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 27 juin 2022

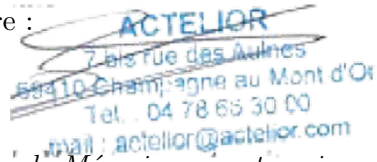
Par : Solange BOYER
Titre : Intégration de la base Open Damir dans des analyses techniques santé

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 : Actélior
Signature :



Directeur de Mémoire en entreprise :
Nom : Elodie PAGET
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

La transformation digitale permet la collecte d'une quantité grandissante de données dans divers secteurs d'activité. Certaines de ces données sont désormais disponibles en Open Data, c'est-à-dire en accès libre et gratuit. Depuis 2015, les bases Open Damir, reprenant l'ensemble des remboursements effectués par l'Assurance Maladie Obligatoire, tous régimes confondus, sont publiées régulièrement. De prime abord, ces bases constituent une source inestimable d'informations pour les différents acteurs. Toutefois, leurs volumétries complexifient l'exploitation et la compréhension des données.

L'intégration de cette base de données au sein des outils pré existants du cabinet de conseil en actuariat Actélior, sont les enjeux de ce mémoire. Comprendre, analyser, exploiter et intégrer les bases de données Open Damir, non utilisées par Actélior avant la rédaction de ce mémoire, en sont les objectifs. Les deux outils sur lesquels portent les travaux sont le suivi technique santé et l'outil de tarification santé.

Après avoir rappelé le fonctionnement du système français de l'assurance santé, nous présentons les bases de données Open Damir et les traitements réalisés via $\mathbb{P}ython$ pour rendre les bases exploitables. Puis, nous détaillons les différentes étapes de la construction des deux outils incluant les bases de données Open Damir traitées. Tout d'abord, le suivi technique Damir permet d'ajouter aux analyses actuelles une dimension nationale permettant aux clients de positionner le niveau de prestations santé de leur portefeuille par rapport aux données nationales réelles. Quant à la tarification santé Damir, l'outil est réalisé par le biais d'une modélisation « Coût × Fréquence », implémentée à l'aide de Modèles Linéaires Généralisés sur *R*. Finalement, les deux tarifications santé, Actélior et Damir, sont liées à partir d'un modèle de crédibilité adapté au cadre de l'étude. Cette méthode a pour objectif d'établir un tarif plus ajusté en prenant en compte des données externes riches en information.

Mots-clés : *Open Damir*, *Ameli*, *MLG*, *crédibilité*, *tarification santé*, *Machine Learning*.

Abstract

Digital transformation is leading to increasing volumes of data collection in various sectors of activity. Some of this data, called Open Data, is now freely accessible. Since 2015, the Open Damir database, which include all healthcare reimbursements made by the French National Health Insurance System, has been published regularly. At first sight, this database is an invaluable source of information for all stakeholders. However, its volumes makes it complex to handle and understand the data.

This database has never been studied nor used by the actuarial consulting company Actélior so far. Integrating it within the pre-existing tools of Actélior is the subject of this memoir. More precisely, the objectives are first to understand and analyse the Open Damir database, then to use and integrate it into two of Actélior's actuarial tools : the health technical risk monitoring tool, and the health risk pricing tool.

After describing how the French National Health Insurance System works, we introduce the Open Damir database as well as the extraction process for our study, carried out via $\mathbb{P}ython$. We then detail the different steps involved in the enrichment of the existing Actélior tools with the processed Open Damir database. As for the health technical risk monitoring tool, adding the Damir technical monitoring opens up a whole new level of analysis by enabling clients to position the level of health consumption in their portfolio in relation to actual national data. As for the health risk pricing tool, the aim is to determine a better-adjusted pricing by taking into account relevant, recent and sizeable external data. This is done by, first, modelling Damir data using a "Cost x Frequency" approach implemented using Generalized Linear Models on R, and then integrating it within the pre-existing Actélior pricing tool using a credibility model developed specifically for this purpose.

Keywords: *Open Damir*, *Ameli*, *GLM*, *credibility theory*, *healthcare pricing*, *Machine Learning*.

Note de Synthèse

Contexte

L'assurance santé occupe une place importante dans le secteur de l'assurance. En effet, la part des dépenses liées aux soins de santé parmi les dépenses nationales est élevée. D'après la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES), ces dépenses santé 2019 s'élevaient à 208 milliards d'euros, ce qui représente 11,3% de la richesse nationale. La prise en charge des dépenses santé est assurée majoritairement par l'Assurance Maladie (part de 78,2% en 2019). Les dépenses restantes sont ensuite prises en charge par différents organismes complémentaires comme les mutuelles ou les institutions de prévoyance, ou bien en dernier recours par les ménages. Pour assurer la continuité de cette prise en charge, ces organismes doivent maitriser le risque santé de leur portefeuille d'assurés, notamment en analysant leur consommation santé. Dans ce contexte, l'utilisation de données externes permettrait d'enrichir ces analyses et d'affiner la maitrise de ce risque. Aujourd'hui, le principal enjeu des organismes complémentaires consiste donc à utiliser des données nationales afin de positionner la consommation santé de leur portefeuille par rapport à la consommation santé nationale.

Depuis quelques années, le système de santé tend à se digitaliser, entrainant alors la collecte de données de plus en plus riches. De nombreuses bases de données santés numériques constituent alors des sources d'informations non négligeables à destination des organismes assureurs, des professionnels de santé, mais aussi pour les personnes souhaitant réaliser des analyses sur le marché de la santé. Depuis 2015, certaines de ces bases numériques sont accessibles par tous et gratuitement. Il s'agit de données Open-Data. La mise en place de ces données Open Data a été possible en étroite lien avec le Règlement de la Protection des Données (RGPD).

Objectifs

Dans cette logique de transformation digitale, le cabinet de conseil en actuariat Actélior souhaite proposer à ses clients une analyse du risque santé complète en intégrant une vision nationale aux analyses préexistantes à l'échelle du portefeuille d'assuré. Pour cela, Actélior a pour objectif de faire évoluer ses solutions d'analyses techniques du risque santé (outil de tarification et suivi technique santé) et de les enrichir de données santé nationales telles que les bases Open Damir.

Ce mémoire vise donc à exploiter les bases volumineuses Open Damir, afin de les intégrer aux outils de tarification de contrat complémentaire santé et d'analyse du risque santé d'Actélior.

Données

Les bases Open-Data des Dépenses d'Assurance Maladie Inter Régimes (DAMIR) sont des bases regroupant toutes les prestations prises en charge par la Sécurité Sociale. Elles sont anonymisées afin de préserver l'identité des bénéficiaires et des professionnels de santé.

L'idée a été dans un premier temps de traiter ce grand volume de données afin de constituer une base de données finale exploitable. Cette étape est l'une des plus importantes. Sans traitement préalable, l'intégration de ces données dans les outils de gestion du risque serait impossible. En effet, dans le cadre de ce mémoire, seules les bases de données des années 2018 et 2019 ont été étudiées, ce qui représente au total 806 734 365 lignes. Les hypothèses prises pour les divers traitements réalisés afin de réduire ces bases seront explicitées au sein de ce mémoire.

Premièrement, de nombreuses variables, considérées comme non utiles pour la suite de l'étude, ont été supprimées. Puis, l'anonymisation des bases Open Damir a dû être étudiée. En effet, les données des bases Open Damir ont été agrégées afin de préserver l'anonymat des bénéficiaires et des professionnels de santé. Une ligne de la base Damir représente alors un nombre N d'actes santé réalisés pour des individus de profil similaire. De plus, des regroupements par tranche d'âge et par région ont été réalisés préalablement sur ces bases. Le nombre de bénéficiaires associé à chaque ligne est donc inconnu. Or, cette information est nécessaire pour la réalisation d'une modélisation sur la fréquence définie par

$$Frequence_i = \frac{Nombre\ d\ actes_i}{Nombre\ de\ personnes\ concernees_i},$$

avec i représentant une ligne de la base Open Damir et donc un profil caractéristique donné.

Cette information a été reconstituée à partir de données démographiques de l'INSEE. Ces données sont réparties selon l'année de soin, du sexe, de la tranche d'âge et de la région. Pour ne pas considérer l'ensemble de la population française (hypothèse forte), elles ont été par la suite couplées aux chiffres clés du pourcentage de personnes couvertes par l'Assurance Maladie Obligatoire.

Enfin, une table de correspondance a été construite afin de lier le référentiel de tarification santé d'Actélior avec les codes actes santé présents dans la base Open Damir. De nombreuses hypothèses, appuyées d'analyses chiffrées, ont été émises pour la construction de cette table, jointe ensuite à la base de données traitée.

Des correctifs ont été appliqués sur les anomalies présentes dans le calcul des montants de remboursement, de dépassement et de la dépense. Une provision a également été additionnée à ces montants. Elle permet de prendre en compte les prestations réalisées en 2019 et remboursées en 2020 qui ne sont donc pas présente dans nos bases de remboursements Open Damir dans le cadre de l'étude (seules les bases de remboursements des années 2018 et 2019 ont été choisies).

Finalement, l'ensemble des données ont été agrégées pour obtenir une base de données dont la volumétrie a été réduite de 80% (soit 10 040 857 de lignes).

Méthodes

Deux outils d'analyses techniques santé d'Actélior ont été alimentés de ces données finales :

- La tarification de contrat complémentaire santé avec la base Damir, jointe à la tarification santé déjà existante d'Actélior grâce au principe de crédibilité
- Le suivi technique santé d'Actélior avec l'ajout d'une vision nationale

Le schéma 1 résume les différents utilisations et liens effectuées entre les travaux de ce mémoire et du cabinet de conseil Actélior.

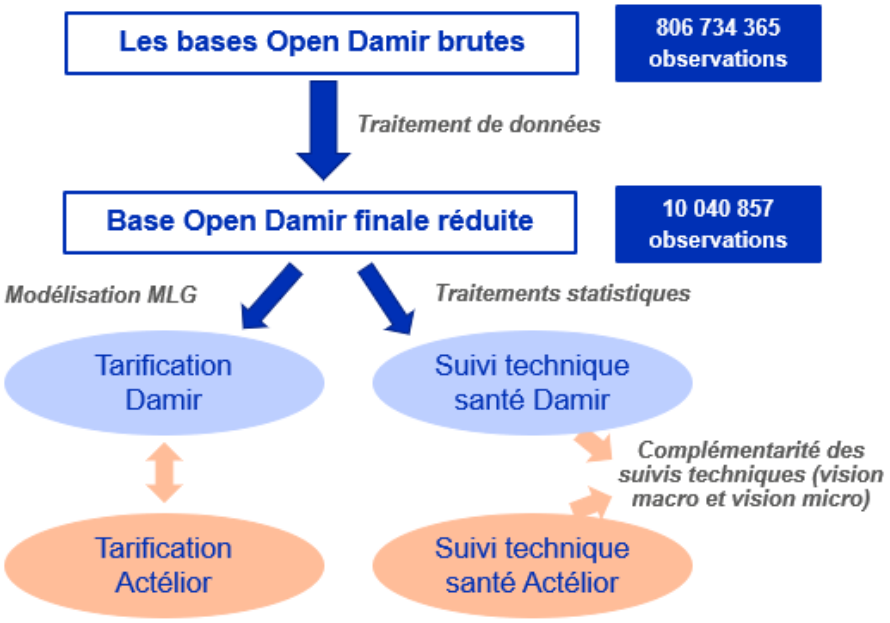


FIGURE 1 : Schéma récapitulatif des travaux effectués à partir des bases Open Damir.

En premier lieu, il est important d'établir une tarification juste des contrats complémentaires santé pour éviter à l'organisme complémentaire de subir des pertes. L'objectif de ce mémoire consiste donc à ajuster la tarification santé d'Actélior avec une tarification modélisée sur les données nationales Open Damir.

La première étape consiste à construire la tarification Damir selon la méthode Coût moyen $\times$ Fréquence. Dans le cadre de cette étude, le coût moyen représente la dépense moyenne (les frais réels des actes de soin). Ces deux variables d'intérêts seront modélisées par un Modèle Linéaire Généralisé. Cette même méthode alimente le moteur de la tarification santé actuelle d'Actélior. Le choix d'une méthode similaire permettra plus aisément de lier ces deux tarifications, et de les comparer à un même niveau de granularité.

La deuxième étape consiste à utiliser le principe de crédibilité pour ajuster la tarification santé d'Actélior grâce à la tarification Damir construite sur des données de dépenses nationales de santé. Pour un libellé brochure donné, un facteur de crédibilité est associé aux deux tarifications. La définition de ce facteur crédibilité est toutefois différente de la définition théorique initiale. Dans le cadre de cette étude, le facteur de crédibilité est défini selon la précision des coefficients GLM obtenus pour chaque variable explicative de la tarification Damir. Si la taille des intervalles de confiance de ces coefficients

GLM est très importante, alors, pour un libellé brochure donné, la tarification Damir influencera très peu la tarification Actélior. Un intervalle de confiance est défini comme très grand lorsqu' il est considéré comme outlier parmi les autres intervalles de confiance. En d'autres termes, le facteur de crédibilité est défini tel que :

$$\mathbf{1 - \alpha_{i,j}} = \frac{Nombre\ de\ variables\ considérées\ comme\ non\ outliers\ dans\ le\ modèle\ j}{Nombre\ de\ variables\ utilisées\ dans\ le\ modèle\ j}$$

où $(1 - \alpha_{i,j})$ désignent respectivement le coefficient de crédibilité associé à la prime pure du libellé brochure n^o i et pour la modélisation de la variable cible j, obtenue avec la tarification Damir.

Le troisième et dernier objectif de ce mémoire consiste à intégrer au suivi technique santé d'Actélior les données de la base Open Damir retraitée en première partie. Cet outil d'analyse technique permet aux clients d'Actélior de connaitre l'évolution du risque santé de leur portefeuille d'une année à l'autre. A ce jour, cet outil d'analyse technique santé permet uniquement aux clients d'Actélior d'avoir une vision de l'évolution la charge (montant de remboursement) et de la consommation des prestations santé à l'échelle de leur portefeuille d'adhérents. Or, une augmentation importante de la consommation santé au sein du portefeuille du client peut être mis en avant, mais pour autant, est-ce cohérent par rapport à la tendance nationale ? C'est alors dans ce contexte que les bases Open Damir ont été intégrées dans cet outil d'analyse technique santé afin d'apporter une vision nationale de la consommation santé. Ce suivi technique santé Damir est construit de la même manière que le suivi technique santé actuel d'Actélior. Les analyses vont d'une maille macroscopique (analyse globale, analyse par bénéficiaire) puis seront détaillées à une maille plus fine (analyses par type de bénéficiaire, par famille d'actes, par type d'actes). Enfin, une analyse supplémentaire a été construite sur les trois postes santés impactés par la réforme 100% santé afin de comprendre et d'analyser l'évolution de la consommation nationale en soins dentaires, optiques et d'audiologie. Cet outil a été construit afin d'automatiser les futurs traitements. Seule la base finale retraitée Damir doit être mise à jour (table de correspondance par exemple). Le tableau de bord du suivi technique sera alors mis à jour automatiquement.

Principaux résultats

Ce mémoire intègre diverses méthodes et outils avec pour objectif l'optimisation, l'accélération et l'automatisation de différents processus essentiels lors de la construction d'une tarification.

En tarification santé, les deux variables cibles sont modélisées pour chaque libellé brochure du référentiel de tarification défini préalablement. Pour chaque modélisation, les variables disponibles au sein des données font alors l'objet d'une sélection par le biais d'une méthode de Gradient Boosting : la méthode d'apprentissage non supervisé Light GBM. Il s'agit d'une méthode adaptée aux bases de données volumineuses, pour laquelle les résultats sont obtenus avec une précision similaire à d'autres méthodes de Gradient Boosting, mais avec un temps d'exécution réduit. L'efficacité de cette méthode est vérifiée dans le cadre de ce mémoire, avec un temps d'exécution pour le LGBM très largement inférieur à celui du Gradient Boosting et avec une précision équivalente (c.f. tableau 3.2).

TABLE 1 : Performances des méthodes de Machine Learning XGBoost et LightGBM.

Dépense moyenne			
	Score - MSE	Score - RMSE	Temps d'exécution
XGBoost	66.68	8.17	00:35:54
LightGBM	46.66	6.83	00:01:36
Quantité d'actes			
	Score - MSE	Score - RMSE	Temps d'exécution
XGBoost	9917.50	99.59	00:34:28
LightGBM	10000.35	100.00	00:01:21

Finalement, les variables retenues sont l'âge du bénéficiaire (AGE_BENEF), la région du bénéficiaire (REGION_BENEF) et l'année de survenance du soin (SOI_ANN).

De plus, l'utilisation de packages R spécifiques permet de détecter automatiquement la loi d'ajustement la plus adaptée parmi les lois possibles (Lois Gamma, Weibull et Lognormale pour la dépense moyenne, lois Binomiale négative et Poisson pour la quantité d'actes).

Dans le cadre de ce mémoire et de la construction de la tarification Damir, seuls les actes d'anesthésie sont tarifiés et seuls les résultats de ce libellé brochure sont exposés. Après exécution des différents traitements, le Modèle Linéaire Généralisé optimal obtenu est satisfaisant, pour lequel l'ensemble des variables explicatives sont significatives et les hypothèses des résidus vérifiées (c.f. graphique 3.22).

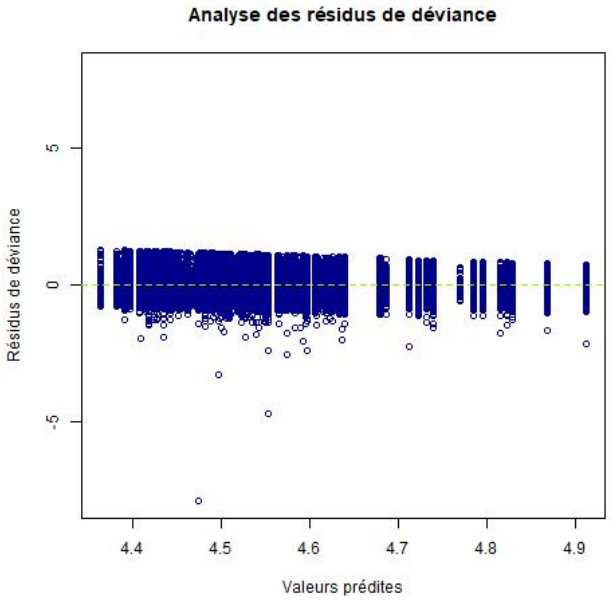


FIGURE 2 : Graphique de vérification de l'hypothèse de linéarité des résidus de Déviance.

Le tarif est alors calculé pour un bénéficiaire de référence, défini comme une personne de 40 ans habitant en Ile-de-France, à partir des coefficients GLM et d'autres paramètres extraits. Dans le cadre de la tarification des actes d'anesthésie, la tarification Damir s'élève à 0,00917€. La tarification Actélior s'élève quant à elle à 0,001€pour ce type d'actes, un tarif proche de celui de la tarification Damir. La modélisation du remboursement complémentaire moyen est donc satisfaisante pour ce type d'actes (c.f. schéma 3).

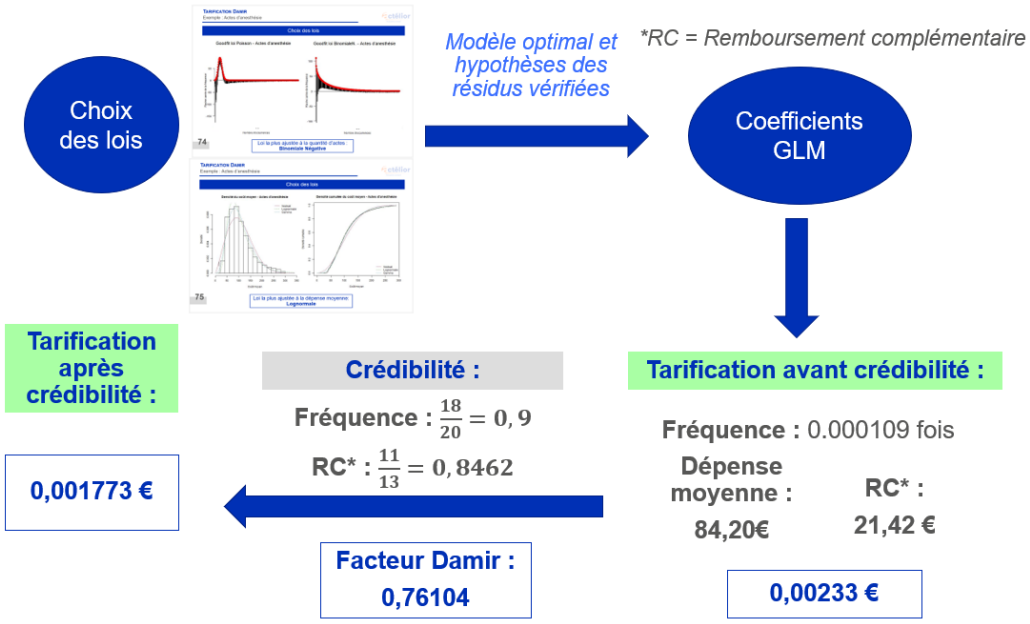


FIGURE 3 : Application du facteur de crédibilité à la tarification Damir des actes d'anesthésie.

Conclusion

Les bases de données Open Damir sont riches en information et permettront une meilleure analyse du marché de l'assurance santé et plus particulièrement de la consommation santé des Français. De plus, les organismes d'assurances et autres organismes d'assurance complémentaire devront à l'avenir faire face à une augmentation majeure du volume des données. Le traitement de ces dernières sera alors inévitable. Ce mémoire confirme la faisabilité du traitement de bases de données volumineuses à l'aide d'outils open source tels que Python et Rstudio, disponible pour tous, mais aussi l'exploitation de ces dernières au sein de méthodes innovatrices utilisées au sein de travaux actuariels.

