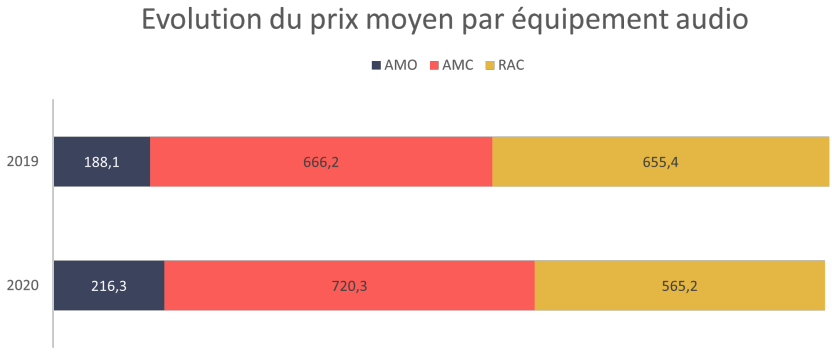


FIGURE 3.8 : Évolution du montant de remboursement de chaque acteur sur les actes optiques depuis 2019

Sur le domaine de l'audiologie, le montant moyen d'un équipement a diminué de 0.5%. Le reste à charge dans ces mêmes actes a pour sa part diminué de 14%. Comme la variation du montant total d'un équipement est très proche de 0, on peut dire que cette baisse du reste à charge est compensée par l'augmentation de la prise en charge de l'AMC et de l'AMO. Le secteur des appareils auditifs est divisé en deux catégories : les appareils de classe I, qui sont ceux concernés par la réforme du 100% santé et les appareils de classe II, qui sont libres. La réforme s'est faite par étape, avec l'instauration d'un prix limite de vente au 1^{er} janvier 2019, qui a été réduit au 1^{er} janvier 2020 puis au 1^{er} janvier 2021. De la même manière, la BRSS sur les appareils de classe I a évolué à la hausse. De plus au 1^{er} janvier 2021, les appareils de classe I doivent être remboursés entièrement par l'AMO et l'AMC. Tout cela explique que sur le graphique, le RAC reste important (565€ en 2020) mais aussi que la part prise en charge par l'AMO et l'AMC a augmenté.

Ceci complète et termine l'étude statistique des premiers résultats de la réforme. Dans la suite, nous analysons les différences par le biais de la tarification de référence, de son application aux individus du portefeuille qui sont présents en 2020 et de l'étude des ratios $\frac{\text{Charge Sinistres}}{\text{Primes}}$.

3.2.2 Analyse de la réforme sur la tarification

Les premiers résultats de la réforme ont donc été évalués domaine par domaine. Nous faisons donc maintenant l'analyse des résultats de la réforme d'après la tarification que nous avons réalisé. Pour cela, les domaines médicaux non-concernés par la réforme ne seront pas regardés.

Le domaine Audiologie ne sera pas non plus étudié dans cette section, le domaine médical comprenant les données liées aux appareils auditifs ayant été mis de côté à cause du manque de données. De plus, les données ne concernent que les années 2019 et 2020 et sur le domaine audiologie, seul un prix limite de vente a été fixé sur cette période. Le remboursement à 100% des prothèses de catégories I n'est entré en vigueur qu'au 1^{er} janvier 2021.

Dentaire

Nous commençons donc l'étude des effets du 100% santé sur le domaine dentaire. Nous avons réalisé la tarification sur deux domaines qui sont liés au dentaire : « Prothèses dentaires » et « Soins dentaires ». Cependant, la réforme ne concerne que les couronnes, les bridges ainsi que les prothèses amovibles (à partir de 2021 pour celles-ci). Tous ces actes sont regroupés dans le domaine « Prothèses dentaires » c'est donc celui-ci que nous allons analyser.

Dans la table 3.2, nous pouvions voir que pour ce domaine, le $\frac{S}{P}$ a diminué de 12,2% entre 2019 et 2020. Ceci est inférieur à la diminution du ratio global qui était de l'ordre de 18%. Nous pouvons donc en conclure que le 100% santé a conduit à une modification de la consommation de soins dans ce domaine. En détaillant l'évolution du coût moyen et de la fréquence de consommation entre les deux années étudiées, on observe une diminution de la fréquence de 13% dans ce domaine (contre 12% de diminution au global) et de 5% de diminution pour le coût moyen (contre une augmentation globale de 4%). Plusieurs conclusions semblent donc importantes à noter, pour ce domaine, la réforme ne semble donc pas avoir modifié la quantité de nombre d'actes consommés pour l'année 2020. Cependant, lorsque une augmentation du remboursement moyen de 4% pour l'ensemble du portefeuille est observée, le

remboursement moyen diminue pour ce domaine de 5%. Ceci peut être dû à la réforme et notamment à la mise en place des prix limites de vente et de la hausse du remboursement de la sécurité sociale.

Optique

Nous allons donc maintenant nous concentrer sur les domaines lié à l'optique, que ce soit les montures ou les verres. Tout d'abord il est remarquable que l'évolution des $\frac{S}{P}$ a été assez différentes entre les deux domaines :

- Une forte baisse pour les montures : près de 30%
- Une baisse plus modérée pour les verres : environ 10%

L'évolution de la fréquence et du coût de chacun des domaines seront analysées séparément. Nous commençons donc par regarder ce qu'il s'est passé sur la fréquence de consommation entre 2019 et 2020 pour les montures et pour les verres optiques. Pour les montures, une diminution de 13% est observable entre les deux années, ce qui est comparable aux prothèses dentaires, et comme il a été vu précédemment à la baisse globale de consommation. La différence de $\frac{S}{P}$ est donc due à une différence d'évolution du coût moyen d'un acte. En effet, en regardant l'évolution des montants moyens des actes, une augmentation de 3% est observée pour les verres optiques mais une diminution de 22% pour les montures.

La forte baisse pour les monture optique s'explique en partie par le changement de plafond de remboursement des montures pour les contrats responsables qui est passé de 150€ à 100€. Ceci conclut l'étude des domaines du 100% santé.

3.3 Conclusion

Les différences entre les deux années on donc été analysées du point de vue d'un évènement que les assureurs pouvaient anticiper comme la réforme du 100% santé ou qu'ils ne pouvaient pas anticiper comme la pandémie de la COVID-19. Pour finir cette étude, nous soulevons quelques questions sur la gestion du futur par les assureurs. Enfin, les limites de l'étude ainsi que les points non abordés seront évoqués.