Les analyses actuarielles effectuées dans le cadre de ce mémoire à partir des bases Open Damir seront utilisées, ajustées et approfondies pour les prochaines années selon les nouvelles données et les évolutions réglementaires éventuelles. De plus, seule la tarification Damir des actes d'anesthésie est présentée dans ce mémoire. En revanche, l'étude est étendue à l'ensemble des libellés brochures afin d'identifier les libellés candidats à l'ajustement de la tarification d'Actélior. En effet, certains actes sont très peu remboursés par la Sécurité Sociale, ou pour lesquels la consommation est très peu élevée. Leur présence est donc limitée au sein des bases de données Damir. Or, un faible nombre de lignes impacte les résultats des Modèles Linéaires Généralisés mais aussi des méthodes de Machine Learning. Finalement, certains libellés brochures ne seront pas utilisés pour l'ajustement de la tarification Actélior. Cet ajustement s'effectue à l'aide du principe de crédibilité. Cependant, la définition du facteur de crédibilité utilisé dans ce mémoire pourra être adaptée selon le besoin et les travaux actuariels futurs du cabinet de conseil Actélior.

Enfin, le suivi technique santé réalisé à partir des bases Open Damir sera alimenté de nouveaux tableaux statistiques et graphiques en fonction des demandes futures et des évolutions réglementaires. Il permettra de mesurer l'impact de ces réformes sur la consommation santé nationale et donc sur les portefeuilles des clients.

Synthesis note

Context

Health insurance is an important part of the insurance industry. Indeed, the share of health care expenditures among national expenditures is high. According to the Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES), this health care expenditure in 2019 amounted to 208 billion euros, which represents 11.3% of the national wealth. The majority of health care expenditure is covered by the Assurance Maladie (78.2% in 2019). The remaining expenses are then covered by various complementary organizations such as mutual insurance companies or provident institutions, or as a last resort by households. To ensure the continuity of this coverage, these organizations must control the health risk of their insured portfolio, in particular by analyzing their health consumption. In this context, the use of external data would enrich these analyses and refine the control of this risk. Today, the main challenge for complementary health insurance organizations is to use national data to position their portfolio's health consumption in relation to national health consumption.

In recent years, the healthcare system has tended to become digitalized, resulting in the collection of increasingly rich data. Numerous digital health databases are therefore important sources of information for insurance companies, health professionals, but also for people wishing to perform analyses on the health market. Since 2015, some of these digital databases have been accessible to all and free of charge. This is Open-Data. The implementation of these Open Data has been possible in close connection with the Data Protection Regulation (RGPD).

Target

As part of this digital transformation, the actuarial consulting firm Actélior wants to offer its clients a complete health risk analysis by integrating a national vision with pre-existing analyses at the level of the insured portfolio. To this end, Actélior's objective is to develop its technical health risk analysis solutions (pricing tool and technical health monitoring) and to enrich them with national health data such as the Open Damir databases.

The purpose of this thesis is to exploit the large Open Damir databases in order to integrate them into Actélior's complementary health contract pricing and health risk analysis tools.

Data

The Open-Data databases of Inter-Scheme Health Insurance Expenditures (DAMIR) are databases that group together all the benefits covered by Social Security. They are anonymized in order to preserve the identity of beneficiaries and health professionals.

The idea was first to process this large volume of data in order to build a final usable database. This step is one of the most important. Without prior processing, the integration of these data into the risk management tools would be impossible. Indeed, in this thesis, only the databases for the years 2018 and 2019 were studied, representing a total of 806,734,365 rows. The hypotheses taken for the various treatments will be explained in this thesis.

First, many variables, considered not useful for the rest of the study, were removed. Second, the anonymization of the Open Damir databases had to be studied. Indeed, the data in the Open Damir databases were aggregated in order to preserve the anonymity of the beneficiaries and the health professionals. A line in the Damir database represents a number N of health care procedures performed for individuals with a similar profile. In addition, groupings by age group and by region were previously carried out on these databases. The number of beneficiaries associated with each line is therefore unknown. However, this information is necessary for frequency modelling

$$Frequency_i = \frac{Number\ of\ acts_i}{Number\ of\ people\ involved_i},$$

with i representing a line of the Open Damir database and thus a given characteristic profile.

This information was reconstructed from INSEE demographic data. These data are broken down by year of care, sex, age group and region. In order not to take into account the entire French population (strong hypothesis), they were then coupled with key figures on the percentage of people covered by compulsory health insurance.

Finally, a correspondence table was constructed in order to link the Actélior health pricing reference system with the health procedure codes present in the Open Damir database. Numerous hypotheses, supported by numerical analyses, were made for the construction of this table, which was then attached to the processed database.

Corrections have been applied to the anomalies in the calculation of the reimbursement, overrun and expense amounts. A provision has also been added to these amounts. It allows us to take into account the services performed in 2019 and reimbursed in 2020, which are therefore not present in our Open Damir reimbursement bases in the framework of the study (only the 2018 and 2019 reimbursement bases were chosen).

Finally, all the data were aggregated to obtain a database whose size was reduced by 80% (i.e. 10,040,857 lines).

Methods

Two of Actélior's technical health analysis tools were fed with these final data:

- The pricing of complementary healthcare products with the Damir base, joined to the existing Actélior healthcare pricing thanks to the credibility theory.
- Actélior's health technical risk monitoring with the addition of a national vision

Figure 4 summarizes the various uses and linkages made between the work of this thesis and Actelior.

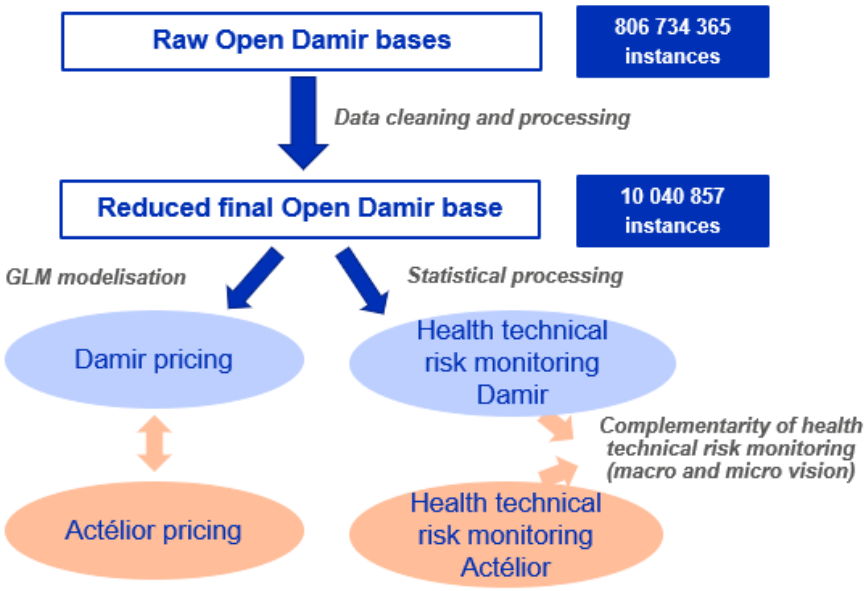


Figure 4: Summary diagram of the work carried out using the Open Damir databases.

First of all, it is important to establish a fair pricing of complementary health contracts to avoid losses for the complementary organization. The objective of this paper is therefore to adjust Actelior's health pricing with pricing modeled on national Open Damir data.

The first step is to construct Damir pricing using the Average Cost x Frequency method. In this study, the average cost represents the average expenditure (the actual cost of the treatment). These two variables of interest will be modeled by a Generalized Linear Model. This same method powers the current Actélior health pricing engine. The choice of a similar method will make it easier to link these two pricing systems, and to compare them at the same level of granularity.

The second step consists in using the credibility principle to adjust the Actélior health pricing with the Damir pricing built on national health expenditure data. For a given brochure wording, a credibility factor is associated with both pricing schemes. However, the definition of this credibility factor is different from the initial theoretical definition. In this study, the credibility factor is defined according to the precision of the GLM coefficients obtained for each explanatory variable of the Damir pricing. If the size of the confidence intervals of these GLM coefficients is very large, then for a given brochure wording, Damir pricing will have very little influence on Actélior pricing. A confidence interval is defined as very large when it is considered an outlier among the other confidence intervals. In other words, the credibility factor is defined as :

$$\mathbf{1 - \alpha_{i,j}} = \frac{Number\ of\ variables\ considered\ as\ non-outliers\ in\ the\ model\ j}{Number\ of\ variables\ included\ in\ the\ model\ j},$$

where $(1 - \alpha_{i,j})$ denote respectively the credibility coefficient associated with the pure premium of the brochure label i and for the modeling of the target variable j, obtained with Damir pricing.

The third and final objective of this thesis consists of integrating the data from the Open Damir database reprocessed in the first part into Actélior's health technical monitoring. This technical analysis tool allows Actélior's clients to know the evolution of the health risk of their portfolio from one year to the next. To date, this technical health analysis tool only allows Actélior's clients to have a vision of the evolution of the cost (amount of reimbursement) and consumption of health services at the level of their portfolio of members. A significant increase in health consumption within the client's portfolio can be highlighted, but is this consistent with the national trend? It is in this context that the Open Damir databases have been integrated into this technical health analysis tool in order to provide a national vision of health consumption. This Damir technical health monitoring is built in the same way as the current Actélior technical health monitoring. The analyses go from a macroscopic level (global analysis, analysis by beneficiary) to a more detailed level (analyses by type of beneficiary, by family of procedures, by type of procedure). Finally, an additional analysis has been built on the three health items impacted by the 100% health reform in order to understand and analyze the evolution of national consumption in dental, optical and audiology care. This tool was built to automate future processing. Only the final reprocessed Damir base needs to be updated (correspondence table for example). The technical monitoring dashboard will then be updated automatically.

Main results

This thesis integrates various methods and tools with the objective of optimizing, accelerating and automating various essential processes during the construction of a health care pricing.

In health pricing, the two target variables are modeled for each brochure wording of the pricing reference frame defined beforehand. For each model, the variables available in the data are then selected using a Gradient Boosting method: the Light GBM unsupervised learning method. This method is adapted to large databases, for which the results are obtained with a similar accuracy to other Gradient Boosting methods, but with a reduced execution time. The efficiency of this method is verified in this thesis, with an execution time for LGBM that is much lower than that of Gradient Boosting and with an equivalent accuracy (c.f. figure 2).

Table 2: Performance of the Machine Learning methods XGBoost and LightGBM.

Average expenditure			
	Score - MSE	Score - RMSE	Running time
XGBoost	66.68	8.17	00:35:54
LightGBM	46.66	6.83	00:01:36
Quantity of health procedures			
	Score - MSE	Score - RMSE	Running time
XGBoost	9917.50	99.59	00:34:28
LightGBM	10000.35	100.00	00:01:21

Finally, the variables retained are the age of the beneficiary (AGE_BENEF), the region of the beneficiary (REGION_BENEF) and the year of occurrence of the care (SOI_ANN).

Moreover, the use of specific R packages allows us to automatically detect the best fitting law among the possible laws (Gamma, Weibull and Lognormal probabilities laws for the average expense, negative Binomial and Poisson laws for the quantity of health procedures).

Within the framework of this thesis and the construction of the Damir pricing system, only anaesthesia procedures are priced and only the results of this brochure are presented. After performing the different treatments, the optimal Generalized Linear Model obtained is satisfactory, for which all the explanatory variables are significant and the hypotheses of the residuals are verified (c.f. figure 3.22).

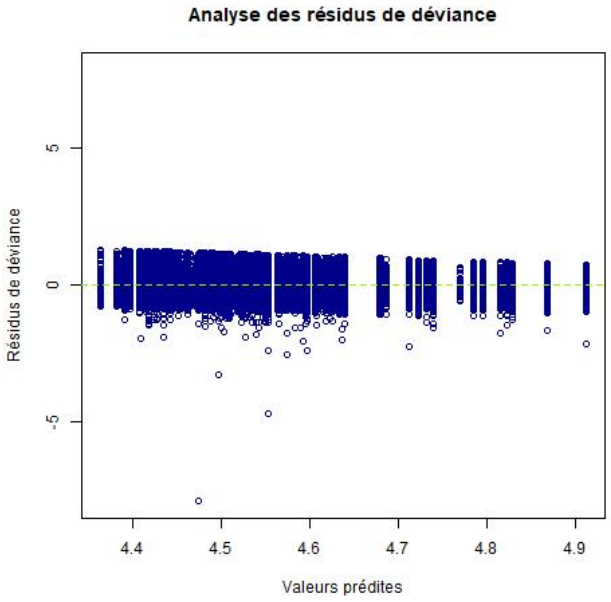


Figure 5: Graph to verify the linearity hypothesis of the Deviance residuals.

The price is then calculated for a reference beneficiary, defined as a 40 year old person living in Ile-de-France, using the GLM coefficients and other extracted parameters. In the context of the pricing of anesthesia health procedures, the Damir pricing amounts to 0.00917€. Actélior pricing is 0.001€for this type of health procedure, which is close to the Damir pricing. The modeling of the average complementary reimbursement is therefore satisfactory for this type of health procedure (c.f. figure 6).

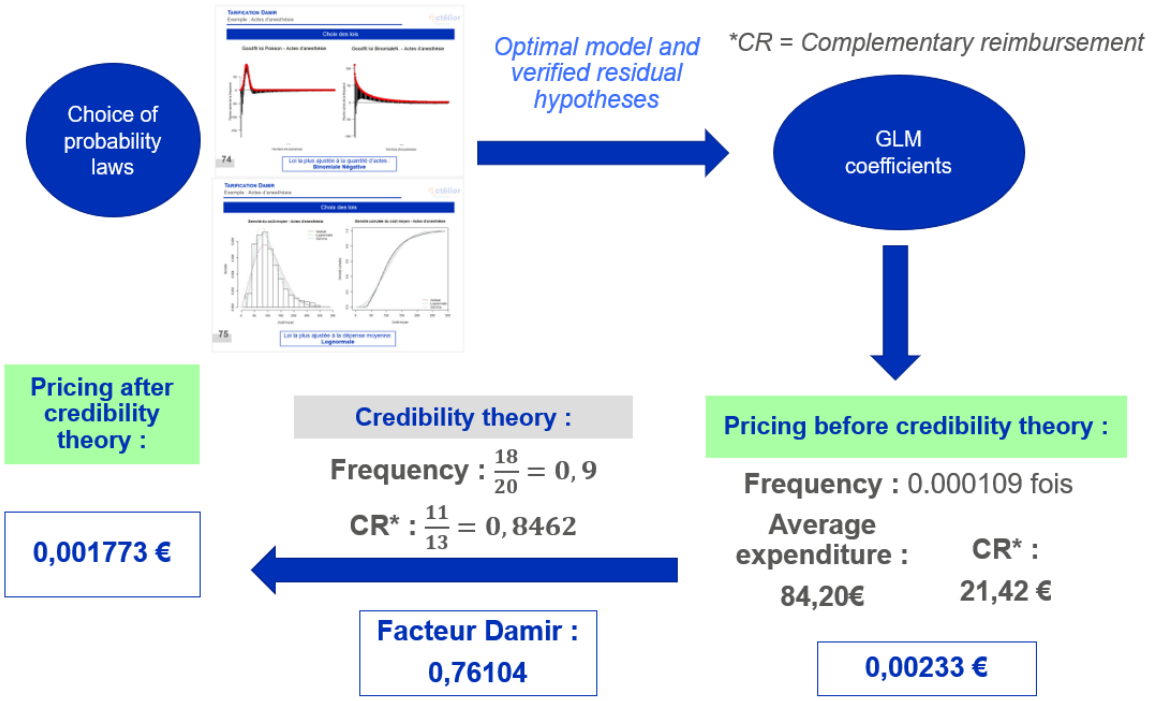


Figure 6: Application of the credibility factor to the Damir pricing of health anaesthetic acts.

Conclusion

The Open Damir databases are rich in information and will allow a better analysis of the health insurance market and more specifically of the French health consumption. Moreover, insurance companies and other complementary insurance companies will have to face a major increase in the volume of data in the future. The processing of these data will then be inevitable. This thesis confirms the feasibility of processing large databases using open source tools such as Python and Rstudio, which are available to everyone, but also the exploitation of these databases within innovative methods used in actuarial analyses.

The actuarial analyses performed in this thesis from the Open Damir databases will be used, adjusted and deepened for the next years according to new data and possible regulatory changes. In addition, only Damir pricing of anesthesia procedures is presented in this brief. However, the study is extended to all the brochure titles in order to identify the candidate titles for adjustment of the Actélior pricing. In fact, some procedures are not reimbursed very much by Social Security, or for which consumption is very low. Their presence is therefore limited in the Damir databases. However, a low number of lines impacts the results of the Generalized Linear Models but also of the Machine Learning methods. Finally, some brochure labels will not be used for the Actelior pricing adjustment. This adjustment is performed using the credibility principle. However, the definition of the credibility factor used in this brief may be adapted according to the need and future actuarial work of Actelior.

Finally, the technical health monitoring carried out from the Open Damir databases will be fed with new statistical tables and graphs according to future requests and regulatory changes. It will make it possible to measure the impact of these reforms on national health consumption and therefore on client portfolios.

Remerciements

Je souhaite en tout premier lieu remercier ma tutrice entreprise Madame Elodie PAGET, Directrice Générale Adjointe et Actuaire IA au sein du cabinet de conseil Actélior, pour son suivi permanent, ses précieux conseils et le partage de son expérience en assurance santé qui m'a permis d'appréhender plus facilement ce mémoire. Je la remercie également pour le soutien non négligeable qu'elle a pu m'apporter tout au long de cette épreuve.

Mes remerciements s'adressent notamment à Monsieur David ECHEVIN, Directeur Général d'Acté--lior, et Monsieur Romain GRACZ, Directeur Associé et Actuaire IA chez Actélior, et Madame Elodie PAGET, de m'avoir laissé l'opportunité d'intégrer Actélior et de travailler sur ce sujet passionnant.

Je remercie également l'ensemble des collaborateurs d'Actélior pour leur bienveillance, leur soutien et leurs différents conseils sur les travaux et la rédaction de mon mémoire.

De plus, j'aimerais remercier ma tutrice académique, Madame Claire LAMON, Actuaire IA chez Allianz, pour son suivi de qualité et les différentes explications techniques à propos différents sujets qu'elle a pu m'enseigner lors de ce suivi.

Je voudrais aussi adresser mes remerciements à l'équipe pédagogique et administrative du Master 2 Actuariat de l'Université Paris-Dauphine, qui m'ont permis d'intégrer cette formation actuarielle de grande qualité, malgré les difficultés liées à la crise sanitaire, ainsi que pour leur contribution à l'atteinte de mes objectifs professionnels en actuariat.

Enfin, je remercie particulièrement ma famille, mes proches et mes amis pour leur soutien inconditionnel tout au long de mon parcours scolaire et de la rédaction de mon mémoire.

Table des matières

Introduction

Le monde de l'assurance, et principalement la réglementation qui l'encadre, sont en constante évolution. Afin de respecter ces nouveaux enjeux, les différents acteurs de l'assurance doivent perpétuellement s'adapter. Des méthodes et outils de gestion des risques doivent être construits pour pouvoir analyser, comprendre les portefeuilles et ainsi quantifier et maîtriser les risques. La digitalisation des processus permet une collecte de données de plus en plus riche et pertinente. De nombreuses bases numériques sont à présent publiques et accessibles gratuitement. Il s'agit de données Open-Data, dont la mise à disposition est encadrée par le Règlement de la Protection des Données (RGPD). Dans le cadre de la santé, les bases Open Damir ont été mises à disposition dès 2015. Elles sont construites à partir de données extraites du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIR-RAM) et concernent l'ensemble des remboursements effectués par l'Assurance Maladie, tous régimes confondus. Bien qu'anonymisées, ces bases comportent de nombreuses informations sur les remboursements et constituent donc un volume important de ressources exploitables. Cependant, malgré la richesse des informations pouvant favoriser la recherche, l'innovation ou bien contribuer à l'approfondissement de travaux pré existants, ce volume de données n'est pas ou très peu exploité. L'enjeu pour les organismes assureurs consiste à intégrer ces nouvelles données au sein de leurs outils afin d'enrichir leurs analyses et affiner leur maîtrise des risques.

Afin d'accompagner ses clients dans cette transformation digitale, le cabinet de conseil en actuariat, Actélior, souhaite faire évoluer ses outils de tarification et d'analyse du risque santé. Pour cela, Actélior souhaite enrichir ses solutions à l'aide de bases Open Data disponibles dans le secteur de la santé. Les bases Open Damir ont donc été sélectionnées. N'étant pas encore utilisées avant la rédaction de ce mémoire, l'exploitation de ces données constituerait alors une véritable valeur ajoutée aux différents travaux actuariels, notamment par l'intégration d'une analyse nationale de la consommation santé, complémentaire à l'analyse du portefeuille client. Cette comparaison permettrait donc aux clients de se positionner et d'améliorer leur maîtrise du risque. De plus, l'objectif de la tarification santé étant de réaliser un tarif le plus proche possible de la consommation réelle des adhérents et selon diverses caractéristiques, les données nationales Open Damir permettrait donc d'établir une tarification plus ajustée. C'est dans cette optique que ce mémoire tentera de répondre aux différentes interrogations du cabinet sur le contenu et l'utilisation des bases Open Damir en assurance santé : Quelles sont les différents types de données disponibles ? Sont-elles exploitables ? Comment les retraiter afin de les intégrer au sein des différents outils d'analyse du risque et de tarification ?

Pour répondre à ces diverses problématiques, les étapes et toutes les hypothèses sous-jacentes seront détaillées au sein de ce mémoire. Dans un premier temps, les principes de l'Assurance Maladie Obligatoire et des organismes complémentaires santé seront présentés, ainsi que les règlements et réformes qui régissent l'assurance santé française et l'Open Data.