3.3.1 Préparation du futur

Les assureurs sont impactés par le contexte social. Ils doivent ainsi anticiper dans la mesure du possible ce qui pourrait impacter leur activité.

Un évènement prévisible

La réforme du 100% santé a donc été anticipée par les assureurs. Comme il a été vu dans la section 3.2.1, dans les domaines du dentaire et de l'audiologie la part remboursée par les assureurs a augmenté. Ceci a un coût important qui doit se répercuter sur les primes pures. La mise en place de la réforme est

prévisible, les assureurs savent à l'avance qu'elle sera mise en place, cependant ils doivent en estimer les effets à l'avance afin d'ajuster leurs tarifs. Pour la réforme du 100% santé, cela implique une hausse des remboursements afin de permettre un reste à charge nul, mais dans le même temps, avec la mise en place des prix limites de vente, les remboursements devront diminuer. Ils ont donc quantifier ces variations année par année en fonction des étapes de mise en place de la réforme. Enfin, dans le même temps, des restes à charges nuls sur des actes très onéreux par le passé vont conduire a une augmentation du nombre d'actes consommés dans ces domaines et donc conduire encore une fois à une augmentation de la charge sinistre de l'assureur.

L'analyse de ce qu'il s'est passé dans les domaines ≪ Dentaire ≫, ≪ Monture Optique ≫ et ≪ Verre Optique ≫ a donc été réalisée. Pour le dentaire, le $\frac{S}{P}$ était de 107% pour l'année 2019 et est passé à 94% pour l'année 2020. Pour estimer le montant de prime recueilli, le modèle de l'année 2019 a été appliqué. Dans ce domaine, celui-ci n'est pas adapté à une période de pandémie. Pour le domaine ≪ Monture optique ≫, le $\frac{S}{P}$ a diminué de 30% entre 2019 et 2020. Enfin, le domaine ≪ Verre optique ≫ a vu une diminution de son $\frac{S}{P}$ de 10%. Comme nous pouvons le voir, bien que les assureurs aient anticipé une hausse des dépenses suite à la réforme, les ratios de chacun des domaines ont grandement diminué. Puisque que l'analyse statistique de la section 3.2.1 indique une hausse des montants moyens remboursés par les assureurs (au moins pour le domaine dentaire), la diminution du ratio $\frac{S}{P}$ peut s'expliquer par la pandémie de COVID-19 et donc la diminution du nombre d'actes consommés. Ceci repose sur l'hypothèse forte que la pandémie n'a eu de l'importance que sur la quantité d'actes médicaux consommés et que la réforme n'a affecté que les montants moyens de remboursement.

Ceci termine l'analyse des effets de la réforme. Cependant il apparait que ceux-ci sont difficilement observables dans la mesure où la réforme a été largement mise dans l'ombre du fait de la COVID-19. Il apparait alors une question naturelle : quels auraient été les résultats de la réforme dans un monde non touché par une pandémie mondiale ? En effet, il est légitime de supposer que les résultats auraient été différents. La réforme n'est pas très connue du grand public, ce qui aurait sûrement été différent si les médias n'avaient pas eu à parler d'une pandémie mondiale, ils auraient communiqué sur la mise en place de la réforme. De plus, sans pandémie, il n'y aurait pas eu de confinement et donc une plus grande consommation de soins dans les domaines touchés par le 100% santé.

Un évènement imprévisible

Dans cette section, les effets de la pandémie sont abordés. Comme il a été vu dans la section 3.1.2, une pandémie impacte le système de santé différemment en fonction de la localisation géographique et de l'âge des assurés.

Nous avons également vu qu'une pandémie n'impacte pas tous les domaines médicaux de la même manière. On rappelle que l'évolution des $\frac{S}{P}$ entre 2019 et 2020 est donnée dans la figure 3.2. Sur celle si, on observe le S/P par domaine obtenu en utilisant le modèle de l'année 2019 sur les données de l'année 2019 (assurés et remboursements). Comme le modèle est calibré sur ces données, chacun des $\frac{S}{P}$ doit être proche de 100%. On a également le $\frac{S}{P}$ obtenu en utilisant le modèle issu des données de 2019 et appliqué sur les données de 2020. Les $\frac{S}{P}$ devraient être aux alentours de 103% si le modèle est adapté et les 3% d'augmentation de la CSBM retenus comme hypothèse sont justes et adaptés à l'année 2020. Enfin, l'évolution en pourcentage entre les deux années a également été rajoutée.

Chacun des domaines n'a pas évolué de la même manière suite aux troubles de l'année 2020.

Il faut donc se poser la question de l'anticipation d'une telle pandémie pour les assureurs. Pour cela, il faut se poser la question de comment anticiper. Les assureurs ont réalisés un gain technique de court terme suite à la pandémie. Il pourrait être envisagé par exemple une clause dans le contrat d'assurance qui restitue d'une manière ou d'une autre une partie de la prime si l'état de pandémie est décrété par l'OMS. On pourrait également prendre en compte la probabilité d'occurrence d'une pandémie ainsi que le gain associé et soustraire cela à la prime pure. Certains assureurs ont d'ailleurs mit en place une contribution exceptionnelle (exonération des primes pendant le confinement) lié au COVID dans le cas où l'entreprise avait déclaré du chômage partiel pour ses employé. Ceci confirme le bénéfice réalisé par les assureurs pendant la pandémie d'un point de vue métier.

3.3.2 Sous-module catastrophe du SCR

Un des plus grands enjeux pour les assureurs aujourd'hui, est le calcul du SCR (Solvency requiered capital) dans le cadre de la directive Solvabilité II. En effet, les assureurs doivent être en capacité de répondre à leurs engagements dans 99,5% des cas. Pour cela, l'EIOPA a développé une formule de calcul du SCR dite « formule standard ». Celle ci est une approche modulaire, comme on peut le voir sur la figure 3.9 (PRAS, 2020).