Ensuite, le retraitement de la base Open Damir, nécessaire pour la rendre exploitable et faciliter son intégration au sein des outils Actélior, sera réalisé à partir de différentes analyses présentées dans ce mémoire. Des réflexions approfondies et l'utilisation de méthodes innovantes, de Machine Learning par exemple, alimenteront le retraitement de ces bases de données volumineuses. Enfin, le deuxième chapitre de ce mémoire, enrichi d'une analyse descriptive de la base traitée, permettra de visualiser les données obtenues.
La troisième partie de ce mémoire exposera les démarches et les résultats de la tarification santé établie sur notre base de données Open Damir retraitée. Elle sera composée d'une partie théorique rappelant le concept du Modèle Linéaire Généralisé (MLG) utilisé dans le cadre de cette tarification et souvent utilisé dans le domaine de l'assurance. En effet, ces résultats explicites et facilement exploitables sont appréciés des différents acteurs. Puis, nous interprèterons les résultats obtenus avec l'exemple des actes d'anesthésie.
Finalement, la quatrième et dernière partie jouera un rôle central dans l'intégration des informations contenues dans la base de données retraitée, mais aussi de la tarification santé Damir au sein des outils actuellement utilisés par Actélior. Le suivi technique santé d'Actélior sera donc enrichi des données nationales Open Damir. Enfin, la tarification santé actuelle d'Actélior sera associée à la tarification Damir par la mise en place d'un modèle de crédibilité sur-mesure. L'objectif consistera à appliquer un facteur de crédibilité à ces deux tarifications, dont la définition sera détaillée au sein de ce chapitre.
A présent, découvrons l'ensemble des travaux effectués pour l'intégration de la base Open Damir en assurance santé.

Chapitre 1

L'apport de l'Open Data en assurance santé

1.1 Le système français de l'assurance santé

Au sein du système français de l'assurance santé, de nombreux acteurs se mobilisent afin d'assurer la pérennité de celui-ci. Il permet de préserver la santé des résidents français et de garantir l'accès aux soins pour tous. Les acteurs qui coexistent et interagissent entre eux sont l'Assurance santé Maladie Obligatoire intégré dans le système de la Sécurité Sociale, mais aussi les organismes complémentaires santé, sans oublier les professionnels de santé. Ce mémoire étant axé sur l'assurance santé, nous détaillerons plus amplement l'organisation et le rôle de chaque acteurs dans les parties suivantes.

1.1.1 La Sécurité Sociale et l'Assurance Maladie

Pierre Laroque, directeur de la Sécurité Sociale de 1944 à 1951, décrivait l'objectif du projet de Sécurité Sociale lors des Ordonnances des 4 et 19 octobre 1945 : "Les caisses ont pour rôle de garantir des moyens d'existence à tous les travailleurs qui se trouvent privés de ressources par suite de maladie, de maternité, d'invalidité ou de vieillesse."

L'organisation de la Sécurité Sociale

En effet, au lendemain de la deuxième la guerre mondiale, le projet de construire un vaste système d'entraide collective et obligatoire naît. Ce projet permettrait de consolider la solidarité entre les travailleurs et les personnes non actives, les personnes saines et malades, etc. Créée donc il y a plus de 70 ans, en 1945, la Sécurité Sociale a permis aux français de vivre plus longtemps et surtout dans des conditions meilleures. Elle fait partie d'un immense système de solidarité collective composé de nombreuses institutions. L'objectif est de protéger l'intégralité des résidents français face à divers risques auxquels ils peuvent être confrontés tout au long de leur vie et pouvant les impacter fortement financièrement. En retour, la population française participe au financement de ce système. Le principe de financement est celui du pot commun, un principe de solidarité.

La Sécurité Sociale se compose de diverses institutions tels que les caisses de la Sécurité Sociale de chaque régime et branches. Il existe aussi les organismes de tutelle comme la Direction de la Sécurité

Sociale (DSS) qui gouverne la Sécurité sociale.

Le système de Sécurité Sociale est subdivisé en trois régimes différents, dont deux régimes généraux, et des régimes spéciaux adaptés aux différentes catégories socioprofessionnelles. Plus particulièrement, ces régimes sont :

- **le régime général**. Il concerne l'ensemble des personnes actives hors salariés agricoles, incluant depuis le premier janvier 2018 le régime des salariés indépendants (plus communément nommés le RSI). Ce régime couvre près de 90% de la population ;
- **le régime agricole**. Il prend en charge l'ensemble des salariés et exploitants agricoles et est géré par la caisse nationale de Mutualité Sociale Agricole (MSA) ;
- **les régimes spéciaux** comme par exemple celui de la RATP, de la SNCF. Ce sont des régimes prenant en charge les salariés de certaines grandes entreprises publiques. Le régime Alsace-Moselle est aussi considéré comme un régime spécial de la Sécurité Sociale.

Plus particulièrement, le régime général est formé de cinq branches faisant référence aux différents risques auxquels un résident français peut faire face lors de période critique de la vie. Quatre branches concernent les dépenses et une branche concernent les recettes de la Sécurité Sociale. Ces branches sont :

- **la branche cotisation et recouvrement**. Composée des Unions de Recouvrement des cotisations de Sécurité Sociale et d'Allocations Familiales (URSSAF), elle constitue l'ensemble des recettes alimentées par les cotisations versées par les particuliers et les professionnels et reversées aux caisses de la Sécurité Sociale. Cela permet de financer les diverses prestations des autres branches ci-dessous ;
- **la branche famille**. Dirigée par la Caisse d'Allocations Familiales (CAF), elle couvre différents types de prestations comme la naissance, la garde d'enfants, les aides à l'éducation ou au logement ;
- **la branche retraite**. Dirigée par l'Assurance Retraite, elle a pour objectif le suivi régulier des salariés tout au long de leur vie pour la contribution à la retraite mais aussi lors du versement de celle-ci ;
- **la branche maladie**. Elle couvre les dépenses d'hospitalisation, de médicaments et de consultations auprès de professionnels de santé. Elle est dirigée par l'Assurance Maladie ;
- **la branche Accident de Travail (AT) et Maladies Professionnelles (MP)**, dirigée par l'Assurance Maladie.

L'organigramme 1.1, extrait du site Ameli (2021), résume les explications ci-dessus.

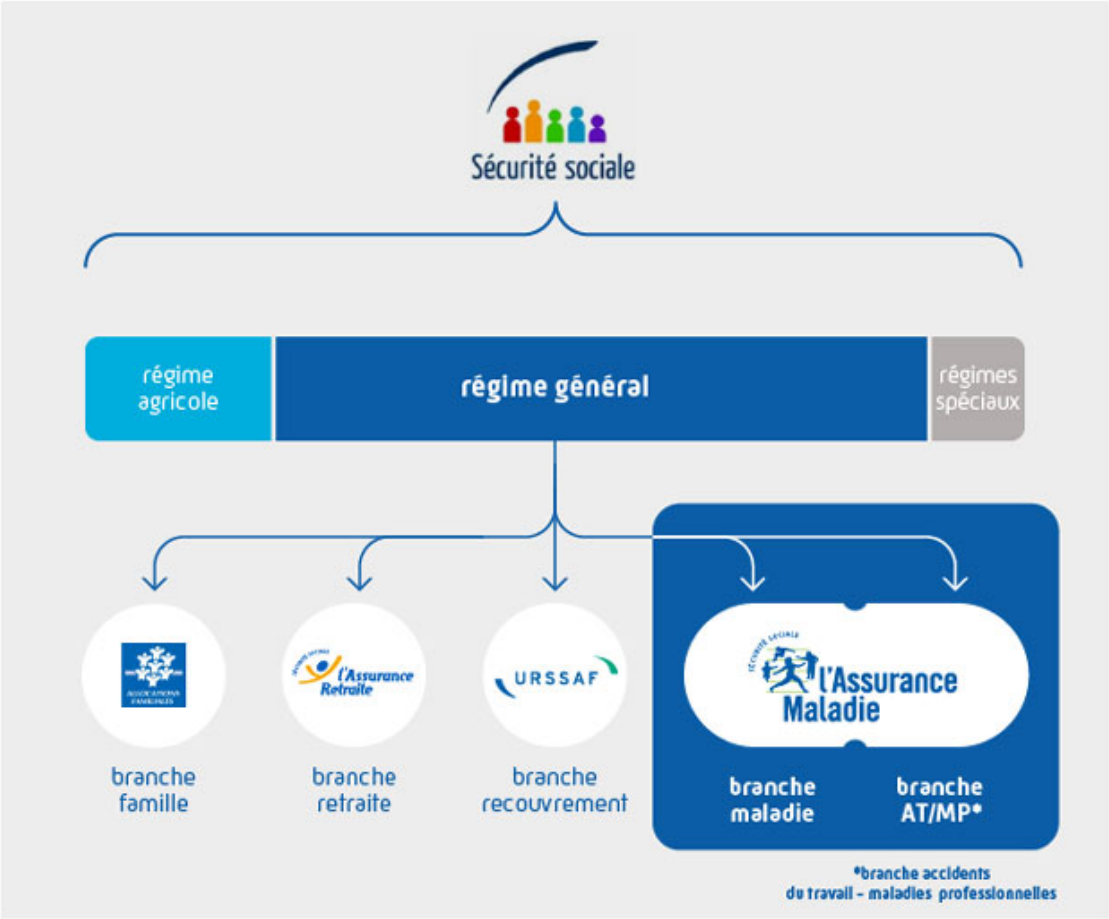


Figure 1.1 : Décomposition de l'organisation de la Sécurité Sociale

Dans le cadre de l'étude, nous étudierons les remboursements de prestations de soins de la Sécurité Sociale. Cependant, le périmètre étudié sera limité au Régime Général. Nous analyserons donc les deux branches liées aux soins de santé du Régime Général, c'est-à-dire l'Assurance Maladie Obligatoire. La prochaine partie explicitera alors les principes de l'Assurance Maladie et donc des deux dernières branches du Régime Général.

Principes de l'Assurance Maladie

L'Assurance Maladie Obligatoire (AMO) est donc une composante de la Sécurité Sociale, prenant en charge les frais des soins subis par les assurés, et liés à différents risques comme la maladie, la maternité, l'invalidité et le décès. Elle participe notamment au financement d'aide directement aux professionnels. Elle gère deux des cinq branches du régime général de la Sécurité Sociale : la branche maladie et la branche accidents de travail et maladies professionnelles. Ces branches sont gérées par un réseau de proximité de caisses. En effet, la Caisse Nationale de l'Assurance Maladie (CNAM) impose la stratégie au niveau national et coordonne 152 organismes au niveau régional et local. Son réseau se compose des Caisses Primaires d'Assurance Maladie (CPAM), des Caisses Générales de Sécurité Sociale (CGSS) dans les départements d'outre-mer, des Directions Régionales du Service Médical

(DRSM), des Caisses d'Assurance Retraite et de la Santé Au Travail (CARSAT), ainsi que des Unions de Gestion des Etablissements de Caisse d'Assurance Maladie (UGECAM) (c.f. figure 1.2).

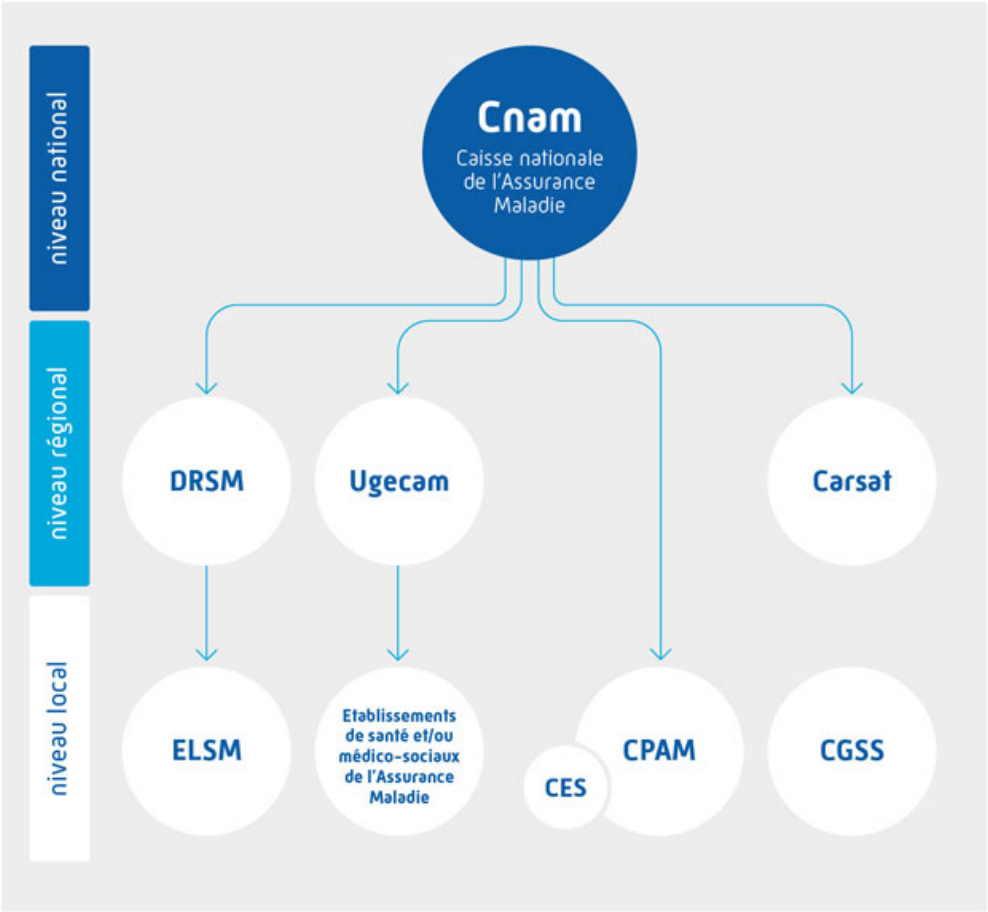


FIGURE 1.2 : Réseau de l'Assurance Maladie Obligatoire.

Le déploiement de ces nombreuses CPAM au niveau local (102 caisses déployées dans toute la France) permet d'accompagner individuellement les assurés sur les risques maladie et professionnel afin d'identifier ces risques éventuels et de mettre en place des actions de prévention. Cela permet de préserver la santé des français et donc de contribuer à l'amélioration de l'efficacité du système de santé, notamment en rendant les soins de plus en plus accessible à la population. Cependant, la gestion de ce système est remise en question, notamment dans le rapport HCAAM (2021). En effet, de nombreuses disparités existent au sein du système, comme une prise en charge par la Sécurité Sociale quasi nulle sur certains postes, une couverture segmentée en fonction du statut professionnel et de l'âge, etc.

Quelques chiffres sur l'Assurance Maladie Obligatoire

Cette partie présentent les chiffres clés sur la Sécurité Sociale et l'Assurance Maladie Obligatoire. Nous avons décidé de présenter les résultats jusqu'à 2019 et exclure l'année 2020. En effet, l'année 2020 étant particulière à cause du confinement et de la crise sanitaire du Covid-19, les évolutions ne suivent donc pas la tendance habituelle et les analyses seraient alors impertinentes. Cependant, les données 2020 ont été tout de même mises à jour depuis la rédaction du mémoire. Il est possible de consulter ces mises à jour sur les sites de la Sécurité Sociale et de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES).

La Sécurité Sociale est un acteur majeur au sein du système français. 160 000 hommes et femmes assurent le bon fonctionnement de ce système de santé. Ils gèrent un budget de 470 milliards d'euros de prestations, ce qui représente un quart de la richesse nationale. De plus, la Sécurité Sociale participe au financement de près de 78,2% des dépenses liées à la consommation de soins et des biens médicaux sur le marché de l'assurance santé (chiffres provenant du rapport DREES (2019b)). Les dépenses restantes sont financées par les organismes complémentaires majoritairement, mais aussi les dispositifs de solidarité et les ménages. Ce rôle prépondérant de la Sécurité Sociale apparaît nettement sur le graphique 1.3, issu du rapport DREES (2019b).

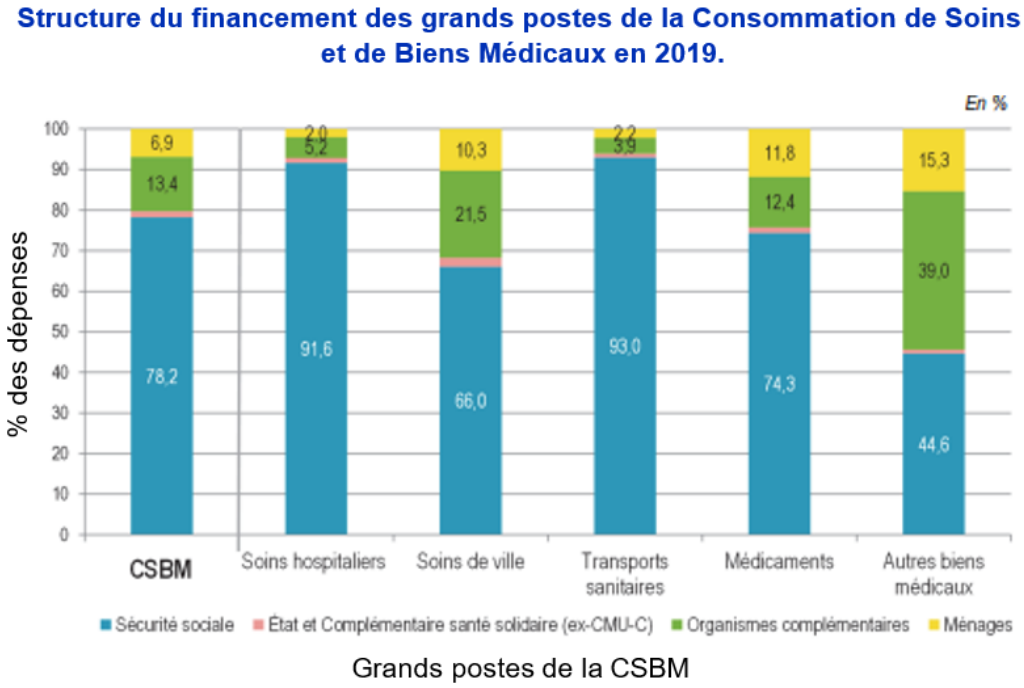


FIGURE 1.3 : Structure de financement des grands postes de consommation santé en 2019.

Malgré la présence majoritaire de la Sécurité Sociale dans les remboursements des soins, nous pouvons remarquer la présence d'une certaine disparité au niveau de la prise en charge de certains postes de soins. En effet, les autres biens médicaux, intégrant le dentaire et l'optique, sont les postes de soins les moins pris en charge par l'Assurance Maladie. Nous verrons au fil de ce mémoire que la prise en charge de ces postes est majoritairement assurée par les organismes complémentaires santé.

De plus, concernant le régime général de la Sécurité Sociale, près de 59,2 millions de personnes bénéficiaient de ce système de couverture en 2019, ce qui représente environ 88% de la population française. Pour l'Assurance Maladie, 202,8 milliards d'euros de prestations ont été gérées et versées

en 2019, ce qui représente un niveau de 11% du Produit Intérieur Brut (PIB) (c.f. le rapport de Direction de la Sécurité Sociale (2019)).

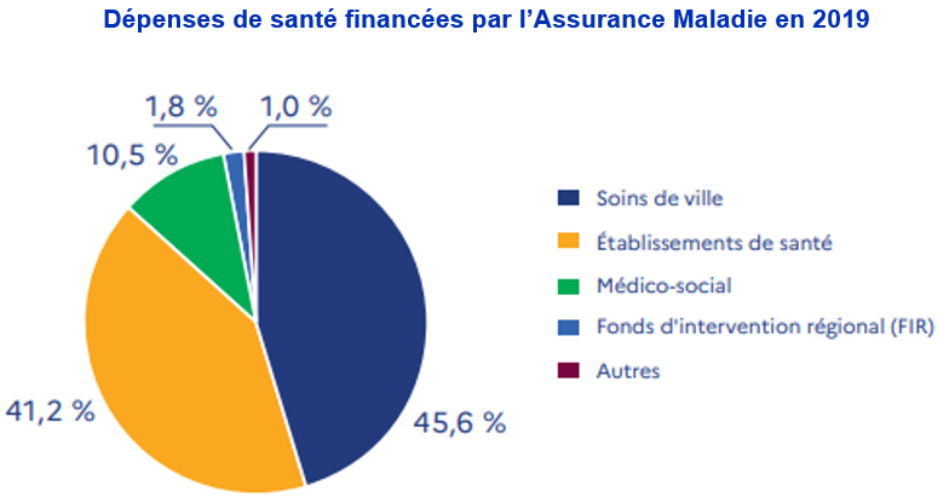


Figure 1.4 : Les dépenses de santé financées par l'Assurance Maladie en 2019.

Comme nous pouvons le voir sur le graphique 1.4, les postes les plus importants en termes de dépenses en consommation santé et couverts par l'Assurance Maladie sont les établissements de santé et les soins de ville. Les soins de villes concernent les honoraires médicaux, les indemnités journalières, les médicaments et dispositifs médicaux, ainsi que les transports. Les établissements de santé sont le plus souvent des hôpitaux ou cliniques, publics ou privés, qui assurent la détection et le traitement de maladies, le soin de personnes blessées et de femmes enceintes. En ce qui concerne les services médico-sociaux, ils accompagnent les personnes handicapées, dépendantes ou en situation d'exclusion sociale. Enfin, les Fonds d'Intervention Régional (FIR) sont des organismes de financement de recherches et d'applications en santé. Ils sont gérés par les agences régionales de santé afin d'améliorer, par exemple, la performance, la qualité ou bien la prévention du système de santé Français.