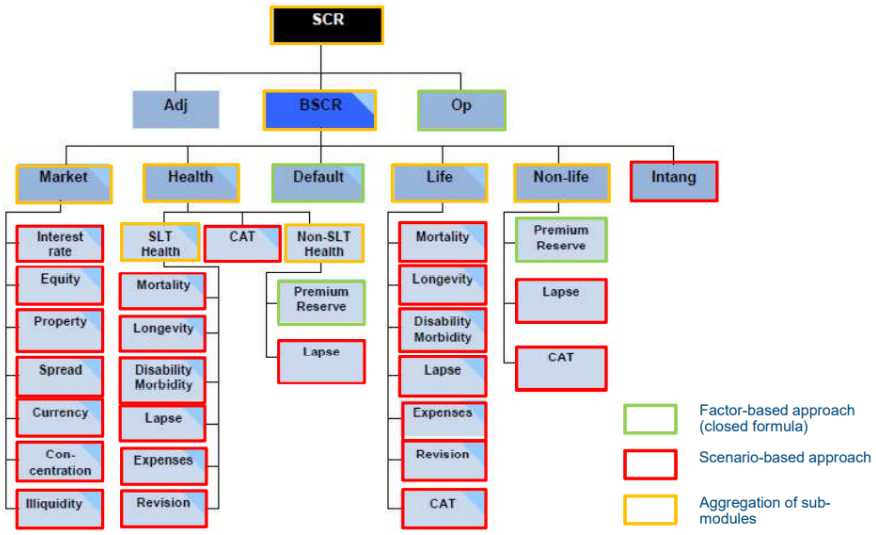


FIGURE 3.9 : Formule standard pour le calcul du SCR

Le risque « santé » est donc calculé par l'agrégation des différents modules qui le compose : Santé SLT, Santé Non SLT et catastrophe. Afin de ne pas être trop restrictif, la formule standard utilise également une matrice de corrélation (fournie par le régulateur) entre ces différents risques. Les différents sous-SCR sont alors agrégés afin d'obtenir le SCR final.

Nous nous intéressons ici au module catastrophe puisqu'il prend notamment en compte le risque pandémique et concerne donc directement la pandémie de COVID-19. L'enjeux de cette courte section

est donc de voir s'il est nécessaire de prendre en compte le risque pandémique du fait du double effet contradictoire vu dans la section 3.1.2. En effet, certes les dépenses de santé augmentent pendant une pandémie sur certains domaines du fait de l'augmentation de la prise en charge des malades touchés par la maladie. Cependant dans le même temps, les différentes mesures gouvernementale ou les éventuels changements de comportement peuvent être un frein à la consommation de soins.

De plus, dans un tel contexte, les soins utilisés pour lutter contre la maladie sont souvent pris en charge en intégralité par la sécurité sociale. De ce fait, une partie de l'augmentation de la quantité d'actes médicaux ne concerne pas les organismes d'assurance complémentaire. Par exemple en France, les différents tests PCR et la vaccination contre la COVID-19 ont été prit en charge à 100% par la sécurité sociale. Cependant, ils ne seront plus pris en charge par la sécurité sociale à partir de octobre 2021, les assureurs complémentaires les prendront alors peut être en charge. Ces dépenses n'affectent donc pas les assureurs et donc la quantité de capital nécessaire à leur bon fonctionnement. Tout ce paragraphe est cependant à nuancer dans la mesure où la sécurité sociale en France est particulièrement protectrice envers la population et que la directive Solvabilité II s'applique aux assureurs de toute l'Europe. Les conclusions ainsi tirées ici ne sont donc valables que pour la France et pour pouvoir obtenir une analyse plus précise, il faudrait faire une étude issue de données provenant de toute l'Europe.

3.3.3 Limites de l'étude

Comme dans le chapitre 2, les limites de l'étude sont étudiées afin d'apporter un regard critique sur celle-ci. Les limites décrites dans ce paragraphe s'ajoutent à celles décrites précédemment.

Différenciation pandémie/réforme

Une première limite de l'étude est que nous avons analysé les impacts de la pandémie via l'évolution du nombre d'actes entre 2019 et 2020. Ceci est une hypothèse forte dans la mesure où il est peu probable que le nombre d'actes n'ait pas été impacté par la mesure du 100% santé. Son but était d'ailleurs d'offrir l'accès à certains soins à des personnes qui se les refusaient pour des raisons financières. A l'inverse, il est également possible que la pandémie de COVID-19 ait eu un impact sur le montant de remboursement des soins. En effet, la consommation de soins a sûrement été différente et donc les remboursements ont pu être différents également. De plus, pour les évolutions constatées au sein des domaines du 100% santé il est difficile de savoir lesquelles sont dûes à la réforme ou à la pandémie. De plus ce sont deux effets contradictoires, le confinement mis en place en réponse à la pandémie va réduire le nombre d'actes consommés quand la réforme a pour but de les augmenter. Il est donc possible que les effets de ces deux évènements s'annulent ou atténuent l'impact de l'autre pour cette étude.

Effets de long terme

Une autre limite de cette étude est qu'elle ne prend pas en compte les effets de long terme. En effet, dans la figure 3.3 il a été montré que le nombre d'actes médicaux consommé pendant le confinement de mars à mai 2020 a drastiquement chuté. De plus dans cette même figure, il est visible que la consommation a augmenté par la suite, c'est un effet de rattrapage. Cependant cette augmentation ne compense pas la diminution des mois de confinement. Nous pouvons donc en déduire que certains soins

n'ont pas été rattrapé et donc que la prévention de certaines pathologie simple n'a pas été réalisée. Ceci peut entrainer une augmentation du nombre de pathologie plus sévère qui seront donc aussi plus couteuse pour les assureurs. Nous pouvons donc supposer que le gain technique réalisé pendant la pandémie de COVID-19 est un gain temporaire qui pourrait même à plus long-terme se transformer en perte technique pour les assureurs.

Conclusion

Dans cette étude, nous nous sommes donc intéressé à l'anticipation pour les assureurs santé de plusieurs évènements qui vont impacter leur activité. Ces évènements peuvent être de nature différentes : prévisibles ou imprévisibles. Ces évènements peuvent donc avoir des répercussions sur la consommation de soins et donc sur les remboursements réalisés par les assureurs. Si les remboursements sont modifiés en fréquence ou en coût par rapport au moment de la tarification, il faut donc ajuster les primes que les assurés paieront sous peine de réaliser une perte technique ou subir une perte de compétitivité qui entrainerait une perte de part de marché.