Aujourd'hui, la branche Maladie de la Sécurité Sociale dispose d'une balance budgétaire déficitaire. En effet, le solde, défini comme la différence des recettes et des dépenses, est négatif depuis quelques années. Néanmoins, ce déficit diminuait sur les dernières années (c.f. le graphique 1.5 issu du rapport de la Direction de la Sécurité Sociale (2019)).

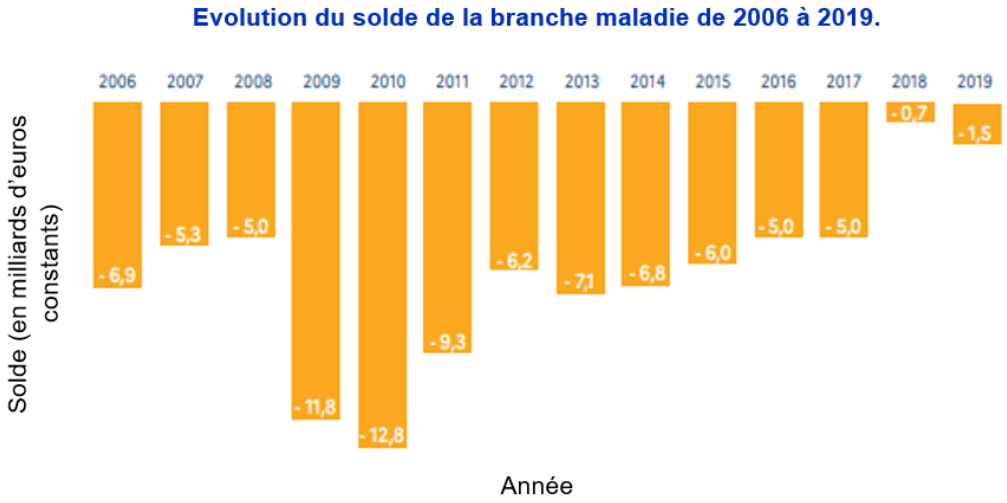


FIGURE 1.5 : Evolution du solde de la branche maladie de 2006 à 2019.

La tarification est donc une étape essentielle. Tarifer correctement des contrats d'assurance permet alors d'obtenir une balance budgétaire optimale. Cet enjeu est notamment valable pour les contrats complémentaires santé. Ce mémoire permettra d'ajuster une tarification de contrat complémentaire santé à l'aide de données nationales afin d'aider les organismes complémentaires santé à réduire leur déficit budgétaire.

1.1.2 La complémentaire santé

Le principe de la complémentaire santé

L'étude réalisée dans ce mémoire consiste à lier une tarification santé des dépenses de soins de la Sécurité Sociale et la tarification santé de contrat complémentaire santé d'Actélior. C'est la raison pour laquelle nous nous attacherons à expliquer les principes du marché de la complémentaire santé.

L'Assurance Maladie Complémentaire (AMC) a pour objectif de compléter les garanties de base prises en charge par l'Assurance Maladie Obligatoire. Elle peut notamment couvrir des prestations non remboursées initialement par l'AMO afin d'assurer une couverture totale à ses adhérents. L'adhérent peut donc, sur certains soins, obtenir un reste à charge nul grâce à sa complémentaire santé. Pour cela, l'Assurance Maladie Complémentaire se constitue d'un ensemble d'organismes privés proposant à ses adhérents (particuliers ou entreprises), des contrats obligatoire ou facultatif, à titre individuel ou collectif (LAZIC, 2020) :

- **les contrats individuels :** Ces contrats sont souscrits librement par un particulier. Une complémentaire santé est donc optionnelle/ facultative pour la souscription individuelle ;
- **les contrats collectifs :** Ces contrats sont souscrits par les entreprises. En effet, depuis la mise en vigueur de l'Accord National Interprofessionnel (ANI), les employeurs ont l'obligation de proposer une complémentaire santé collective à leurs salariés, qui peut notamment être appliquée aux membres de leurs familles (conjoints et enfants). Le salarié choisit par la suite d'accepter ou non la proposition de son employeur.

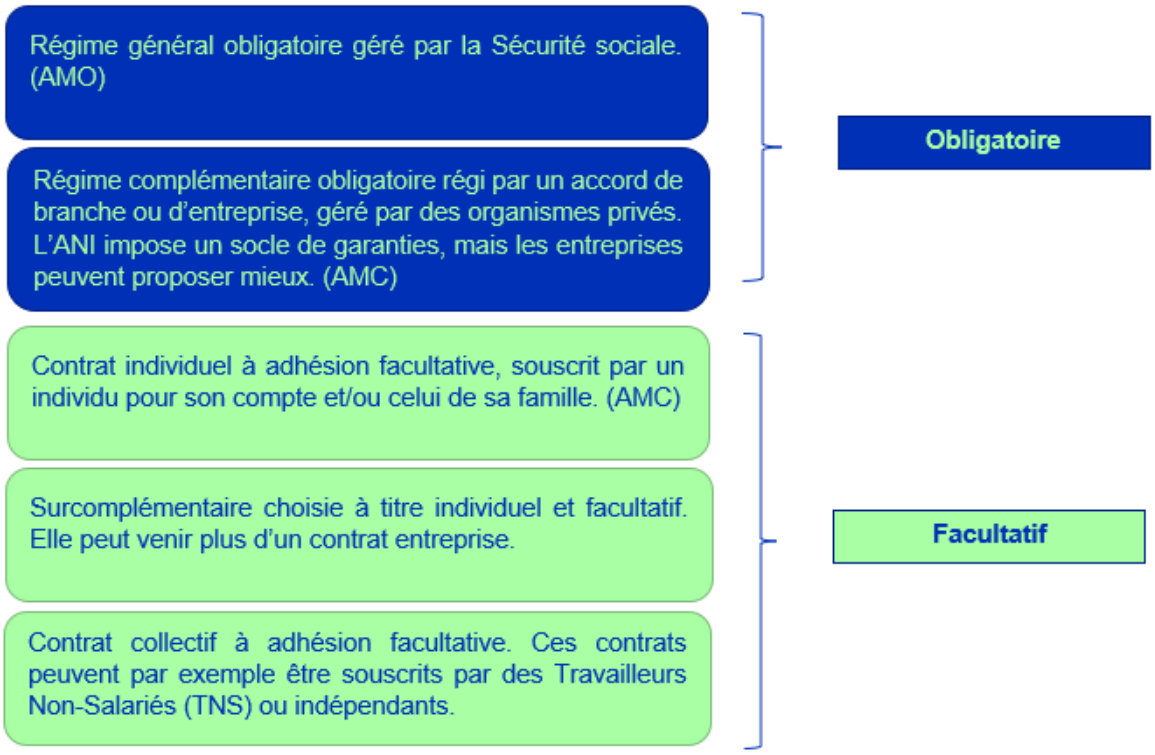


FIGURE 1.6 : Les différents types de contrats souscrits en assurance santé.

Ces contrats (c.f. leur résumé en figure 1.6) peuvent être souscrits dans divers organismes privés, explicités ci-après.

Les organismes complémentaires santé

Les organismes privés de l'assurance complémentaire couvrant les risques du domaine de la santé peuvent être :

- **des compagnies d'assurances**. Elles sont régies par le Code des assurances et peuvent être juridiquement des Sociétés Anonymes (SA) ou bien des Sociétés d'Assurance Mutuelle (SAM). Elles agissent sur tous les domaines en termes de couverture des assurés ;
- **des mutuelles**. Elles sont régies par le Code la mutualité et sont des groupements à but non lucratif, gouvernés par leurs adhérents. Leur domaine d'action est assez large, touchant les frais de santé mais aussi la prévoyance telle que les arrêts de travail ;
- **des Institutions de Prévoyance (IP)**. Elles sont régies par le Code de la Sécurité Sociale et sont des organismes paritaires et à caractère non lucratif. Cela signifie qu'elles ne possèdent pas d'actionnaires et que les décisions sont votées par des personnes morales de droit privé que sont ses adhérents, et ceci de manière paritaire. Elles se concentrent donc sur la qualité des services et des garanties proposées aux adhérents. De plus, leur domaine d'action est concentré sur les contrats collectifs de prévoyance, souscrits par les entreprises.

Remarque : Souvent, les termes de complémentaire santé et mutuelles sont confondus. La mutuelle est un organisme, alors que la complémentaire santé est un produit proposé par cet organisme.

Quelques chiffres sur le marché de l'assurance santé complémentaire

Cette partie présentent les chiffres clés sur la complémentaire santé. Comme précédemment, nous détaillerons les résultats et chiffres jusqu'à l'année 2019, l'année 2020 étant exclue.

Aujourd'hui près de 94% de la population française est couverte par une assurance complémentaire santé. Cette couverture est assurée, comme vu précédemment, par divers organismes (c.f. le rapport de la DREES (2019a)).

Le nombre de mutuelles a diminué depuis les années 2000 (diminution de 67% environ). En effet, les 1 158 mutuelles existantes dès 2006 prenaient en compte les petites tout comme les grandes mutuelles (c.f. graphique 1.7). Cependant, avec l'instauration du nouveau régime prudentiel Solvabilité 2 (S2) ou bien la mise en place des contrats collectifs obligatoires pour les entreprises, ces organismes ne pouvaient assurer une telle charge supplémentaire, plus contraignante qu'auparavant. Cette diminution n'est donc pas due à la disparition de ces mutuelles, mais surtout à la concentration et à la fusion de celles-ci.

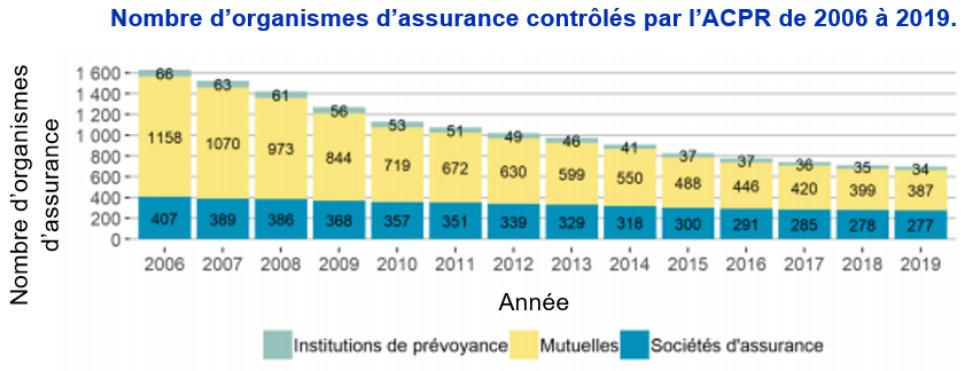


FIGURE 1.7 : Nombres d'organismes d'assurances contrôlés par l'Autorité de Contrôle Pudentiel et de Résolution de 2006 à 2019.

Les mutuelles sont majoritairement présentes sur le marché de la santé par rapport aux deux autres types d'organismes. Son activité de la santé représente 85% des cotisations collectées (c.f. graphique 1.8).

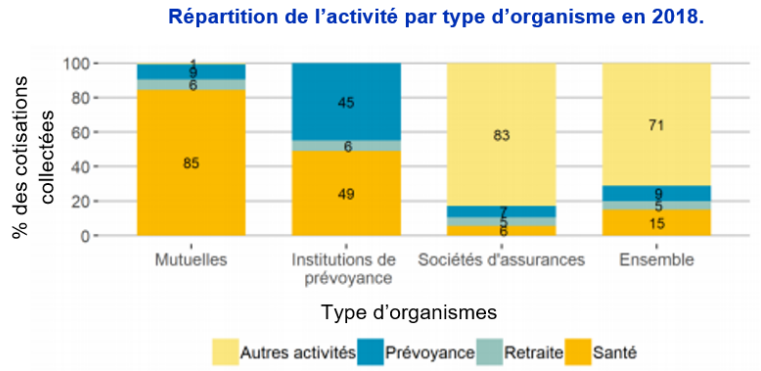


FIGURE 1.8 : Répartition de l'activité par type d'organisme en 2018.

De plus, le marché des mutuelles se concentre principalement sur des contrats individuels, qui représente 70% des cotisations individuelles. Elles ont plus de difficultés à se tourner vers le marché du collectif, majoritairement présent pour les institutions de prévoyance. En effet, les institutions de prévoyance ne sont pas autorisées à commercialiser des contrats individuels, hormis les surcomplémenta--ires santé. Elles se concentrent donc principalement sur les contrats collectifs (c.f. graphique 1.9).

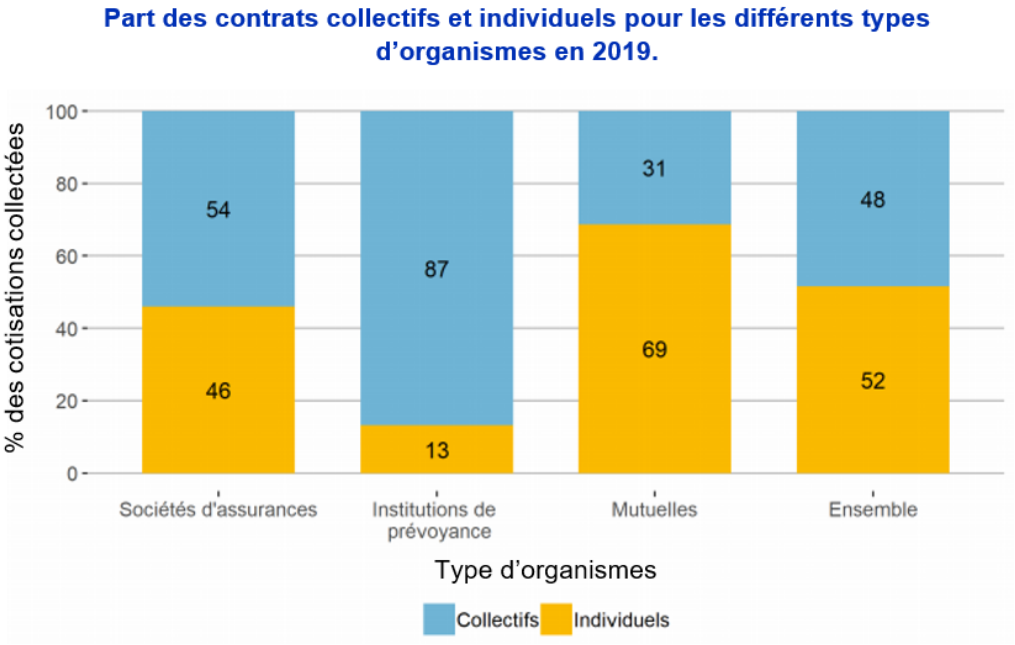


FIGURE 1.9 : Part des contrats collectifs et individuels sur l'ensemble des cotisations collectées en santé par les différents types d'organismes en 2019.

Le graphique 1.10 ci-dessous montre la répartition des prestations par postes de soins en % des cotisations collectées. Nous remarquons que la prise en charge en optique et dentaire est largement supérieure pour les IP et les contrats collectifs. Cela provient du fait que les IP propose essentiellement des contrats collectifs obligatoires. Ces contrats bénéficient d'une mutualisation plus forte que les contrats individuels puisqu'un niveau de garantie est imposée au salarié et que ce dernier ne fait pas la démarche de choisir un niveau en cohérence avec ses besoins.

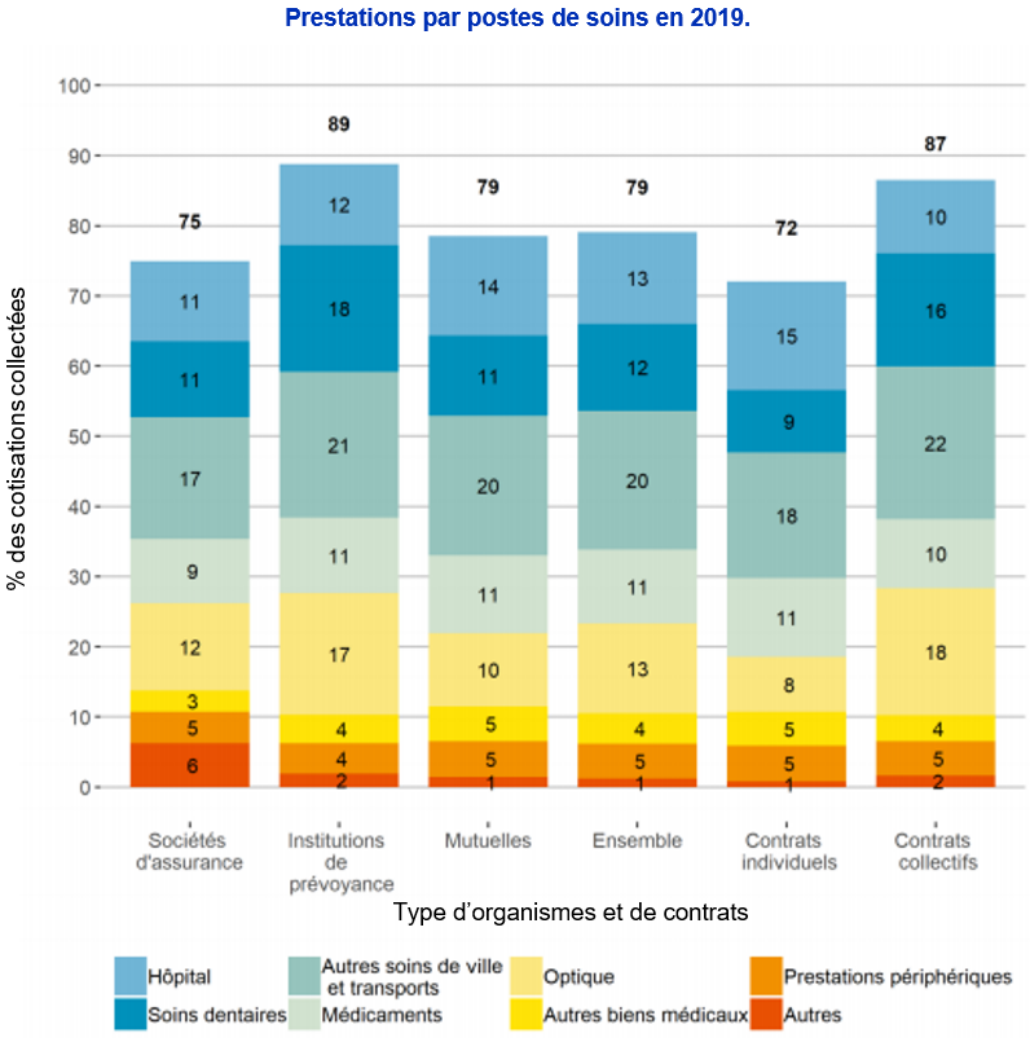


FIGURE 1.10 : Prestations par postes de soins en 2019.

1.1.3 Les réglementations et dispositifs en assurance santé

De nombreuses réglementations et dispositifs ont été mis en place ces dernières années au cœur du système de l'assurance santé. Ces évolutions réglementaires ont pour objectif de faire évoluer le système de santé et de rendre plus accessible les soins. Les réglementations et dispositifs abordés dans cette partie sont :

- l'Option Pratique Tarifaire Maîtrisée (OPTAM) et l'Option Pratique Tarifaire Maîtrisée Chirurgie et Obstétrique (OPTAM-CO) ;
- le contrat responsable ;
- la réforme 100% santé ;
- la Couverture Maladie Universelle Complémentaire (CMU-C) ;
- l'ANI et les contrats complémentaires collectifs obligatoires.

Les dispositifs OPTAM et OPTAM-CO

L'Option Pratique Tarifaire Maîtrisée (OPTAM) est un accord réalisé entre l'Assurance Maladie et les médecins, remplaçant le Contrat d'Accès aux Soins (CAS) depuis 2017 (c.f. graphique 1.11). Ce dispositif a été mis en place afin de :

- limiter les dépassements d'honoraires des médecins devenant trop importants, pendant une durée de trois ans ;
- favoriser les consultations au tarif conventionnel (alignés à la base de remboursement de la Sécurité Sociale) ;
- réduire le renoncement aux soins du fait des restes à charge trop importants pour les individus.

En effet, certains médecins appliquent des dépassements d'honoraires qui ne sont pas remboursés par le Régime Obligatoire. Selon les garanties proposées par le contrat complémentaire santé de l'adhérent, les dépassements peuvent être partiellement ou entièrement remboursés par l'organisme. L'application de ces dépassements impacte l'accès aux soins aux individus pour des raisons financières. *A contrario*, les médecins signataires à l'OPTAM reçoivent en retour une diminution des cotisations sociales.

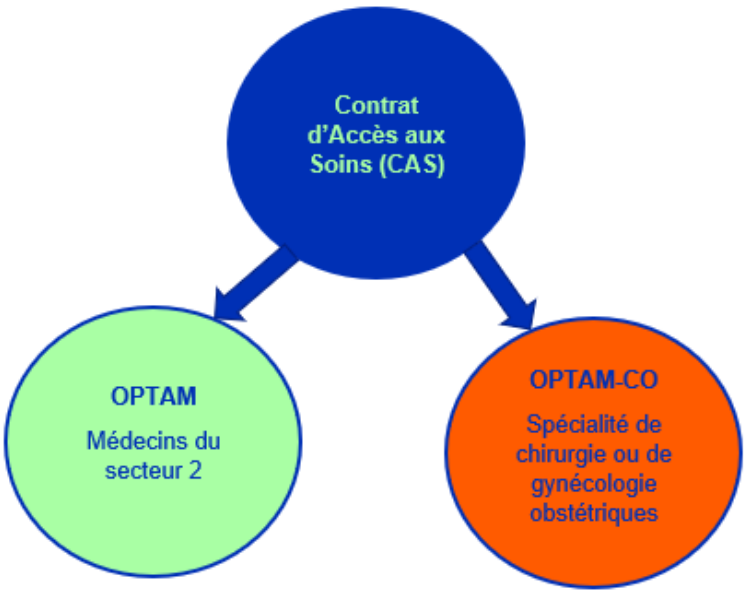


FIGURE 1.11 : Evolution du CAS à l'OPTAM et l'OPTAM-CO.