Il a également été vu que même si la réforme avait pu être anticipée par les assureurs qui ont majoré les primes entre 2019 et 2020 en réponse à la mise en place du 100% santé, ceux-ci se sont heurtés à un évènement plus important : la pandémie de COVID-19. En effet, un pandémie mondiale sans précédent s'est déclarée début 2020 et les états ont du réagir face à celle-ci en empêchant certaines activités non-essentielles dont faisaient partie les dentistes et opticiens notamment. De plus, les populations qui étaient plus réticentes à sortir et donc à s'exposer au virus, ont moins consommé de soins sur l'année 2020. Cette baisse globale de la consommation fait donc diminuer grandement la prime pure pour chaque individu. Ceci soulève donc une question sur la prise en compte de l'année 2020 pour les tarifications futures. Celle-ci doit elle être prise en compte ? Prise en compte mais avec des coefficient de correction ? Simplement écartée des données ? Cela dépendra du choix de la personne en charge de la tarification. D'autres questions peuvent également êtres soulevées comme ce qu'il aurait été possible de faire avec des données plus fines avec par exemple, une vision produit par produit. La modélisation aurait donc été plus fine et aurait peut -être permit de mieux capter certains effets.

Cette diminution de la prime pure réelle peut sembler ne pas être un problème dans la mesure où elle augmente la rentabilité des contrats d'assurance et donc les assureurs ne réalisent pas de perte technique, ils ne perdent pas non plus de part de marché dans la mesure où aucun autre assureur pouvait prévoir la pandémie et donc aucun d'entre eux n'a pu anticiper une baisse des cotisations. Cependant elle pose un problème dans la mesure où les assureurs doivent présenter un tarif à leurs assurés pour les années post-covid. En effet, il faudra alors faire attention à ne pas être biaisé par les années troublées par la pandémie tout en intégrant les coûts issus du 100% santé qui aura alors atteint son rythme de croisière et donc les premiers résultats sont biaisés par la modification de la consommation dûe à la COVID-19.

Ceci pose donc plusieurs questions, une fois la situation sanitaire revenue à la normale. Quelle va être la consommation des assurés ? Les assurés vont-ils plus profiter du 100% santé et donc encore augmenter la quantité de remboursements des organismes complémentaire d'assurance maladie. Enfin et surtout, cela pose une question sur le manque de soins pendant la pandémie. En effet, certaines prises

en charge préventives considérées comme moins urgentes ont été reportées ou annulées et donc des pathologies plus sérieuse n'ont pas pu être détectée où se sont aggravées et seront donc plus coûteuses à soigner et donc à rembourser. Il pourrait donc être intéressant de quantifier cette variation à plus long terme.

Bibliographie

1957-1958 Pandemic (H2N2 virus) (2019). URL : https://www.cdc.gov/flu/pandemic-resources/1957-1958-pandemic.html (visité le 14/12/2021).

Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost (oct. 2020). URL : https://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost (visité le 23/07/2021).

Assurance maladie complémentaire (avr. 2021). Page Version ID: 182324552. URL : https://fr.wikipedia.org/w/index.php?title=Assurance_maladie_compl%C3%A9mentaire&oldid=182324552 (visité le 12/05/2021).

Boîte noire (système) (mai 2021). Page Version ID: 182894045. URL : https://fr.wikipedia.org/w/index.php?title=Bo%C3%AEte_noire_(syst%C3%A8me)&oldid=182894045 (visité le 20/09/2021).

Bradley Efron, T. H. (2021). Computer Age Statistical Inference. Cambridge University Press.

Corlouer, A. (mai 2021). Covid-19 : une baisse différenciée des dépenses de santé | SeaBird. Section: On en parle. URL : https://www.seabirdconseil.com/nos-decryptages/on-en-parle/covid-19-des-impacts-differencies-sur-les-depenses-de-sante/ (visité le 12/08/2021).

De Castex, E. (2019). Le problème de la boîte noire : pour faire confiance aux algorithmes, faut-il les comprendre ? URL : https://www.anthropotechnie.com/le-probleme-de-la-boite-noire-faut-il-comprendre-les-algorithmes-pour-leur-faire-confiance/ (visité le 14/12/2021).

Delignette-Muller, M. L. et Dutang, C. (2015). Journal of Statistical Software. *Journal of Statistical Software* 64.4.

Donnet, S. (2019). Cours de modèle linéaire, Université Paris-Dauphine - PSL.

Graves, E. (2017). Package 'onehot'.

Grippe asiatique (mai 2021). Page Version ID: 182932930. URL : https://fr.wikipedia.org/w/index.php?title=Grippe_asiatique&oldid=182932930 (visité le 21/05/2021).

Lazic, S. (2020). Cours Santé - Prévoyance, Université Paris-Dauphine - PSL.

Le marché de la santé et de la prévoyance progresse de 2,8 % en 2018 (2019). URL : https://www.ffa-assurance.fr/etudes-et-chiffres-cles/le-marche-de-la-sante-et-de-la-prevoyance-progresse-de-28-en-2018.

Les branches de la sécurité sociale (2021). URL : https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-branches (visité le 15/09/2021).

Les régimes (2021). URL : https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-regimes (visité le 28/06/2021).

Les régimes de la sécurité sociale (s. d.). URL : https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-regimes (visité le 15/09/2021).

MANDHOUJ, K. (2010). Analyse du Risque Catastrophe d'une Pandémie en Assurance Prévoyance par une Approche Épidémiologique.

PALIS, M. (2021). Impact de la réforme du « 100% Santé » sur les contrats individuels et collectifs de complémentaires santé.

PRAS, I. (2020). Solvabilité II - Cours, Université Paris-Dauphine - PSL.

Réforme 100 % santé (avr. 2021). Page Version ID: 181684383. URL : https://fr.wikipedia.org/w/index.php?title=R%C3%A9forme_100_%25_sant%C3%A9&oldid=181684383 (visité le 17/05/2021).

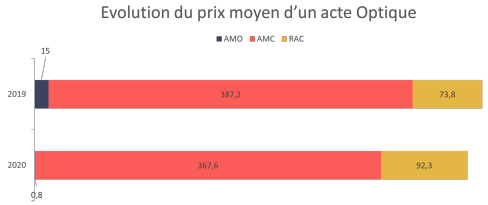
Sécurité Sociale 2021 (2021). URL : https://www.ccomptes.fr/fr/publications/securite-sociale-2021 (visité le 18/12/2021).

Trevor Hastie Robert Tibshirani, J. F. (2009). The Elements of Statistical Learning. Springer, New York, NY.