L'OPTAM est destinée aux médecins du secteur 2. Les médecins de secteur 1 et 3 ne sont pas concernés par cette option. Quant à l'OPTAM-CO, cette option est destinée aux médecins spécialisés en chirurgie ou pratiquant des actes de gynécologie obstétriques. Ils peuvent néanmoins, s'ils le souhaitent, adhérer à l'OPTAM, mais pas aux deux options. La différence entre ces deux dispositifs, pour la chirurgie et l'obstétrie, porte sur la majoration des forfaits modulables, ainsi que la rémunération qui dépend de l'option choisie pour les chirurgiens et obstétriciens. Le tableau 1.1 détaille les différents secteurs et leurs principes de remboursement.

Table 1.1 : Tableau récapitulatif des différents secteurs de conventionnement des médecins.

Secteurs	Conventionné ?	OPTAM	Dépassements d'honoraires
Médecins du secteur 1	Oui	Non signataire à l'OPTAM	Pas de dépassement d'honoraires. Ils sont alignés sur le tarif conventionnel qui est la Base de Remboursement de la Sécurité Sociale.
Médecins du secteur 2	Oui	Signataire à l'OPTAM	Dépassements d'honoraires maitrisés. Ils ne s'alignent pas sur le tarif conventionnel.
Médecins du secteur 3	Non	Non signataire à l'OPTAM	Honoraires libres. Aucune convention signée avec l'Assurance Maladie.

Les médecins sont dits **conventionnés** lorsque un accord est signé avec l'Assurance Maladie concernant la maîtrise des dépassements d'honoraires appliqués. La liste des médecins signataires à l'OPTAM/OPTAM-CO est disponible dans l'Annuaire du site de l'Assurance Maladie (2021). Cette liste est obtenue selon divers critères comme le lieu de résidence, la spécialité ou le type d'honoraires. Ils sont identifiables selon la mention "Honoraires avec dépassements maîtrisés (OPTAM)".

Le contrat responsable et solidaire

Le 13 août 2004, une loi a été introduite afin de définir les contrats dits « responsables ». L'objectif de l'introduction de cette loi est multiple :

- réduire le reste à charge des assurés et maîtriser l'évolution des pratiques tarifaires ;
- favoriser la prise en charge des assurés auprès des praticiens faisant partie intégrante du Dispositif de Pratique Tarifaire Maîtrisée (DPTAM) et non auprès de ceux qui ne l'ont pas intégré ;
- améliorer la prise en charge minimum pour certains postes de soins comme le dentaire ou bien l'optique.

Afin que ces objectifs soient effectifs, les organismes assureurs bénéficient d'avantages fiscaux et sociaux. De plus, pour l'employeur d'une entreprise, les contrats collectifs responsables sont exonérés de cotisations sociales pour la charge de l'employeur et sont déductibles de l'impôt sur le revenu pour la charge de l'employé. Puis, cette loi a subi une modification lors de la parution du Décret le 18 novembre 2014. Ce décret prévoit les critères que doivent respecter les contrats complémentaires santé responsables. Un contrat responsable doit donc (Vernin-Biancale, 2020) :

- prendre en charge au minimum :
  - l'intégralité du ticket modérateur pour l'ensemble des actes santé remboursés par la Sécurité Sociale (consultations du médecin, pharmacie, frais de laboratoire, optique, etc.) sauf exceptions (cures thermales, homéopathie, médicaments dont le service médical est classé comme faible ou modéré),
  - l'intégralité du forfait journalier hospitalier sans aucune limite de durée d'hospitalisation,
- prendre en charge les dépassements d'honoraires des médecins n'ayant pas adhérés à l'OPTAM dans la limite de 100% du tarif de la Sécurité Sociale. Cette prise en charge s'effectue à condition que ce remboursement soit inférieur d'au moins 20% au remboursement susceptible d'être perçu par un médecin adhérent à l'OPTAM.

Concernant l'optique, des planchers et plafonds sont mis en place en fonction des types de correction. Une limite de la fréquence du renouvellement des montures est notamment instaurée et limitée à un an pour les mineurs, une fois tous les deux ans pour le reste de la population, ou en cas d'évolution de la vue. Le tableau 1.2 récapitule les plafonds et planchers appliqués en optique.

TABLE 1.2 : Tableau récapitulatif des planchers et plafonds en optique pour un contrat individuel complémentaire santé responsable selon la correction nécessaire.

Type de corrections / Monture	PLANCHER	PLAFOND
Monture	150 € (100 € pour un contrat collectif)	
Verres simples + monture	50 € (100 € pour un contrat collectif)	470 €
Verres complexes + monture	200 €	750 €
Verres très complexes + monture	200 €	850 €
Verre simple + Verre complexe + monture	125 €	610 €
Verre simple + Verre très complexe + monture	125 €	660 €
Verre complexe + Verre très complexe + monture	200 €	850 €

Pour illustrer ces critères, des exemples de contrat non-responsable et responsable sont présentés ci-après afin de comprendre de manière plus concrète les critères demandés. Le tableau 1.3 sera choisi comme base de garanties d'un produit qu'une mutuelle souhaite commercialiser. Elle souhaite savoir si ce contrat est responsable ou non, et les raisons pour lesquelles il ne l'est pas.

TABLE 1.3 : Garanties d'un contrat individuel complémentaire santé non-responsable.

Postes	**Remboursements y compris SS**
HOSPITALISATION	
Honoraires chirurgicaux	**250 % BR**
Forfait journalier	**De type frais réels & limité à 110 jours par an**
Frais de séjour	**80 % BR**
Frais de transport	100 % BR
SOINS COURANTS	
Consultation généraliste	**250 % BR**
Consultation spécialiste	**250 % BR**
Auxiliaires médicaux	100 % BR
Analyses	100 % BR
Pharmacie	100 % BR
OPTIQUE	
Montures + verres	**650 € par an par bénéficiaire**
Lentilles	350 € par an par bénéficiaire
Chirurgie réfractive	800 € par an par bénéficiaire
DENTAIRE	
Prothèses acceptées (par la SS)	100 % BR
Orthodontie acceptée (par la SS)	100 % BR
Soins dentaires	100 % BR
AUTRES	
Cures thermales	100 % BR
Prévention	100 % BR
Médecine douce	40 € par an

Après analyse des garanties proposées, il est évident que ce contrat n'est pas responsable. En effet :

- les remboursements des honoraires médicaux et des consultations généralistes / spécialistes ne sont pas différenciés selon si l'exécutant adhère ou non au DPTAM ;
- il ne devrait y avoir aucune limite de durée pour la prise en charge intégrale du forfait journalier ;
- la prise en charge du montant des montures et verres optiques doit être détaillée pour chaque type de verres, dans le cadre d'un contrat collectif.

Le tableau 1.4 correspond au détail des garanties corrigées afin que le contrat devienne responsable.

TABLE 1.4 : Garanties d'un contrat individuel complémentaire santé responsable.

Postes	**Remboursements y compris SS**
HOSPITALISATION	
Honoraires chirurgicaux	250 % BR OPTAM / 200 % Non-OPTAM
Forfait journalier	De type frais réels & pas de limite de durée
Frais de séjour	100 % BR
Frais de transport	100 % BR
SOINS COURANTS	
Consultation généraliste	250 % BR OPTAM / 200 % Non-OPTAM
Consultation spécialiste	250 % BR OPTAM / 200 % Non-OPTAM
Auxiliaires médicaux	100 % BR
Analyses	100 % BR
Pharmacie	100 % BR
OPTIQUE	
Montures + verres	
dont 2 verres simples	360 € par an par bénéficiaire
dont 2 verres complexes	600 € par an par bénéficiaire
dont 2 verres très complexes	700 € par an par bénéficiaire
Lentilles	350 € par an par bénéficiaire
Chirurgie réfractive	800 € par an par bénéficiaire
DENTAIRE	
Prothèses acceptées (par la SS)	100 % BR
Orthodontie acceptée (par la SS)	100 % BR
Soins dentaires	100 % BR
AUTRES	
Cures thermales	100 % BR
Prévention	100 % BR
Médecine douce	40 € par an

Les organismes complémentaires santé ont tout intérêt à appliquer ces critères à leurs contrats. En effet, s'ils ne respectent pas ces conditions, l'organisme ne bénéficie plus d'aides fiscales et sociales. La mise en place de ce nouveau contrat responsable s'accompagne notamment d'une nouvelle réforme, votée par les décrets de 2018 et 2019 : la réforme 100% santé.

La réforme 100% santé

La réforme 100% santé constitue une évolution importante pour le système de santé français. Sa mise en place, effective lors de la rédaction de ce mémoire, a eu un effet non négligeable sur le secteur de la santé, notamment sur la consommation en soins des assurés, mais aussi pour les professionnels exécutants et les organismes assureurs. Suite à cette réforme, un certain nombre de points importants au niveau de la gestion du système a été modifié. Ces modifications complètent les applications définies initialement pour les contrats responsables.

En effet, il a été observé que le reste à charge était très important pour trois postes : optique, dentaire et audiologie. Cette tendance pourra être observée et étudiée au sein de ce mémoire. L'objectif de la mise en place de la réforme 100% santé est donc d'améliorer l'accès aux soins, avec la diminution du reste à charge, mais aussi de responsabiliser les consommateurs et les professionnels de santé. Elle concerne donc l'optique, le dentaire et l'audiologie. La mise en place de cette réforme s'établie

progressivement. Les différentes étapes de la mise en place auront lieu entre le 1^{er} janvier 2019 et le 1^{er} janvier 2023, et diffèrent selon le poste de soins. L'introduction de la réforme débute :

- au 1^{er} janvier 2019 pour l'audiologie ;
- au 1^{er} avril 2019 pour le dentaire ;
- au 1^{er} janvier 2020 pour l'optique.

Le schéma 1.12 ci-dessous résume cette chronologie.

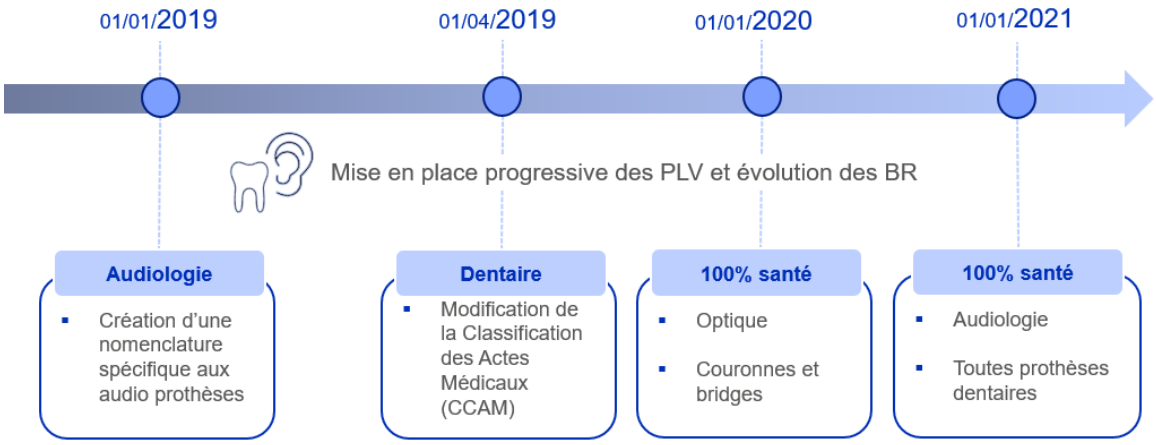


FIGURE 1.12 : Chronologie de la mise en place progressive de la réforme 100% santé.

Concernant les modifications appliquées pour les trois postes, la réforme 100% santé a permis la création de panier de soins (dont le panier 100% santé), la modification de la nomenclature et la création de Prix Limites de Vente (PLV) des équipements (optiques et auditifs) ou des prothèses dentaires. Pour l'optique et l'audiologie, ce sont deux paniers de soins qui ont été créés :

- **panier A :** Il s'agit du panier 100% santé. Les équipements optiques et auditifs intégrés dans ce panier sont intégralement pris en charge par la Sécurité Sociale et par les contrats complémentaires santé responsables (y compris la CSS (c.f. section 1.1.3)) ;
- **panier B :** Il s'agit du panier aux tarifs libres. Les bases de remboursement ont été cependant revues à la baisse de cinq centimes avec la réforme 100% santé.

Pour les prothèses dentaires, trois paniers de soins ont été créés :

- **panier 1 :** Il s'agit du panier 100% santé. Les actes/ prothèses dentaires intégrées dans ce panier sont intégralement pris en charge par la Sécurité Sociale et par les contrats complémentaires santé responsables (y compris la CSS (c.f. section 1.1.3)). Des honoraires de facturation sont notamment appliqués à chacun des actes de ce panier ;
- **panier 2 :** Il s'agit du panier pour lequel des prix limites de vente sont intégrés sur la période de mise en place progressive de la réforme. Néanmoins, le reste à charge n'est pas obligatoirement nul pour ce panier ;
- **panier 3 :** Il s'agit du panier aux tarifs libres, sans prix limite de vente.

Etant donnée la mise en place progressive de la réforme, la mesure du reste à charge zéro ne sera effective qu'à partir de 2020, d'après le calendrier. Avant cette date, le panier 100% santé peut inclure des dépenses de soins et d'équipement avec un reste à charge non nul.

Plusieurs mesures concernent aussi les professionnels de santé. Dans un premier temps, un devis, incluant au moins une prestation du panier 100% santé, devra être établi par le professionnel de santé. L'assuré n'a pas l'obligation de choisir cette prestation. De plus, comme vu dans les descriptions des paniers de soins, des prix limites de vente sont appliqués. Les équipements et soins concernés par un prix limite de vente ne pourront pas être vendus à un prix supérieur à la limite. Si l'assuré souhaite un équipement/soin de valeur plus élevée, il pourra alors les choisir parmi ceux proposés dans le panier "Tarifs libres".

Enfin, les montants des montures et équipements optique, définis dans le tableau 1.2, ont été impactés par la réforme 100% santé. Pour les contrats individuels complémentaires santé et responsables, les planchers ne sont pas modifiés. Néanmoins,les montants des plafonds ont été rabaissés de 50 €. Enfin, après réforme, la monture peut être remboursée au maximum à hauteur de 100 €, contre 150 € avant la mise en place de la réforme 100% santé.

Détaillons à présent le dispositif solidaire mentionné au sein de cette partie et bénéficié par une partie de la population française : la Complémentaire Santé Solidaire (CSS).

Un dispositif de solidarité récemment naissant : la CSS et la PUMA

La Complémentaire Santé Solidaire (CSS) résulte de la fusion de la Couverture Maladie Universelle Complémentaire (CMU-C) et de l'Aide à la Complémentaire Santé (ACS) au 1^{er} novembre 2019 (c.f. graphique 1.13). Avant la fusion, environ 7,4 millions de personnes étaient couvertes par ces deux dispositifs, contre 10 à 12 millions aujourd'hui après la naissance de la CSS.

La CMU-C est une complémentaire santé gratuite pour l'adhérent, et dont l'adhésion est conditionnée au niveau de ses ressources, de son lieu de résidence et de la composition de son foyer. Elle vient compléter la Couverture Maladie Universelle (CMU) de base. La gestion de ces contrats par les organismes est assurée sur la base du volontariat. L'ensemble des adhérents à la CMU-C a basculé de façon automatique et gratuite vers l'adhésion à la CSS. Quant à l'ACS, il s'agit d'une aide du régime général de santé versée à l'adhérent pour payer sa complémentaire santé s'il dispose de ressources qui satisfassent la condition ci-dessous

$$Plafond_{CMU} < Ressources_{adhérent} < Plafond_{CMU} \times (1 + 35\%).$$

Le montant de cette aide dépend de l'âge du bénéficiaire.

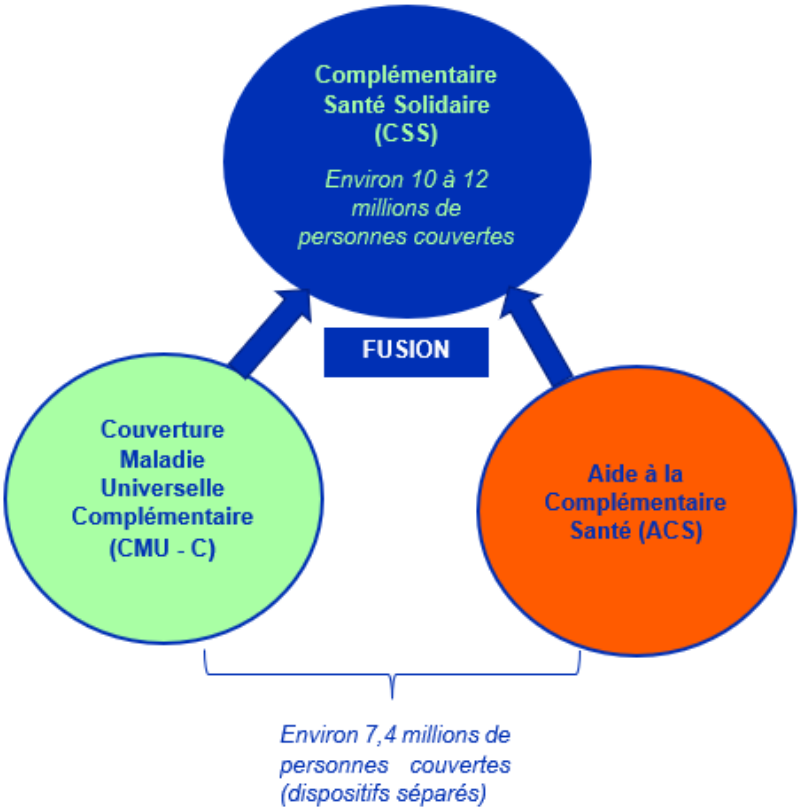


FIGURE 1.13 : Fusion de la CMU-C et de l'ACS le 1^{er} novembre 2019.

Il existe notamment un dispositif de solidarité de base qui est la Protection Universelle Maladie (PUMA), remplaçant la Couverture Maladie Universelle (CMU) depuis le 1^{er} janvier 2016 (c.f. graphique 1.14). La CMU a été initialement créée afin de respecter l'objectif principal du système d'assurance santé français qui est l'accès aux soins pour tous. Elle permet le remboursement de frais de santé des résidents réguliers français aux faibles ressources à condition de ne pas être couvert par un autre régime obligatoire.

Remarque : Ces dispositifs n'intégreront pas le périmètre d'étude de ce mémoire et ne seront donc pas traités par la suite. Néanmoins, il était important de souligner leur existence au vu de leur importance pour l'accès aux soins à tout individu.

L'ANI et les contrats obligatoires à titre collectif

D'après l'Article L. 911-7 du Code de la Sécurité Sociale et de l'Accord National Interprofessionnel (ANI), "Tout employeur du secteur privé, entreprise et association, a l'obligation depuis le 1^{er} janvier 2016 de proposer une couverture complémentaire santé collective à ses salariés, en complément des garanties de base de l'Assurance Maladie Obligatoire de la Sécurité Sociale."

Initialement, les objectifs de l'ANI se concentraient sur la sécurité de l'emploi et du parcours professionnel des salariés lors de contrat au sein de l'entreprise, mais aussi lors de rupture de contrat professionnel. En effet, L'ANI assure la portabilité des droits du salarié. Cela signifie que la complémentaire santé, bénéficiée par le salarié avant son départ de l'entreprise, est maintenue à titre gratuit pendant une durée maximale de 12 mois ou jusqu'à l'attribution d'un nouveau poste avant les 12 mois.

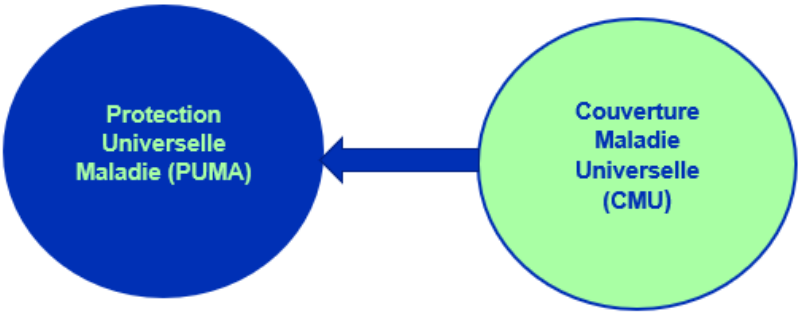


FIGURE 1.14 : Passage de la CMU à la PUMA au 1^{er} janvier 2016.

Depuis 2016, la complémentaire santé collective est obligatoirement proposée par les employeurs pour les entreprises du secteurs privés, mais aussi pour le salarié. Seules les garanties optionnelles, souscrites par l'employeur sont à caractère facultatif pour le salarié.

1.1.4 Le principe de remboursement

Comme vu précédemment, l'assuré bénéficie d'une prestation santé. Par la suite, différents acteurs interviennent lors de la prise en charge de cette prestation. Quelles sont les caractéristiques de cette prise en charge ?