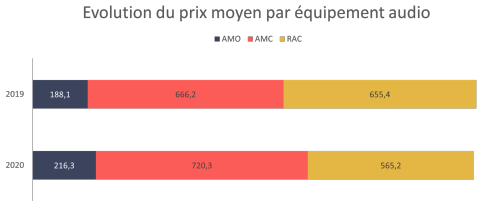
Villarmé, L. (2020). Tarification d'une complémentaire santé et impact de la réforme 100% santé.

Annexe

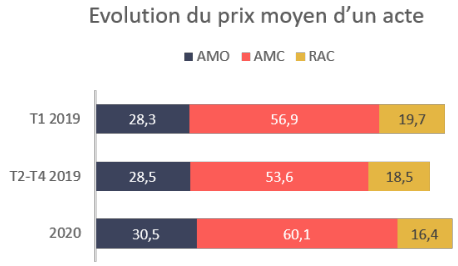
Résultats du 100% santé



(a) Évolution du montant de remboursement de chaque acteur sur les actes optiques depuis 2019

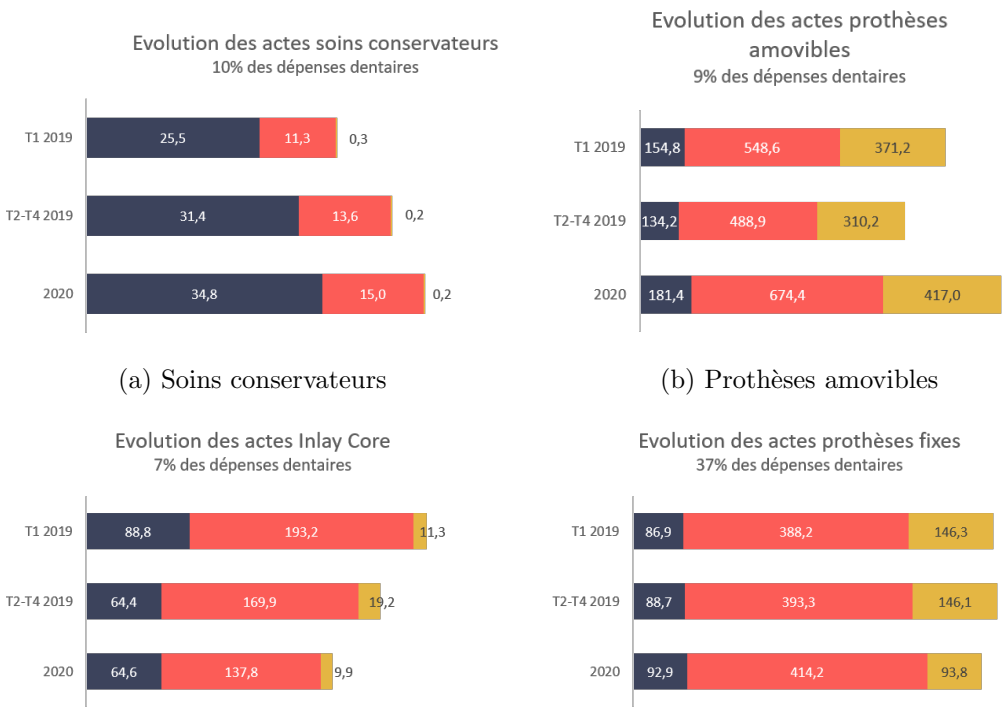


(b) Évolution du montant de remboursement de chaque acteur sur les actes optiques depuis 2019



(c) Évolution du montant de remboursement de chaque acteur sur les actes optiques depuis 2019

FIGURE 10 : Evolution des répartition des remboursements sur les domaines du 100% santé



(a) Soins conservateurs

(b) Prothèses amovibles

(c) Inlay Core

(d) Prothèses fixes

FIGURE 11 : Évolution par grande catégories d'actes du domaine dentaire

La loi de Poisson appartient à la famille exponentielle

Nous montrons que la loi de Poisson appartient à la famille exponentielle. La fonction de masse de la loi de poisson de paramètre λ est :

$$f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Nous montrons maintenant que cette fonction appartient à la famille exponentielle.

$$\begin{aligned} f(k) &= \exp\{-\lambda\} \times \exp\left\{\ln\left(\frac{\lambda^k}{k!}\right)\right\} \\ &= \exp\{-\lambda + k\ln(\lambda) - \ln(k!)\} \end{aligned}$$

En posant : $\theta = \ln(\lambda)$. L'expression devient :

$$f(k) = \exp\{-\exp\{\theta\} + k\theta - \ln(k!)\}$$

En posant

$$\gamma(\phi) = 1, \text{ et } b(\theta) = \exp\{\theta\}, \text{ et } c(k,\phi) = -\ln(k!)$$

Alors la loi de Poisson appartient bien à la famille exponentielle.