Dans un premier temps, l'Assurance Maladie Obligatoire rembourse différentes prestations de soins selon des taux de remboursement bien définis. Celui-ci diffère selon le type d'actes (consultations généralistes, pharmacie, etc.), mais également de la situation de chaque assuré. En effet, pour des personnes appartenant au Régime Alsace Moselle, les taux de remboursement sont supérieurs, et sont proches de 90%. De même, pour des personnes bénéficiant du dispositif de solidarité, la CSS, ou bien en état d'Affections Longues Durées (ALD), les taux de remboursement peuvent atteindre les 100%. Ce taux de remboursement est appliqué à une base de remboursement définie grâce à une classification des prestations santé selon une nomenclature générale, comme la Classification Commune des Actes Médicaux (CCAM). Cela permet donc d'associer une base de remboursement à chaque acte classifié. Cette base de remboursement correspond au tarif fixé par la sécurité sociale. De plus, la prise en charge des prestations est souvent exprimée comme un pourcentage de la base de remboursement. Le principe de remboursement est présenté dans le schéma 1.15 avec

$$Dépassements\ d'honoraires = Frais\ réels - BRSS,$$

$$RSS(Remboursement\ de\ la\ Sécurité\ Sociale) = BRSS \times Taux\ de\ remboursement\ de\ la\ SS,$$

$$TM(Ticket\ Modérateur) = BRSS - RSS - Participation\ forfaitaire,$$

$$RAC(Reste\ à\ Charge) = Frais\ réels - Remboursement\ complémentaire\ y\ compris\ SS.$$

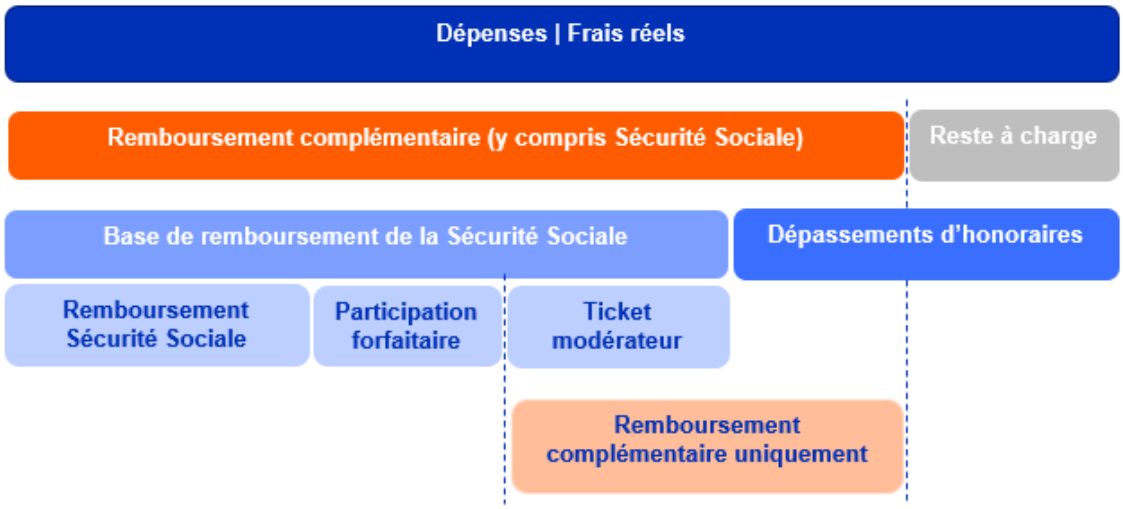


FIGURE 1.15 : Le principe de remboursement des prestations santé.

Le taux de prise en charge par l'Assurance Maladie Obligatoire est compris entre 0% et 100%. Pour les actes non remboursés par l'AMO, tels que les frais d'accompagnant en cas d'hospitalisation, les chambres particulières ou encore la médecine douce, il est courant de trouver des prises en charge par l'AMC dans les tableaux de garanties. Le reste à charge de l'assuré est lui aussi compris entre 0% et 100% de sa dépense en fonction des remboursements effectués ou non par l'AMO et l'AMC. Le niveau de remboursement est défini selon les garanties du contrat complémentaire santé auquel l'adhérent à souscrit. Ces niveaux de garanties diffèrent selon l'organisme d'assurance. Par exemple, Actélior prend en compte trois niveaux de couverture pour les complémentaires santé pour lesquelles les conditions de remboursements sont différentes :

- **entrée de gamme :** Les remboursements sont limités au ticket modérateur et ne prennent pas en compte les dépassements d'honoraires des professionnels de santé de secteur 1 et 2 (Rappel : ce sont des professionnels de santé dit "conventionnés") ;
- **milieu de gamme :** Ce produit prend en charge la majorité des dépenses peu remboursées par la Sécurité Sociale comme l'optique et le dentaire, mais aussi une partie des dépassements d'honoraires ;
- **haut de gamme :** Ce produit prend en charge la majorité des dépassements d'honoraires et les dépenses peu voire non prises en charge par la Sécurité Sociale (médecine douce, diététique, etc.).

Une participation forfaitaire est notamment appliquée au processus de remboursement. Elle est soustraite du montant de remboursement de la Sécurité Sociale. De plus, une partie appelée ticket modérateur fait partie du système de remboursement. Il correspond à la différence de la base de remboursement et du montant remboursé par la Sécurité Sociale. Enfin, les dépassements d'honoraires pratiqués par les professionnels de santé peuvent entraîner un reste à charge pour l'assuré. Ce résidu dépend des secteurs auxquels appartiennent les professionnels de santé et donc du niveau de ce dépassement.

1.1.5 Le principe de la tarification d'un produit complémentaire santé et ses particularités

Le marché de la complémentaire santé en France couvre 94% de la population par le biais des organismes cités précédemment. En contrepartie des prestations que versent ces organismes aux adhérents pour la prise en charge de leurs soins, les adhérents leur versent une prime. Cette prime est le résultat d'une analyse statistique de la consommation santé des assurés. Cette partie introduit la tarification d'un produit complémentaire santé. La partie théorique de la tarification sera plus amplement détaillée dans la section 3.1 consacrée à la réalisation de la modélisation.

Décomposition de la prime commerciale

Pour obtenir le tarif final appelé **prime commerciale**, celle que l'adhérent va finalement payer, plusieurs facteurs doivent être pris en compte dans le calcul (Lazic, 2020) :

- **la prime pure :** Elle correspond au risque auquel est confronté l'adhérent et se calcule par le biais de statistiques telles que la fréquence et le coût ;
- **la marge technique :** Il s'agit d'une partie tarifaire supplémentaire permettant de pallier les erreurs induites par des hypothèses prises préalablement. Elle permet notamment à l'organisme de se constituer des fonds propres ;
- **les chargements :** Il existe les chargements de gestion et les chargements d'acquisition, qui peuvent être additionnels ou proportionnels à la prime pure. Ils représentent tous les frais liés à la vie du produit ;
- **les taxes :** En santé, tous les contrats intègrent une taxe de solidarité additionnelle de 13,27% pour les contrats solidaires-responsables et de 20,27% pour les contrats non-solidaires et non-responsables ;
- **annexe :** Il s'agit de frais liés à des garanties annexes, comme le coût d'une assistance par exemple.

Méthodes de calculs pour la prime pure

La prime pure d'un produit d'Assurance Maladie Complémentaire peut être obtenue selon plusieurs méthodes :

- 1) La méthode Coût × Fréquence (majoritairement utilisée) avec l'application de coefficients selon différents critères choisis ;
- 2) Une prime pure forfaitaire (dans le cas où le nombre de données est faible).

Pour la première méthode, comme son nom l'indique, la prime pure se décompose selon une fréquence moyenne et un coût moyen

$$Prime\ pure = Fréquence\ moyenne \times Coût\ moyen,$$

avec

$$Fréquence\ moyenne = \frac{Nombre\ de\ sinistres}{Exposition\ totale},$$

et

$$Coût\ moyen = Montant\ moyen\ des\ sinistres.$$

La prime Coût × Fréquence permet d'expliquer la consommation santé selon ces deux facteurs. Cependant, les hypothèses prises pour cette méthode sont fortes. L'indépendance entre ces deux facteurs est généralement admise mais peu réaliste en pratique. L'hypothèse d'une distribution du risque, identique pour tous les individus, est notamment peu réaliste en pratique puisque le risque varie en fonction de divers paramètres comme l'âge, la Catégorie Socio-Professionnelle (CSP), etc. Néanmoins, nous verrons dans le chapitre 3 que cette indépendance est vérifiée dans le cadre de l'étude.

Il est donc possible d'appliquer à cette prime de référence des coefficients afin de créer une segmentation du tarif selon divers critères. Ces critères tarifaires dépendent du contrat souscrit par l'adhérent (collectif ou individuel). En collectif, il s'agit donc d'un contrat appliqué pour un groupe de personnes au sein d'une entreprise ou bien une catégorie bien définie. Les variables comme l'âge, le code NAF (Nomenclature d'Activités Française) de l'entreprise, le nombre d'enfants à charge, la situation familiale et bien d'autres sont choisis. *A contrario*, pour un contrat individuel, l'âge de l'assuré, la CSP et la zone géographique sont demandés. L'expression de la prime pure obtenue pour un contrat individuel est donc la suivante

$$Prime\ pure = Prime\ de\ référence \times Coefficent_{Age} \times Coefficient_{Région} \times Coefficient_{CSP}.$$

Cependant, ces caractéristiques sont détenues par l'assuré lui-même (conscient de son âge, son risque, etc.) mais pas par l'assureur. Il s'agit du phénomène d'antisélection.

La tarification et l'antisélection

L'antisélection désigne l'asymétrie d'information entre les assureurs et les assurés. En effet, les assurés disposent des informations sur leurs propres risques non accessibles aux assureurs. Sans ces informations, l'assureur se retrouve dans l'incapacité de déterminer des primes en fonction des risques. Les primes correspondent donc, comme vu précédemment, au coût moyen des sinistres des individus pour un même contrat souscrit. En cas d'absence d'information sur les risques, les mauvais risques seront donc particulièrement avantagés, contrairement aux bas risques qui auront une prime finalement plus élevée du fait du coût moyen faible des sinistres. La catégorie de risque à laquelle appartient l'assuré ne peut donc pas être connue.

Cette asymétrie d'information entre assureurs et assurés et ce manque d'informations pour les assureurs est un frein pour une tarification plus juste. C'est pourquoi faire intervenir des données externes en Open Data permettrait d'obtenir des informations précises sur des profils similaires et donc d'éviter ce phénomène d'antisélection.

1.2 La valeur ajoutée de l'Open Data en santé

1.2.1 Objectif de l'Open Data

L'Open Data désigne un ensemble de données réutilisables et accessibles librement et gratuitement par tous. Ces données peuvent être utilisées dans le cadre d'études, de recherches, de statistiques et pour divers domaines d'activités. L'ouverture de ces données publiquement a pour objectif d'accélérer la transformation numérique des secteurs d'activités mais aussi des organisations publiques et des administrations. Cependant, il est important de noter que ce sont des données anonymisées. Elles ne comportent donc aucune information personnelle sur les individus.

Les données publiques de l'Etat français sont disponibles en libre accès sur le site Datagouv (2021). Sur celui-ci, chaque individu peut les utiliser, mais aussi les améliorer et les partager afin de pouvoir répondre à des questions, dans le cadre de décision, de recherches, ou à titre informatif. Ce site est alimenté principalement par diverses organisations et institutions publiques (ministères, collectivités, établissements scolaires). De plus, ces bases Open Data sont disponibles pour divers secteurs comme les transports, l'environnement, l'économie, la finance, les énergies, l'agriculture, ... mais aussi la santé.

1.2.2 L'Open Data en santé

Aujourd'hui, un grand nombre de données sont collectées dans le domaine de la santé. L'évolution grandissante de cette collecte de données est aujourd'hui un progrès considérable pour la santé. En effet, même si elles sont peu utilisées aujourd'hui, elles seront utilisées en étroite lien avec l'Intelligence Artificielle (IA). Ces données permettront d'améliorer la santé et le travail des médecins mais aussi la vie des patients par anticipation et détection de maladies telles que les cancers.

Ces données santé en open data ont été utilisées plus particulièrement l'année précédente, lors de la crise sanitaire due à la propagation du virus Covid-19. De nombreuses plateformes digitales recueillant ces données provenant de plusieurs sites ont vu le jour. Ceci a permis d'informer l'ensemble de la population sur la progression de la pandémie, et d'appuyer les restrictions et mesures sanitaires avec des données et informations précises les justifiant.

Outre l'Intelligence Artificielle, ces données santé en open data peuvent être utilisées pour la digitalisation des processus d'assurance santé notamment la tarification, mais aussi pour comprendre, au niveau national, le comportement des patients et leur consommation.

1.2.3 L'Open Data en assurance santé

L'Assurance Maladie réalise des études sur le système de soins, l'évolution de la consommation de soins, le suivi des dépenses, l'activité des professionnels de santé, les accidents du travail ainsi que les maladies professionnelles. Elle réalise ces études afin de connaitre les enjeux pour améliorer la qualité du système de santé et maitriser les dépenses. Elle crée notamment des données résumant ces études et les met à disposition sur le site Ameli (2021) en Open Data et principalement au sein du système informationnel de l'Assurance Maladie, le SNIIRAM, géré par la CNAM.

Le SNIIRAM a été créé en 1998 pour collecter de nombreuses informations anonymes et enrichir le système de santé de connaissance à propos des dépenses de l'ensemble des régimes d'Assurance Maladie, mais aussi de la mise en œuvre et de l'évaluation des politiques de santé. Ces données, collectées depuis

2002, sont disponibles, comme toutes données en Open-Data, à tout individu souhaitant réaliser une étude en santé, analyser ou améliorer ses bases de données. Elles concernent les remboursements effectués par l'ensemble des régimes d'assurance maladie pour les soins du secteur libéral. Le schéma 1.16 présente ci-dessous la structure du SNIIRAM.

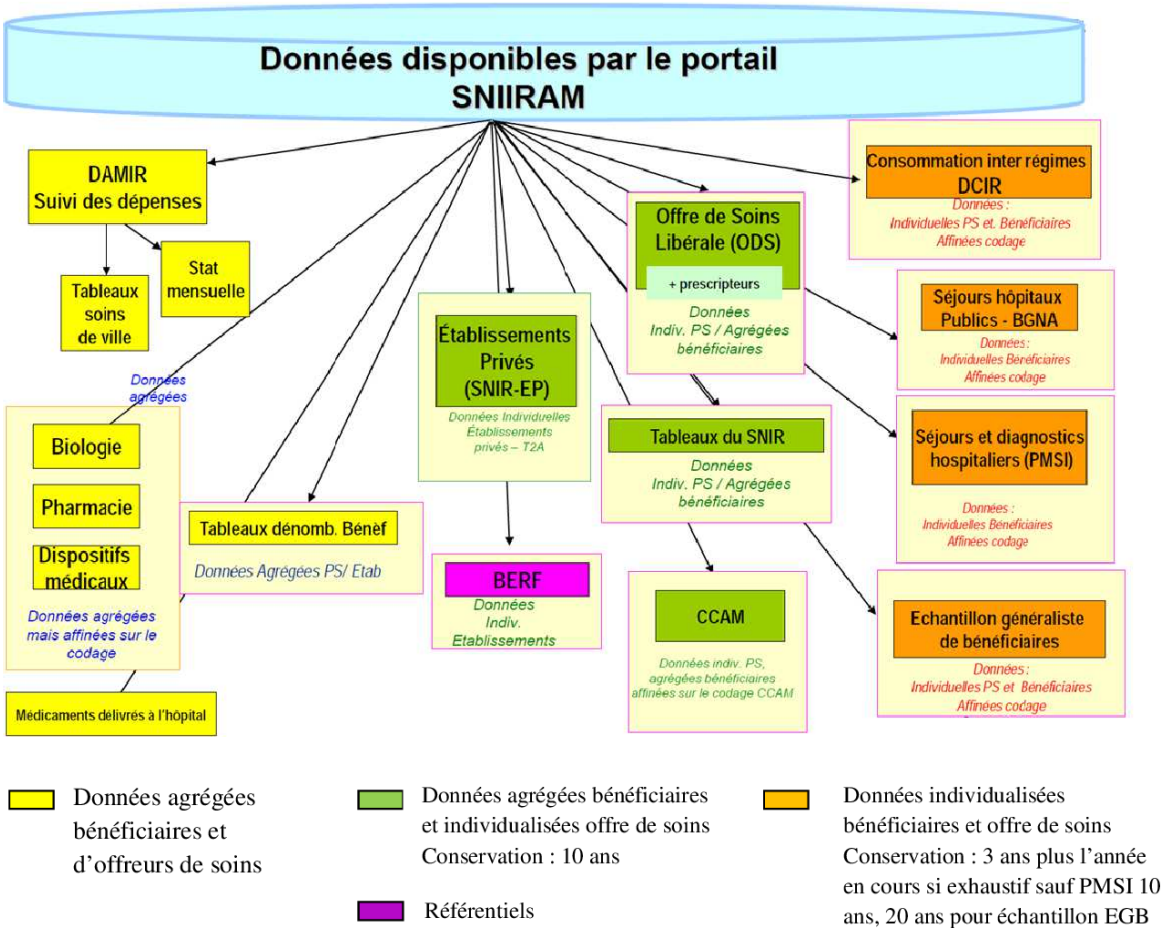


FIGURE 1.16 : Structure du SNIIRAM.

1.2.4 Protection des données par l'anonymisation et RGPD

Le Règlement Général sur la Protection des Données (RGPD), a pour objectif de garantir la protection des données. Il serait donc cohérent de penser que cette protection va à l'encontre du principe de l'Open Data. Cependant, pour que ces données puissent être disponible en libre accès, l'intégration préalable d'une notion de protection de la vie privée lors de l'enregistrement des données et informations personnelles pourrait être effectuée afin d'éviter tout retraitement. Cependant, la Loi Lemaire, permet de contourner ce problème. En effet, ces bases de données en accès public gratuitement se doivent d'être anonymisées afin de protéger les informations personnelles de chaque individu. Cette anonymisation s'effectue de manière à ce qu'aucune personne ne puisse déterminer de nouveau la base originale avec l'ensemble des informations. Pour cela, de nombreux regroupements sont souvent effectués, notamment au niveau de l'adresse (regroupement de départements), d'âge (regroupement en classe d'âges). Au final, nous obtenons des lignes agrégées où le nombre d'individus par ligne est inconnu, rendant cette anonymisation efficace. C'est le cas de la base de données santé Open Damir,

base principale de l'étude, dont les caractéristiques sont présentées ci-après.

Dans le cadre de ce mémoire, nous verrons que l'anonymisation de ces bases de données volumineuses complexifient le traitement de ces dernières. Pour les différents travaux de ce mémoire, les bases de données santé Open Damir, en libre accès sur le site DATAGOUV (2021), ont été sélectionnées. Détaillons à présent les différents retraitements effectués sur ces bases de données, ainsi que la manière dont la volumétrie des bases a été gérée.

Chapitre 2

Etude de la base de données santé Open Damir

L'objectif de ce mémoire est d'utiliser les bases Open Damir dans le cadre de divers travaux actuariels détaillés aux chapitre 3 et chapitre 4. Avant toute chose, cette base doit être analysée et retraitée afin de comprendre son contenu et de l'adapter aux prochaines études de ce mémoire.

2.1 Description générale de la base

Lors de la présentation de la base d'étude, la base Open Damir, de nombreux éléments seront détaillés comme le périmètre. Cette base couvre plusieurs axes d'analyses, plusieurs régimes, et est disponible pour chaque mois de 2009 à 2019.

2.1.1 Périmètre de l'étude

Pour notre étude, des hypothèses ont été prises afin d'en déterminer le périmètre. Ces hypothèses sont importantes puisqu'elles permettent de cadrer l'ensemble des traitements effectués et de pouvoir les interpréter correctement. Notre étude portera donc uniquement sur :

- Régimes : régime général. Attention, les assurés appartenant au Régime des Indépendants (RSI) ne sont pas comptabilisés au sein du régime général pour les années 2018 et 2019 ;
- Année de survenance des prestations santé et leur remboursement (c.f. section 2.2.1 pour l'explication de ce choix) :
  - prestations survenues en 2018 remboursées en 2018,
  - prestations survenues en 2018 remboursées en 2019,
  - prestations survenues en 2019 remboursées en 2019.
- Les personnes affiliées aux dispositifs et régimes suivants sont exclues de l'étude car elles ne sont habituellement pas intégrées dans la tarification santé d'Actélior :
  - les personnes bénéficiant de la CSS (CMU),

  - les personnes faisant partie intégrante du régime Alsace Moselle.

- Prestations prises en compte :
  - seules les prestations de santé des particuliers seront étudiées,
  - l'ensemble des remboursements de l'Assurance Maladie destinés aux professionnels de santé ne seront pas considérés. Les remboursements de ces acteurs sont très élevés et ne reflètent pas la consommation en santé des particuliers (Assurance Maladie Obligatoire, juillet 2021).

A présent, nous allons aborder le cœur du sujet avec la présentation de la base de données qui est le support principal de cette étude.

2.1.2 Présentation de la base Open Damir

Source et extraction de la base Open Damir

Open Damir est une base de données santé mise à disposition en Open Data lors du premier Hackathon, en janvier 2015. En général, cet événement a pour objectif de réunir plusieurs spécialistes afin de travailler sur un projet informatique et de développement numérique. Dans le cadre de l'Hackathon de janvier 2015, de nombreux travaux de recherches, de retraitements de la base ont été réalisés. L'objectif était la compréhension de la base de données en l'exploitant au maximum et de manière approfondie.

Cette base de données a été créée à partir des informations issues du Système National Inter Régimes d'Assurance Maladie (SNIIRAM). L'organisation du SNIIRAM a été explicitée auparavant dans la partie 1.2. La base Open Damir recense l'intégralité des remboursements effectués par l'Assurance Maladie, y compris les prestations hospitalières, tous régimes confondus. Cependant, comme explicité précédemment, nous n'étudierons que le Régime Général. Il s'agit d'une base de données complète, donnant des informations supplémentaires par rapport aux autres bases de données du SNIIRAM qui couvrent des champs différents et moins détaillés. Ces bases de données sont disponibles en accès gratuit et à tout public sur le portail du Système National des Données de Santé (SNDS) ou bien sur le site des données publiques de l'Etat français Datagouv.