□

Choix des lois par domaine pour le GLM

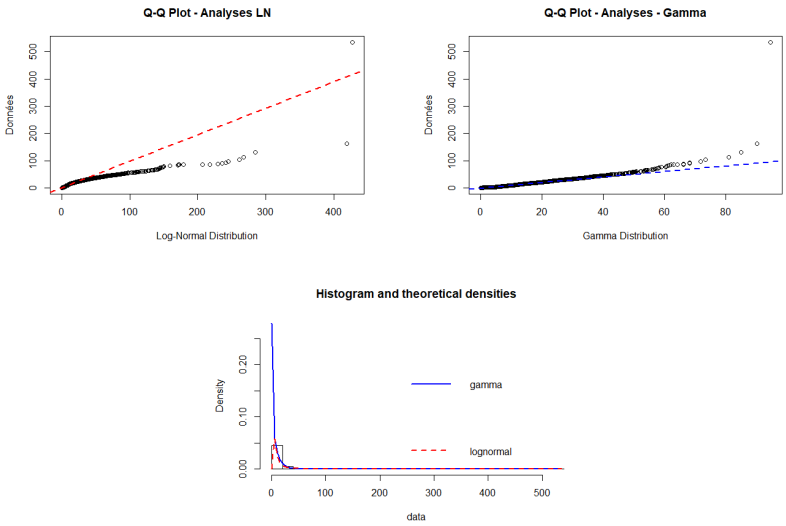


FIGURE 12 : Choix de la loi d'adéquation pour le domaine Analyses

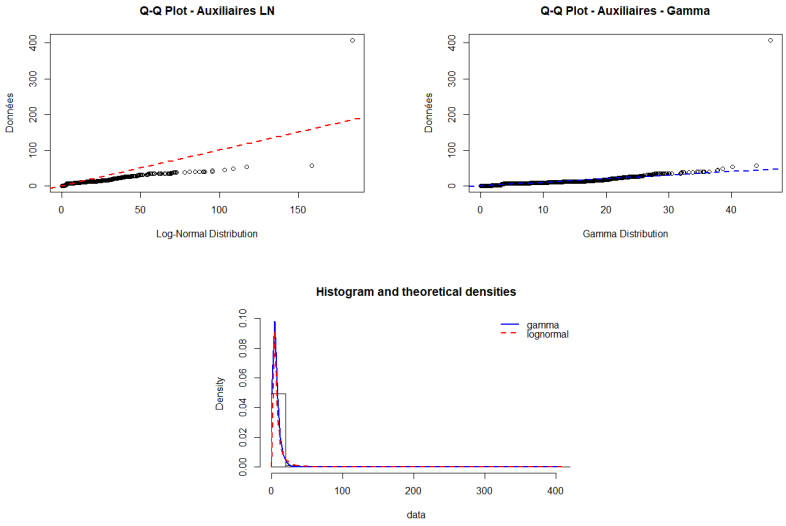


FIGURE 13 : Choix de la loi d'adéquation pour le domaine Auxiliaires

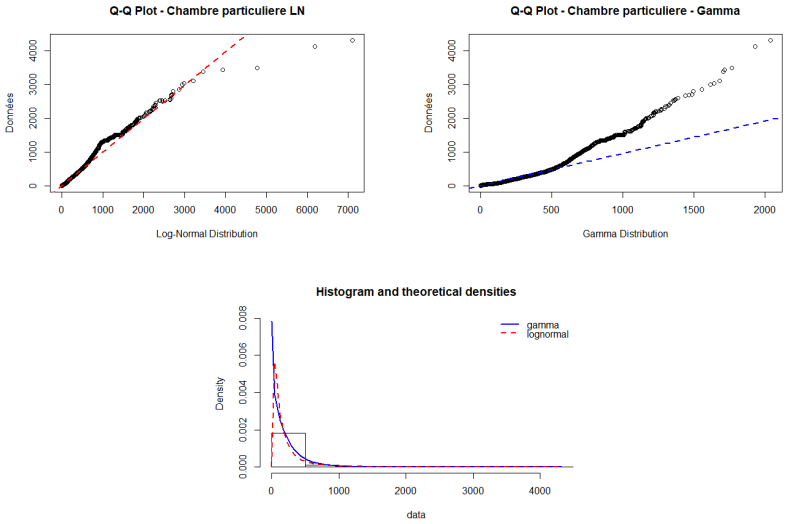


FIGURE 14 : Choix de la loi d'adéquation pour le domaine Chambres particulière

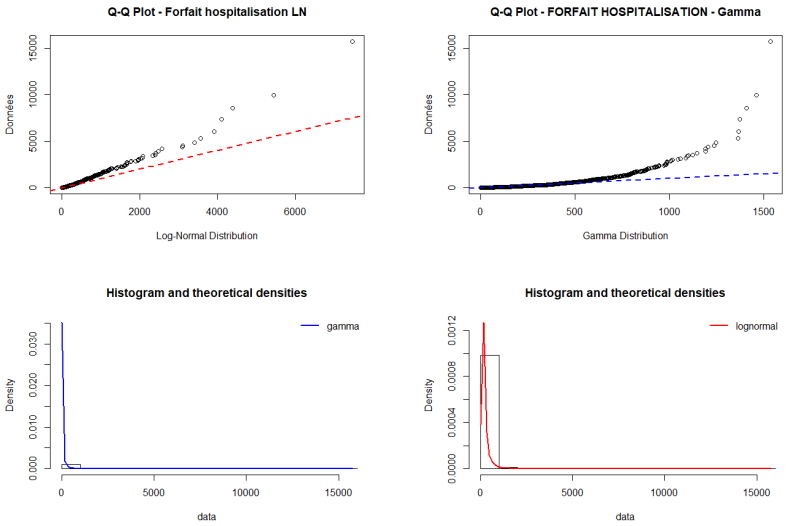


FIGURE 15 : Choix de la loi d'adéquation pour le domaine Forfait hospitalier

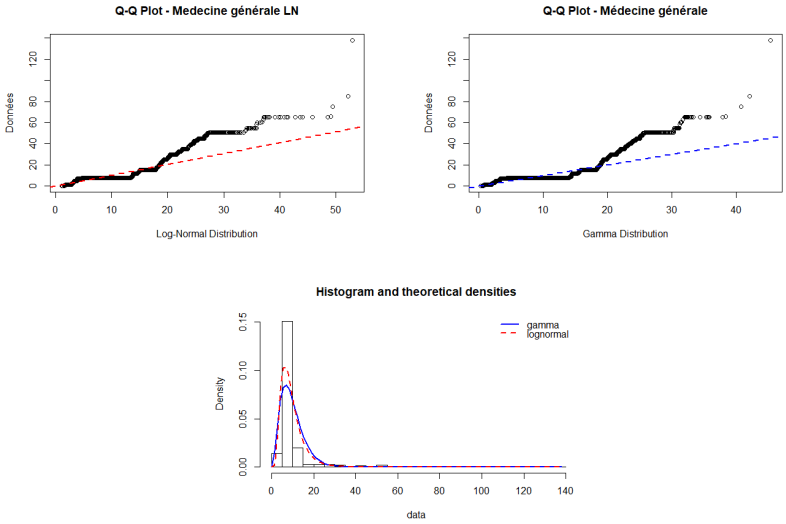


FIGURE 16 : Choix de la loi d'adéquation pour le domaine Médecine générale

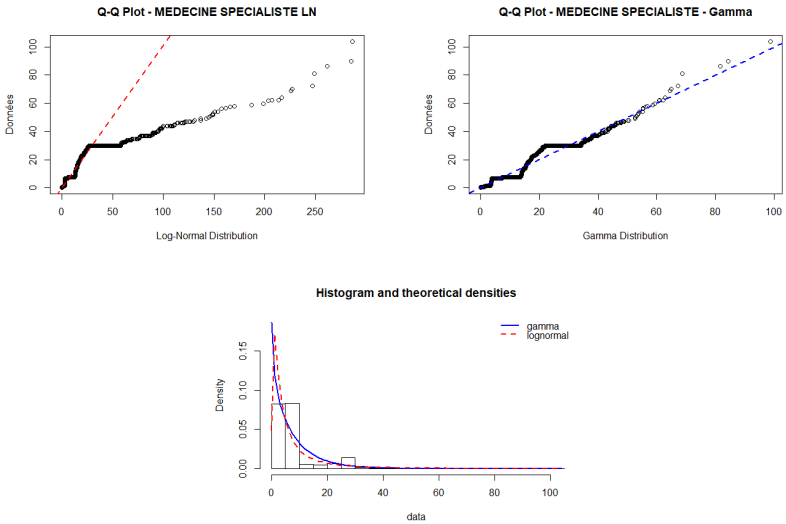


FIGURE 17 : Choix de la loi d'adéquation pour le domaine médecine spécialiste

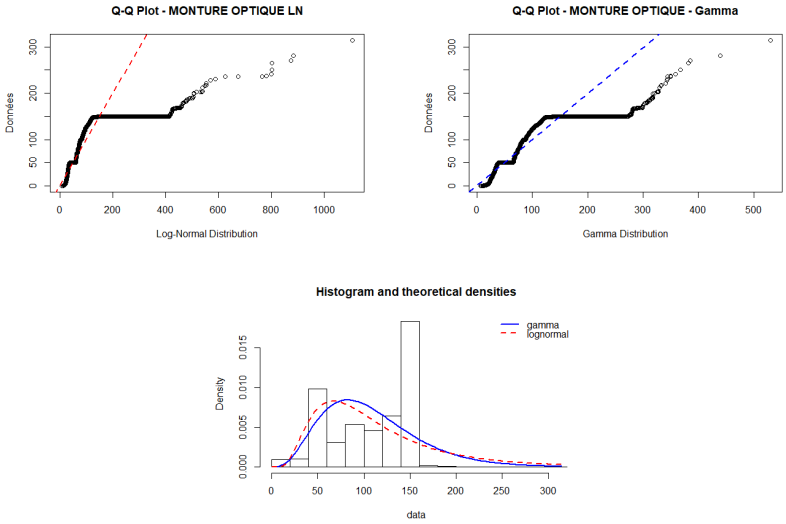


FIGURE 18 : Choix de la loi d'adéquation pour le domaine Monture optique

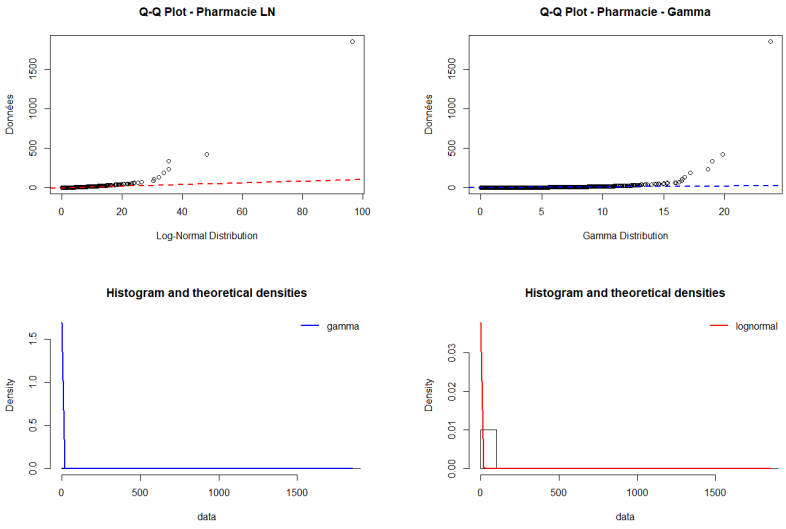


FIGURE 19 : Choix de la loi d'adéquation pour le domaine pharmacie

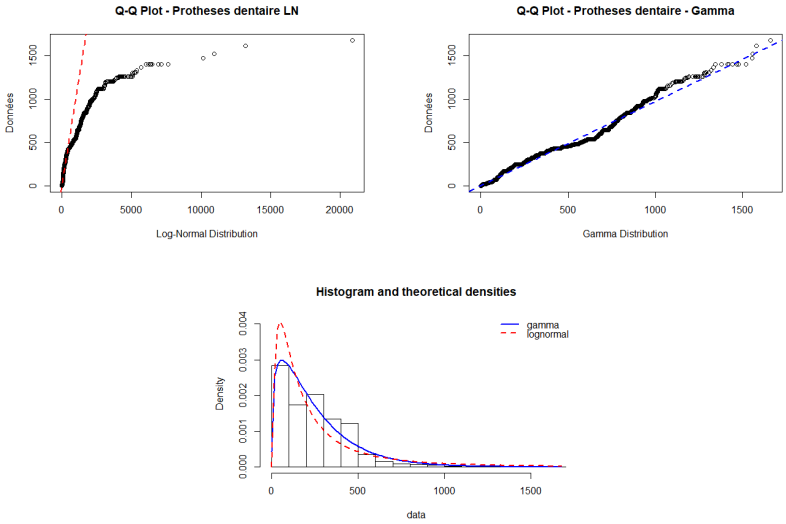


FIGURE 20 : Choix de la loi d'adéquation pour le domaine pharmacie

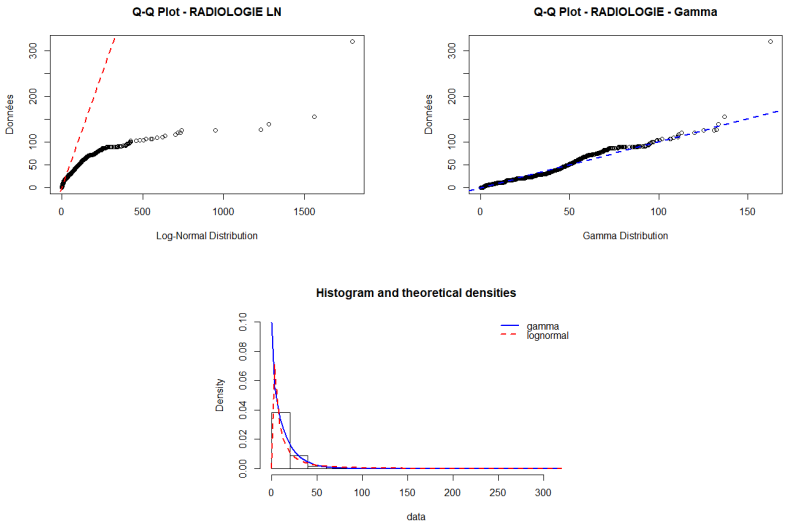


FIGURE 21 : Choix de la loi d'adéquation pour le domaine radiologie

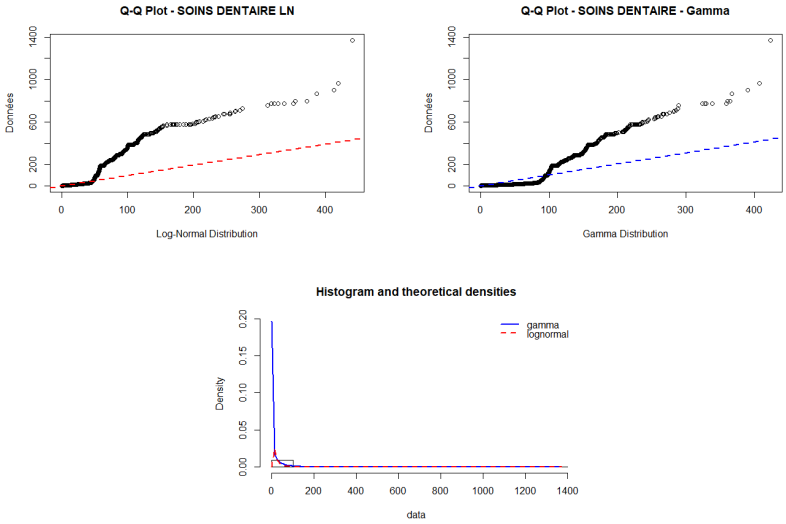


FIGURE 22 : Choix de la loi d'adéquation pour le domaine soins dentaire

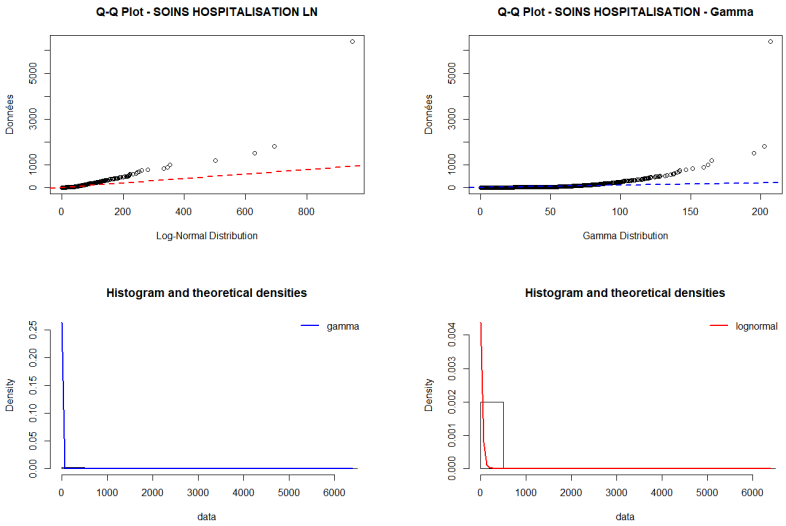


FIGURE 23 : Choix de la loi d'adéquation pour le domaine soins hospitaliers

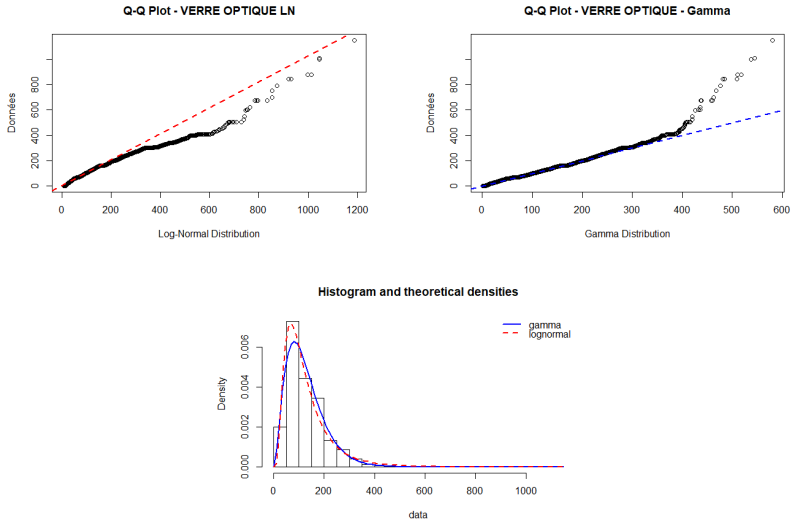


FIGURE 24 : Choix de la loi d'adéquation pour le domaine verre optique

Abréviations utilisées

Abréviation	**Nom complet**
ACP	Analyse en composante principale
AMC	Assurance maladie complémentaire
AMO	Assurance maladie obligatoire
Ana	Analyses
Aux	Auxiliaires
BRSS	Base de remboursement de la sécurité sociale
CAH	Classification ascendante hiérarchique
CART	Classification and regression tree
CDC	Centers for diseases control
CP	Chambre particuliere
CSBM	Consommation de soins et biens médicaux
CSP	Catégorie socio-professionnelle
DH	Dépassements d'honoraires
FH	Forfait Hospitalier
FR	Frais réels
GLM	Generalized linear model (modèles linéaires généralisé)
GM	Gérant majoritaire
iid	Indépendant et identiquement distribué
IP	Institution de prévoyance
MG	Médecine Générale
MS	Médecine Spécialiste
MO	Monture Optique
NAF	Nomenclature d'activités française
OHE	One Hot Encoding
OMS	Organisation Mondiale de la Santé
OPTAM	Option pratique tarifaire maîtrisée
Pha	Pharmacie

Abréviation	**Nom complet**
PrD	Prothèses dentaire
Rad	Radiologie
RMSE	Root mean square error
RSS	remboursement de la sécurité sociale
RAC	Reste à charge
RC	remboursement complémentaire
SA	Sociétés d'assurance
SCR	Solvency requiered capital
SD	Soins dentaire
SH	Soins hospitaliers
S/P	Ratio charge sinistre sur ensemble de primes
SS	Sécurité Sociale
TM	Ticket modérateur
TNS	Travailleur non salarié
VO	Verres optique
XGB	extreme gradient boosting

Table des figures

Liste des tableaux