La base de données Open Damir est destinée aux personnes souhaitant avoir des informations sur le domaine de la santé, que ce soit pour le milieu professionnel ou pour la réalisation de travaux de recherches statistiques et/ou scientifiques. Ils peuvent notamment utiliser cette base en tant que données exogènes afin de compléter leurs recherches déjà existantes, et d'explorer des axes complémentaires. Mais quels sont les axes d'études possibles de la base de données santé Open Damir ?

Description des axes d'analyses de la base Open Damir

Chaque ligne de prestation santé de la base Open Damir est décrite au total par 55 variables (catégorielles et quantitatives). Elles sont articulées autour de six axes d'analyse présentés par le schéma 2.1 (un axe complémentaire correspond au périmètre des prestations santé du dispositif CMU-C).

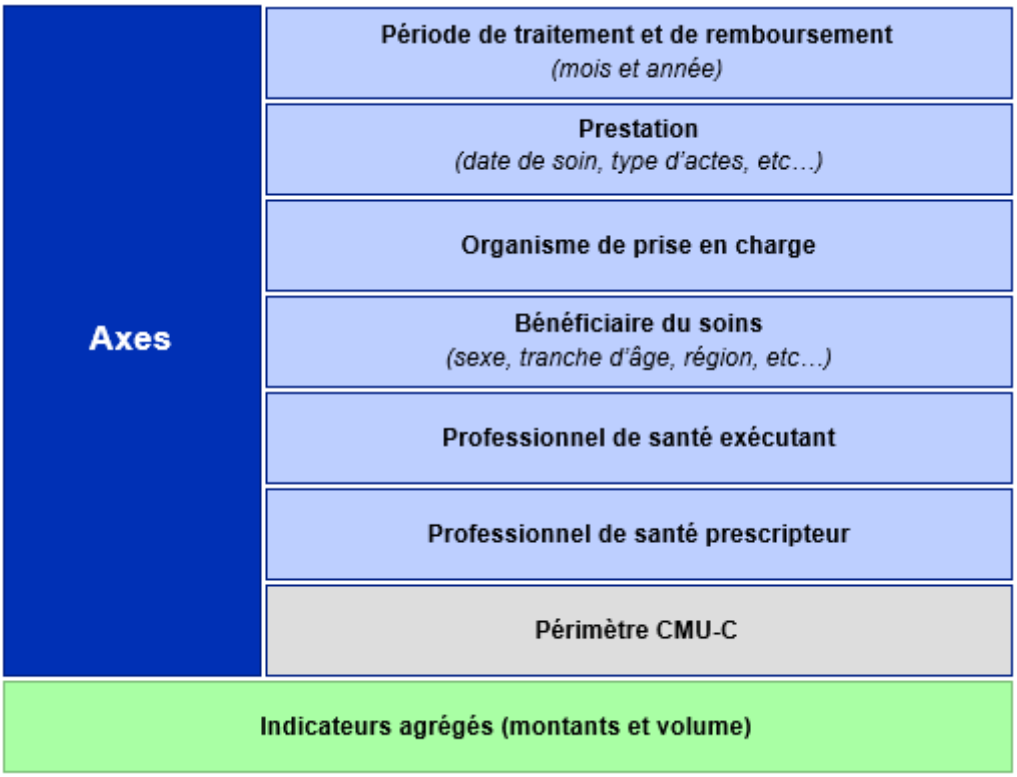


FIGURE 2.1 : Les différents axes de la base Open Damir.

Pour ces six axes d'analyses, de nombreuses variables de type qualitative sont disponibles. Elles donnent des précisions sur la nature de la prestation, le type de remboursement, le lieu de résidence et l'âge du bénéficiaire, mais aussi le type de spécialité du professionnel de santé exécutant et prescripteur, et bien d'autres encore. Le nom de chaque variable ainsi que leur description sont donnés dans le tableau A.7 disponible en annexe.

D'autres variables quantitatives viennent s'ajouter à ces six axes. Ce sont les indicateurs de montant et de volume des remboursements des prestations en santé. Les indicateurs de montants présents sont les suivants :

- la base de remboursement : il s'agit d'un montant défini par la Sécurité Sociale ;
- le montant de la dépense : ils correspondant aux frais réels, ce que coûte la prestation santé ;
- le montant du dépassement : après prise en charge d'une partie des frais réels par la Sécurité Sociale, le dépassement correspond à la différence des frais réels et du montant remboursé partiellement ou intégralement. La base Open Damir contient uniquement les informations sur les remboursement de la Sécurité Sociale. Aucune information sur les remboursements des organismes complémentaires n'est indiqué.
- Le montant du remboursement obligatoire : il correspond au montant remboursé par la Sécurité Sociale (c.f. figure 1.15).

Concernant les indicateurs de volumes de la base Open Damir, il existe :

- la quantité d'actes,
- le dénombrement d'actes,
- le coefficient global.

Ces indicateurs de montants et de volume de la base Open Damir ont notamment deux spécificités, que nous détaillerons juste après : Le régime étudié et l'anonymisation de la base de données conformément au RGPD.

Les indicateurs de montants et de volume en fonction du régime étudié

En effet, chaque indicateur peut être de deux types différents et est donc présent deux fois en tant que variable :

- les indicateurs non préfiltrés,
- les indicateurs préfiltrés.

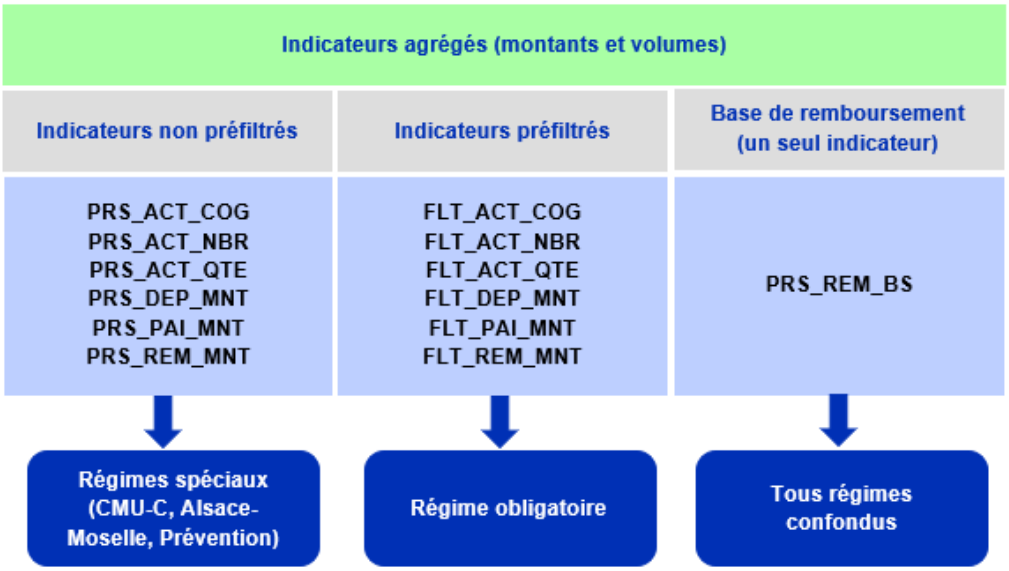


FIGURE 2.2 : Les deux types d'indicateurs de montant et de volume de la base Open Damir.

Les indicateurs non pré filtrés sont utilisés dans le cadre d'études des champs ne relevant pas du Régime Obligatoire (RO), mais plutôt du dispositif de la CMU, du régime spécial Alsace-Moselle, et dans le cadre de la Prévention. L'ensemble des noms de variables débute par « PRS » pour « Prestation » tandis que les indicateurs pré filtrés se distinguent par « FLT » au début du nom de la variable. Ces indicateurs pré filtrés concernent uniquement le Régime obligatoire (c'est-à-dire hors CMU, hors Alsace-Moselle, etc.). Le schéma 2.2 résume l'utilisation de ces deux types d'indicateurs. Nous avons donc décidé d'utiliser uniquement les indicateurs pré filtrés, en raison du périmètre choisi et afin d'éviter notamment les doubles comptes. Parmi ces indicateurs, seule la base de remboursement

est unique et n'est présente qu'une seule fois. En effet, elle est définie pour n'importe quel régime. Ce sont les taux de remboursement pour chaque libellé acte santé qui vont être modifiés selon le régime auquel est affilié le bénéficiaire du soin.

Afin de respecter le RGPD, les informations contenues dans la base Open Damir ont été anonymisées par l'intermédiaire de nombreux traitements. Par exemple, les âges et les régions ont été regroupés, en classe pour les âges et en nouvelle zones géographiques pour les régions, afin de rendre impossible l'identification du bénéficiaire du soin, ou même du professionnel de santé. Le découpage des zones géographiques a également évolué dans le temps : de 2009 à 2014, les bases comprenaient 9 zones alors qu'à partir de 2015, nous en constatons 13 correspondant en grande partie aux régions administratives. La correspondance de chacune de ces zones géographiques est disponible au sein du descriptif officiel des données de la base Open Damir, téléchargeable sur le site DATAGOUV (2021). Ensuite, les lignes ont été agrégées afin de rendre impossible l'identification d'un individu. En effet, au sein des bases Open Damir non retraitées, une ligne représentait la prestation santé, à une date t, et d'un montant n pour un unique bénéficiaire selon des informations qualitatives i. Après les regroupements effectués pour anonymiser la base, une ligne représente la somme des indicateurs de montants et de volumes correspondants à toutes les autres variables catégorielles.

Afin de permettre l'exploitation de ces bases, nous sommes confrontés à deux premières problématiques :

- la problématique de l'anonymisation des données que l'on vient de voir ci-dessus. En effet, l'anonymisation de la base de données ne permet pas aux utilisateurs de celles-ci de connaître le nombre de bénéficiaires concernés. Or, pour cette étude, le nombre de bénéficiaires est indispensable pour le calcul de la fréquence des prestations de santé, pour la mise en place de la tarification. Il faudra donc essayer de réaliser des rapprochements avec d'autres bases de données afin d'obtenir cette information. La méthode utilisée sera explicitée dans la section 2.2.4 du Traitement des données ;
- mais aussi le volume important de données. Un très grand volume de données peut, en effet, rendre le traitement de données très complexe, mais peut aussi avoir un impact lors de la réalisation de méthodes d'apprentissage statistiques.

La volumétrie importante de la base Open Damir : une des principales problématiques

Les données « Open Damir » sont disponibles sous forme de bases mensuelles correspondant à chaque mois de paiement par la Sécurité Sociale. 12 bases Damir sont donc générées pour une année de remboursement. Ces 12 bases sont rendues accessibles généralement courant le mois de juin de l'année $N + 1$ sur le site DATAGOUV (2021). A la date de rédaction de ce mémoire, nous disposons des bases jusqu'en 2020 (ces dernières ont été publiées mi-juillet 2021). Les noms des bases de données mensuelles sont préfixées par P de 2009 à 2014 puis préfixés par A à partir de 2015 (la différence de lettre est due au changement de la délimitation des régions administratives) et sont suffixées par l'année et le mois de remboursement des dépenses de santé. Cependant, lors de l'utilisation de la base pour effectuer des analyses ou des recherches, les utilisateurs peuvent être amenés à en regrouper plusieurs. Toutefois, les bases étant volumineuses (en moyenne 5 Go), cela peut générer des temps de calculs très longs.

Dans le cadre de notre étude, comme explicité auparavant dans la partie 1.1 seules les bases de données des années 2018 et 2019 seront donc exploitées. Les données de 2020 n'étant disponible que courant juillet 2021, et l'étude déjà commencée, il a été choisi de ne pas les inclure. De plus, les données sont spécifiques puisque l'année 2020 concerne l'année touchée par la pandémie Covid-19 mais aussi par la fusion du Régime général avec le Régime des Salariés Indépendants. Le périmètre couvert par

la base 2020 est donc différent de celui des années 2018 et 2019. Cependant, la base de données 2020 sera bien traitée et étudiée dans le cadre de futurs travaux au sein de l'entreprise. C'est donc un total de 120 Go de volume de données pour les 24 bases qui seront étudiées dans le cadre de ce mémoire. Le schéma 2.3 indique le nombre de lignes et le volume de données pour les années 2018, 2019 et le total pour ces deux années.

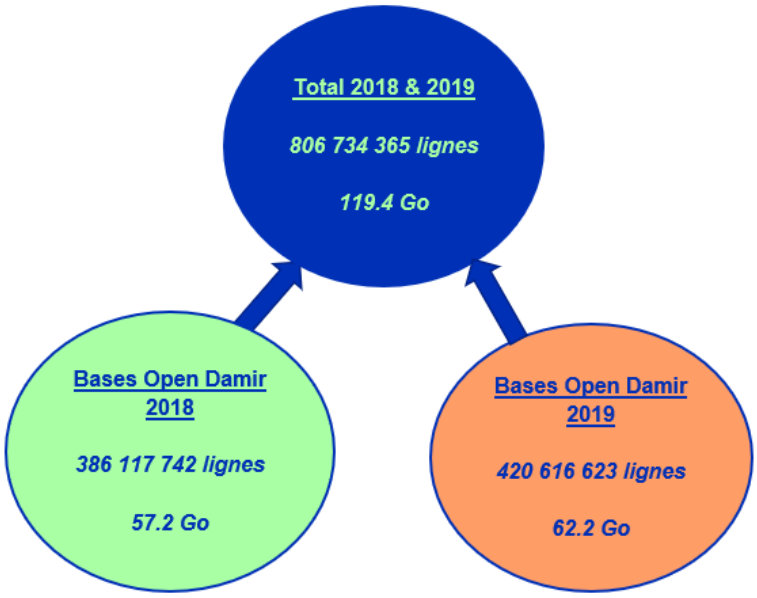


FIGURE 2.3 : Volume des bases de données Open Damir de l'étude.

En général, d'après le Github de la base Open Damir mis à disposition pour le hackathon par CHEVALIER P.A. (23 janvier 2015), « les fichiers compressés fournis par la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS) par année font autour de 7 Go. Décompressé, le total sur 6 ans de 2009 à 2014 fait un peu plus de 200 Go. Une année compte environ 220 millions de lignes. »

Le choix du logiciel est important pour le traitement de ces données. De nombreux utilisateurs ont opté pour des machines virtuelles ou bien le logiciel SAS (Statistical Anlaysis System) pour ces traitements, puis le logiciel *R* (R CORE TEAM, 2022) pour les analyses et l'application de méthodes d'apprentissage statistiques. Dans cette étude, les traitements de chaque base seront réalisés sur $\mathbb{P}ython$, plus particulièrement sur le logiciel Spyder de la solution Anaconda, qui est assez performant, que ce soit pour le développement informatique mais aussi le traitement de données ou bien pour la réalisation de statistiques. Puis, le logiciel *R* sera utilisé pour l'application des méthodes d'apprentissage statistique. Ce sont des logiciels gratuits et donc accessibles facilement pour tout utilisateur.

Objectif de la base Open Damir

La base de données santé Open Damir peut être exploiter par de nombreux utilisateurs (étudiants, doctorants, chercheurs) dans le cadre d'une thèse, d'un mémoire mais aussi pour les particuliers ou les professionnels de santé pour réaliser des recherches plus poussées. Les bases de données étant assez complexes, chacun a à sa disposition un descriptif du jeu de données. Il est téléchargeable au format csv. sur le site DATAGOUV (2021), et contient les significations de chaque variable, ainsi que l'ensemble de leurs modalités avec leur correspondance. Malgré le descriptif, il n'est pas toujours aisé

de comprendre exactement à quoi correspond la variable.

2.1.3 Point d'attention

Aujourd'hui, de nombreux forums voient le jour. Ils permettent aux utilisateurs de poser une question associée à un problème rencontré (scientifique, informatique, et bien d'autres domaines). La plateforme s'enrichit donc avec les anciennes questions et réponses, et permet aux autres utilisateurs ayant rencontré un problème similaire, de trouver la solution rapidement.

Sur la page principale des ressources des bases Open Damir du site DATAGOUV (2021), un forum dédié offre la possibilité à n'importe quel utilisateur, d'interroger des personnes expertes sur ces bases afin d'obtenir des réponses à leurs questions.

Aujourd'hui, les différents types d'informations disponibles sont les suivantes :

- la période de mise en ligne des nouvelles bases ;
- la différence entre les indicateurs pré filtrés et non pré filtrés ;
- les codes des prestations santé pour certains postes (les codes des actes du poste "Hospitalisation" commencent par 2xxx) ;
- les filtres à appliquer pour sélectionner un périmètre en particulier ;
- Etc.

Ce forum a été une aide précieuse pour la réalisation de ce mémoire.

La base de données étant décrite et présentée, la partie suivante détaillera l'ensemble des traitements effectués pour constituer une base d'étude, ainsi que l'ensemble des hypothèses prises, les réflexions associées de leur explications.

2.2 Traitement des données

Compte tenu des volumes évoqués précédemment, il est nécessaire de traiter les données et de ne garder que les éléments essentiels pour les prochaines études. Les traitements appliqués sur chaque base de données mensuelle sont représentés sur le schéma 2.4.

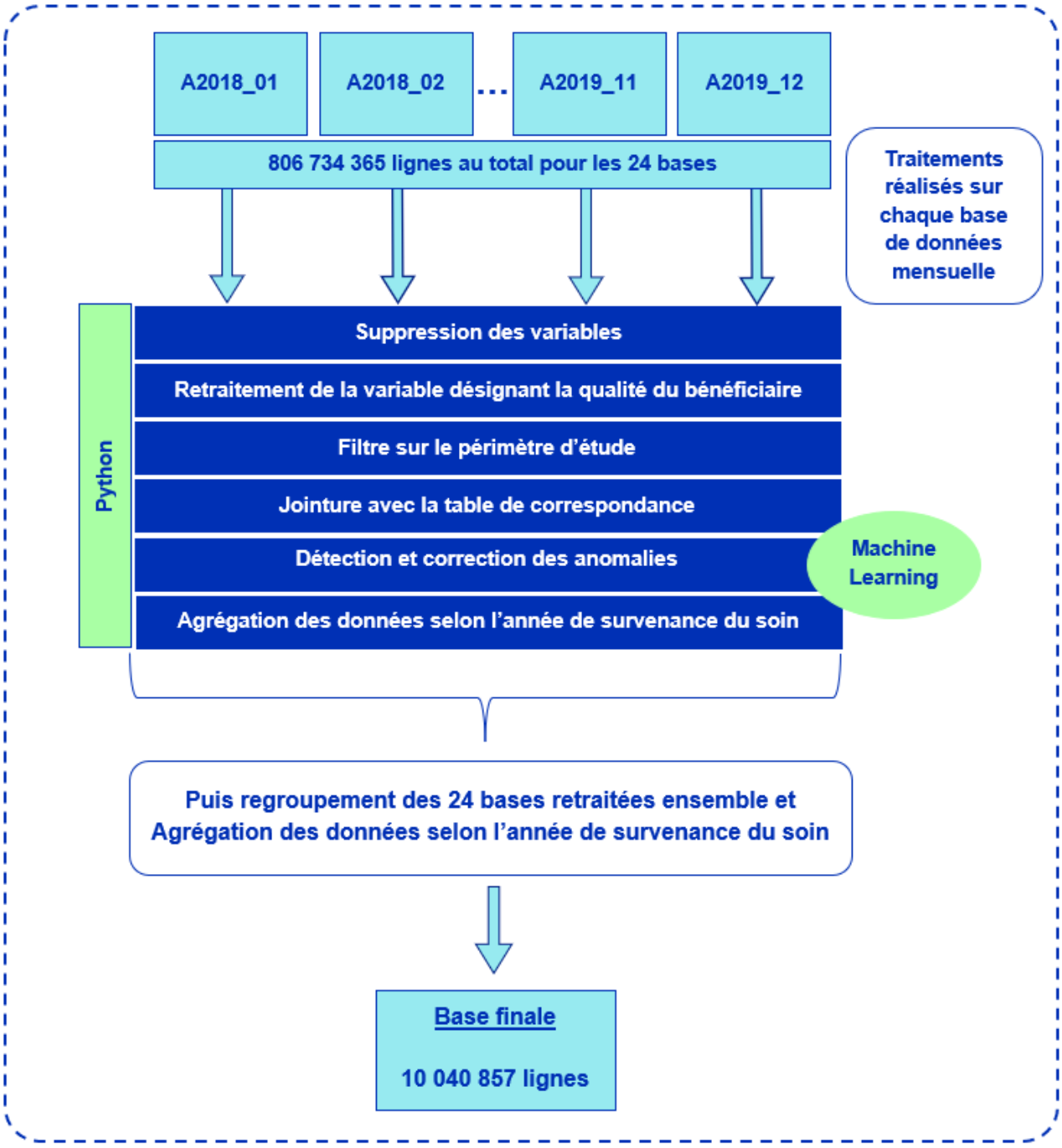


FIGURE 2.4 : Schéma récapitulatif des étapes du traitement des bases de données Open Damir.

En supplément de ce schéma, nous listons ces différentes étapes dans l'ordre d'exécution en apportant quelques justifications :

- suppression d'un grand nombre de variables considérées comme non utiles pour notre étude, et permettant d'alléger la complexité du code. Ce choix de variable a été effectué selon l'index des variables de la base Open Damir. Il s'agit, par exemple, de variables associées aux caractéristiques du professionnel de santé prescripteur ou exécutant ;
- modification de certaines modalités. Par exemple, qualités de bénéficiaires « Conjoint et assimilés » et « Autre ayant droit » ont été regroupés avec la qualité « Assuré » pour former la qualité « Adulte ». En effet, des études (détaillées à la section 4.2) ont montrés que les modalités « Conjoint et assimilé » et « Autre ayant droit » étaient minoritaires au sein de la base. Cela nous parait cohérent puisque la majorité des assurés a son propre numéro de Sécurité Sociale. Une personne est exceptionnellement rattaché à son conjoint, adhérent au régime de l'Assurance Maladie, lorsqu'elle se trouve sans emploi et est concubin ou partenaire pacsé de celui-ci.De plus, il existe une corrélation entre les bénéficiaires « Assuré » et « Conjoint et assimilé ». Le regroupement est donc préférable ;
- filtres pour prendre en compte les hypothèses exposées ci-après à la section 2.2.1 ;
- sélection des lignes associées uniquement aux codes d'actes qui ont été choisis pour l'étude ;
- ajout de la table de correspondance réalisée et exposée ci-après à la section 2.2.2 ;
- détection et traitement des anomalies (c.f. section 2.2.3) ;
- agrégation des données au niveau mensuel puis annuel ;
- ajout des variables cibles (coûts moyen et fréquence) ;
- ajout de ces données finales à la suite de la base mensuelle précédemment traitée.

L'ensemble de ces traitements a été réalisé pour chaque base mensuelle, soit 24 fois, puis nous avons agrégé les tables par année de survenance. Enfin, le nombre de bénéficiaires, issu de données externes provenant de l'Institut National de la Statistique et des Etudes Economiques (INSEE), est rajouté pour chaque ligne (c.f. section 2.2.4).

2.2.1 Hypothèses du retraitement

Les bases de données de l'année 2020 ne sont pas retenues dans le cadre de l'étude pour de plusieurs raisons :

- la base n'était pas disponible au moment du traitement des données ;
- cette base est spécifique puisque le Régime des Salariés Indépendants (RSI) intègre le régime général à partir des bases Open Damir 2020. Cependant, il est dans l'intérêt d'agréger et de comparer des données sur le même périmètre ;
- les données 2020 représentent à nouveau 12 bases, soit environ 70 Go. Le volume de données peut donc être plus important et cela peut rendre le temps de traitement machine plus long.

De plus, il a été considéré que aucun effet d'inflation mensuel n'est présent, mais seulement entre chaque nouvelle année. Cela pourrait faire l'objet d'une étude plus approfondie à l'aide de méthodes de séries temporelles. Cependant, cette problématique ne sera pas l'objet de ce mémoire.

Les étapes les plus importantes du retraitement peuvent à présent être exposées.

2.2.2 Élaboration d'une table de correspondance des actes santé

Il est essentiel de réaliser au préalable un traitement de cette base de données, afin de pouvoir la manipuler par la suite et d'obtenir des résultats cohérents. Cette base nettoyée doit être au format adapté pour les futures études et pour l'obtention de valeurs correctes et cohérentes.

Le premier traitement réalisé concerne le regroupement de codes actes et donc la réalisation de tables de correspondance 2.1 et 2.2 des codes actes de la base Open Damir selon les libellés brochures.

TABLE 2.1 : Table des regroupements selon les libellés brochures santé (1).

Famille d'actes	Sous familles d'actes	Libellé brochure	ID	Remarques
Soins courants	Soins courants	Consultations de généralistes	1	
		Consultations de spécialistes	2	
		Consultations de psychiatres	3	
		Actes techniques médicaux	4	
		Auxiliaires médicaux	5	
		Analyses médicales	6	
Hospitalisation	Hospitalisation	Actes d'anesthésie	7	
		Actes de chirurgie	8	
		Actes d'échographie	9	
		Actes d'imagerie	10	
		Actes d'obstrétrique	11	
		Chambre particulière maternité	12	Non utilisé
		Chambre particulière médicale et chirurgicale	13	
		Frais de séjour médical et chirurgical	14	
		Forfait hospitalier	15	
		Frais de transport	16	
		Frais d'accompagnant	17	Non utilisé
		Autres	18	

TABLE 2.2 : Table des regroupements selon les libellés brochures santé (2).

Famille d'actes	Sous familles d'actes	Libellé brochure	ID	Remarques
Dentaire	Dentaire	Prothèse dentaire - Panier 1	19	
		Prothèse dentaire - Panier 2	20	
		Prothèse dentaire - Panier 3	21	
		Parodontologie	22	
		Implant dentaire	23	
		Soins dentaires	24	
		Orthodontie acceptée / refusée	25	
Optique	Optique	Monture	26	
		Verres	27	
		Lentilles acceptées	28	
		Chirurgie réfractive	29	Non utilisé
		Autres	30	
Pharmacie	Pharmacie	Pharmacie 100%	31	
		Pharmacie 15%	32	
		Pharmacie 30%	33	
		Pharmacie 65%	34	
		Vaccins anti-grippe	35	
		Autres	36	
Appareillage	Appareillage	Accessoires	37	
		Petit appareillage	38	
		Grand appareillage	39	
		Prothèse auditive I	40	
		Prothèse auditive II	41	
Autres	Cure thermale	Cure thermale - Hébergement	42	
		Cure thermale - Transport	43	
		Cure thermale - Soins	44	
	Médecine douce	Médecine douce	45	
	Prévention	Dépistage	46	Non utilisé
		Contraception	47	Non utilisé
		Arrêt du tabac	48	
	Prestations supplémentaires	Soins à l'étranger	49	
		Prime de naissance ou d'adoption	50	Non utilisé
		Maternité	51	Non utilisé
		Autres vaccins	52	Non utilisé

Cette étape est nécessaire puisque de nombreux codes actes existent dans la base Open Damir : 1080 actes sont indexés, mais 1102 actes sont présents au total dans les bases Open Damir 2018 2019. Une mise à jour de l'index sera réalisée lorsque les données 2020 seront disponibles. Plusieurs étapes et hypothèses ont été effectuées afin de réaliser au plus juste cette table de correspondance. Il s'agit de la partie la plus longue du traitement de données, puisque au travers des différents tests, cela a permis d'appréhender de manière plus approfondie la base de données et donc de se l'approprier. L'ensemble des traitements a été réalisé sous $\mathbb{P}ython$ et pour chacune des 24 bases mensuelles. En effet, nous n'avons pas voulu faire l'hypothèse que les résultats trouvés pour une base mensuelle sont généralisables aux autres bases de données. Puis, l'analyse finale des résultats, nécessaire pour valider les hypothèses faites, a été réalisée sous Excel. Les grandes lignes de ce traitement,que nous détaillerons par la suite, sont les suivantes :

- détermination des bases de remboursements moyens pour chaque code acte sur l'ensemble des bases d'études ;
- détermination des codes actes de l'index Open Damir, jamais présents sur les années 2018 et 2019 ;

- détermination des actes concernant des aides / remboursements destinés aux professionnels de santé (Assurance Maladie Obligatoire, juillet 2021) ;
- détermination des actes concernant le secteur de la maladie ;
- détermination des taux de remboursements moyens de chaque code acte sur l'ensemble des bases d'études ;
- définition des codes actes ayant un taux de remboursement moyens à 100% ;
- détermination du nombre moyen de dépassement et du montant de dépassement moyen par code acte sur l'ensemble des bases d'étude ;
- détermination des correspondances à l'aide d'autres tables de correspondance.

L'objectif est de construire une table de correspondance correcte, avec des analyses et des traitements préalables. Les codes actes, pour lesquels il n'était pas possible de déterminer une correspondance, n'ont pas été considérés. De plus, l'ensemble de ces traitements a été réalisé sur le périmètre d'étude défini en section 2.1.1.

Etape 1 : Zoom sur les bases de remboursement de chaque code acte santé.

La détermination des bases de remboursement permet de regrouper les codes actes selon des bases de remboursement similaires. Pour rappel, les bases de remboursement sont des montants agrégés au sein de la base Open Damir. Il est donc essentiel de calculer la base de remboursement moyenne pour un soin. Ce premier traitement a permis d'avoir une première vision générale sur la base d'étude Open Damir. Les bases de remboursement de certains codes actes se situaient entre 10 000 euros et 150 000 euros, ce qui est très élevé pour des actes de santé. En moyenne, les bases de remboursement en santé n'excèdent pas les 2 000 euros. **L'ensemble des codes actes ayant une base de remboursement supérieure à 5 000 euros n'ont donc pas été considérés**. De plus, certains code actes ne possédaient pas de base de remboursement moyen. Deux traitements supplémentaires ont été effectués pour approfondir les résultats suivants :

- l'absence de codes actes au sein de la base d'étude (c.f. section 2.2.2) ;
- la présence de codes actes avec une base de remboursement élevée (c.f. section 2.2.2).

Etape 2 : Zoom sur les codes actes non présents dans les bases d'études.

L'étape précédente a donc permis de se rendre compte que certains codes actes, répertoriés dans l'index de la base Open Damir, étaient absents sur l'ensemble des bases utilisées pour l'étude. Ils n'ont donc pas été considérés pour la suite des traitements. En effet, aucune donnée quantitative ne pourra être analysée. Un traitement a notamment été réalisé pour obtenir le pourcentage que représente chacun des codes actes sur chacune des 24 bases de données sélectionnées. Cela permettra par la suite de supprimer des codes actes qui impacteraient peu les bases de données.

Etape 3 : Détermination des actes concernant des aides / remboursements destinés aux professionnels de santé.

Certains codes actes ont des bases de remboursement moyennes très élevées. Après analyse, il s'agit de codes actes destinés aux professionnels de santé (aides financières pour le matériel, actes effectués, etc.). Ceci explique donc le niveau nettement plus élevé du montant par rapport à des actes de santé traditionnels destinés aux particuliers. Une base de données extraite du site Ameli (2021), recensant l'ensemble des codes actes des remboursements destinés aux professionnels de santé, a permis d'identifier ces codes actes au sein des bases Open Damir utilisées pour l'étude et de les supprimer (Assurance Maladie Obligatoire, juillet 2021). Cette liste, sous format PDF, a été importée sous Excel. Au total, 266 codes actes concernent des remboursements destinés aux professionnels de santé. Seulement 237 de ces codes actes sont présents dans l'index de la base Open Damir, ce qui représente 22% des codes actes de cet index. Les autres codes actes non présents dans l'index sont en faite présents dans les bases Open Damir mais leur libellé n'est pas encore répertoriée dans l'index. Ils ne seront donc pas considérés pour l'élaboration de la table de correspondance.

Etape 4 : Détermination des actes concernant le secteur de la maladie.

Une seconde analyse a été effectuée par la suite, sur le type d'assurances auxquels sont affectés les codes actes. La variable considérée pour les traitements est « ASU_NAT » (c.f. le tableau A.7). Elle possède 11 modalités (c.f. le tableau 2.3).

Table 2.3 : Les différentes modalités de la variable ASU_NAT.

ASU_NAT	Libellé Nature d'Assurance
10	MALADIE
11	MALADIE COURS NAVIGATION > 6 MOIS
12	MALADIE COURS NAVIGATION < 6 MOIS
22	SOINS AUX INVALIDES DE GUERRE (CNMSS)
30	MATERNITE
40	AT ET MP
50	DECES
70	PRESTATIONS SUPPLEMENTAIRES
80	INVALIDITE
90	PREVENTION MALADIE
99	VALEUR INCONNUE

Deux traitements sont réalisés :

- analyse concentrée uniquement sur les prestations relatives à la maladie ;
- analyse de la quantité de chaque acte pour chaque modalité.

Tout d'abord, nous regroupons les modalités relatives à la maladie : Maladie, maladie cours navigation supérieure et inférieure à six mois, prestations supplémentaires et prévention maladie (en bleu dans le tableau 2.3). Ensuite, nous extrayons tous les codes actes pour vérifier qu'il s'agit bien ou non de prestations relatives à la maladie.

Une deuxième analyse est réalisée pour les autres modalités de la variable « ASU_NAT ». Nous déterminons pour chaque code acte, leur répartition sur l'ensemble de ces modalités. Ce traitement concerne notamment l'ensemble des 24 bases de données Open Damir sélectionnées. Les résultats sont sans appel. Les codes actes ne sont pas attribués à un unique type d'assurance. Après l'analyse de ces résultats, il a donc été choisi d'éliminer l'hypothèse de départ qu'était de prendre en compte uniquement les codes actes relatifs à la maladie. Un exemple des résultats obtenus pour la base de données « A2019_08 » lors de ce traitement est disponible en annexe A.1.

Etape 5 : Détermination des taux de remboursement moyens de chaque code acte sur l'ensemble des bases d'études.

Enfin, en plus de l'analyse des bases de remboursements, des tests ont notamment été réalisés pour les taux de remboursement présents au sein des bases sélectionnées pour l'étude. L'objectif était d'analyser les taux de remboursement moyens pour chaque code acte. Une fois ces résultats obtenus, le taux de remboursement moyen de certains codes actes s'élevait à 100%. Il a donc été choisi de ne pas les prendre en compte. En effet, l'objectif de ce mémoire est d'intégrer la base Open Damir en tarification santé pour des contrats de complémentaires santé. Or, la base Open Damir recense l'ensemble des remboursements de la Sécurité Sociale. Si la prise en charge s'élève à 100% pour la Sécurité Sociale, la prise en charge du côté des organismes complémentaires santé est donc nulle.

Pour rappel, l'ensemble de ces traitements a été effectué sur le périmètre d'étude choisi, c'est-à-dire pour les remboursements destinés aux bénéficiaires du régime général uniquement. Si les traitements étaient réalisés sur la base de données globale, le nombre de codes actes avec un taux de remboursement moyen proche des 100% aurait augmenter. En effet, pour le Régime Alsace-Moselle ou les bénéficiaires du dispositif de la CSS, de nombreux soins sont pris en charge en totalité par la Sécurité Sociale. Cela aurait donc biaisé les résultats de ces analyses.

Cependant, une problématique se pose. Pour certains codes actes, le taux de remboursement est proche des 100% sans pour autant l'atteindre. Ces codes actes doivent-ils être supprimés ? Comment définir le seuil de suppression du code acte selon le taux de remboursement moyen ? Deux autres traitements ont donc été effectués pour répondre à cette problématique :

- analyse de la variance du taux de remboursement pour les codes actes pour lesquels ce taux est compris entre 90% et 100%. Cet intervalle a été validé à titre d'expertise (c.f. section 2.2.2) ;
- analyse des dépassements d'honoraires pour les codes actes dont le taux de remboursement moyen est défini à 100% après les analyses précédentes (c.f. section 2.2.2).

Etape 6 : Définition des codes actes ayant un taux de remboursement moyen à 100%.

Ce traitement est effectué sur les codes actes dont le taux de remboursement moyen sur l'ensemble des 24 bases est compris entre 90% et 100%. Pour savoir si le taux de remboursement moyen de ces codes actes doit être considéré à 100%, l'analyse de la variance du taux de remboursement moyen sur chacune des bases est importante. En effet, si pour un code acte donné, la variance du taux de remboursement moyen est faible sur une base donnée et stable sur l'ensemble des 24 bases, alors ce taux de remboursement moyen peut être considéré à 100%. En revanche, si cette variance est importante, le taux de remboursement moyen peut donc varier entre 60% et 100%, par exemple. Ces codes actes ne seront donc pas éliminés. Le choix de considérer ou non un code acte donné dépend donc du niveau

du taux de remboursement moyen et de sa variance. Il s'agit d'une analyse poussée, demandant un certain recul. Le critère de considération du code acte n'est donc pas fixe.

Etape 7 : Analyse des dépassements d'honoraires par code acte sur l'ensemble des bases de l'étude.

De plus, une dernière analyse s'impose pour les codes actes sélectionnés à l'étape précédente, ayant un taux de remboursement moyen considéré à 100%. En effet, comme expliqué par le schéma du principe du remboursement d'un acte santé (c.f. schéma 1.15), des dépassements d'honoraires peuvent être appliqués pour un acte remboursé à 100% par la Sécurité Sociale. Les organismes complémentaires santé peuvent donc intervenir pour prendre en charge une partie ou la totalité du dépassement. Il serait donc incohérent de supprimer les codes actes pour lesquels un dépassement d'honoraire est souvent appliqué. Pour cela, le nombre moyen de dépassements par code acte et pour chacune des 24 bases a été déterminé, ainsi que le montant moyen du dépassement. Si pour un code acte donné, le nombre moyen de dépassements et le montant moyen de ce dépassement est élevé, alors le code acte est gardé. De plus, si le nombre moyen de dépassement d'honoraires pour un code acte donné est élevé mais que le montant moyen de dépassement est très faible, alors le code acte peut être gardé. . Il s'agit d'une analyse poussée, demandant un certain recul. Le critère de considération du code acte n'est donc pas fixe.

Etape 8 : Détermination des correspondances de chaque code acte.

Les codes actes gardés pour cette étude sont à présent déterminés. Les libellés du tableau de correspondance de référence 2.1 peuvent être associés à chaque code acte. Pour cela, plusieurs méthodes sont utilisées :

- utilisation des libellés de chaque code acte issu de l'index Open Damir ;
- utilisation des libellés plus détaillés de chaque code acte issu de plusieurs tables de données ;
- utilisation de tables de correspondance déjà établies pour des clients ;
- recherche sur le site Ameli (2021) de la définition de certains code actes/ libellés.

De plus, pour les différents paniers en dentaire et audiologie une analogie a été faite entre les libellés des codes actes de la base Open Damir et les types de paniers, présentée dans le tableau 2.4.

Table 2.4 : Distinction des différents paniers santé en dentaire et audiologie.

Libellé	Types de paniers (respectivement en dentaire et audiologie)
RAC 0 (Reste à charge 0)	Panier 1 / Panier I
RAC MODERE (Reste à charge maîtrisé)	Panier 2 / Panier II
TARIF LIBRE	Panier 3

Remarque : En optique, la distinction des différents types de paniers pour les code acte de la base Open Damir ne peut pas être effectuée par manque d'informations. L'ensemble des codes actes en optique concerne des équipements optiques destinés aux bénéficiaires du dispositif de la CSS.

Table de correspondance finale

Initialement, l'index Open Damir recensait 1080 codes acte. Après l'ensemble des analyses précédentes, la table de correspondance Open Damir construite ne contient finalement que 319 codes acte. Il est évident que la perte d'information sur les bases de données est minimisée lorsque de nombreux codes actes sont gardés. Cette table de correspondance sera mise à jour les prochaines années, à l'aide de connaissances supplémentaires au fur et à mesure des travaux effectués. Par exemple, les codes actes concernant le poste « Optique » seront analysés afin d'obtenir la distinction des paniers.

Remarque : certains libellés brochure de la table de correspondance de référence ne sont pas affectés à des codes actes de la base d'étude Open Damir. Ceci est cohérent puisque ce sont des soins de santé qui ne sont pas remboursés par la Sécurité Sociale. Aucun remboursement n'apparaît donc dans la base Open Damir. Ces libellés brochures sont associés à la mention « Non utilisé » au sein de la table de correspondance de référence 2.1.

2.2.3 Nettoyage des données et processus automatique de détection d'anomalies

Les principaux traitements réalisés sur les données, comme l'application de filtres pour ne prendre en compte qu'un certain périmètre ou bien la fusion des données avec la table de correspondance des actes, sont à présent effectués. Par la suite, il est indispensable de vérifier la cohérence et l'exactitude des données. En effet, au sein de notre base, certaines variables numériques s'obtiennent par le produit de deux autres. Or, lors des différents tests effectués pour la réalisation de la table de correspondance, de nombreuses incohérences ont été détectées. **Les incohérences relevées sont les suivantes :**

- le montant remboursé par la Sécurité Sociale ne correspond pas au produit de la base de remboursement et du taux de remboursement en pourcentage (le taux est compris entre 0 et 100 au sein de la base) ;
- certaines quantités d'actes nulles sont associées à des montants (montant du dépassement, du montant remboursé, de la dépense) non nuls. Seules les régularisations effectuées par la Sécurité Sociale ne sont pas considérées comme anormales.

L'identification de ces anomalies lors du traitement de données

Ces anomalies ont donc été corrigées manuellement. Pour cela, le montant remboursé a été considéré comme une information fiable puisqu'il correspond au montant payé par la Sécurité Sociale, qui est donc supposé correspondre au montant présent dans les comptes. Seuls la base de remboursement et le taux de remboursement ont donc été modifiés. Pour cela, une nouvelle variable « ANOMALIE » a été créée. La modalité prise pour chacune des lignes de la base est **-1** si l'information, en fonction des tests effectués, est considérée comme une anomalie, +1 dans le cas contraire. Les critères et tests choisis pour renseigner cette variable sont les suivants :

- la ligne est considérée comme une anomalie si le montant remboursé est différent du produit de la base et du taux de remboursement (réexprimé en pourcentage) ;
  - un second test sur ce produit a été effectué. En effet, il a été remarqué que l'égalité était fausse avec une différence au centième près pour les décimales. Par conséquent, il a été défini que ces lignes ne seraient pas des anomalies.

- si les montants de remboursement, de dépense, de dépassement et de base de remboursement sont non nuls dans le cas où la quantité d'actes est nulle, cette ligne est considérée comme une anomalie et l'ensemble de ces montants seront modifiés ;

Deux nouvelles variables ont donc été créés : BASE_REMB_CORR et TAUX_REMB_CORR. Ils représentent respectivement la correction des valeurs prises initialement par les variables de la base de remboursement « PRS_REM_BSE » et du taux de remboursement « PRS_REM_TAU ». Les modifications sont réalisées en considérant le cas où il y a ou non une anomalie, mais aussi dans le cas où le taux de remboursement initial est non nul ou bien nul :

- **dans le cas où la ligne possède une anomalie et que le taux de remboursement est non nul :**
  - la base de remboursement est redéfinie comme le quotient du montant remboursé et du taux de remboursement initial,
  - le taux de remboursement corrigé est égal au taux de remboursement initial.
- **si le taux de remboursement est nul et la ligne possède toujours une anomalie :**
  - le taux de remboursement corrigé est redéfini comme le quotient du montant remboursé et de la base de remboursement initial,
  - la base de remboursement corrigée est égale à la base de remboursement initiale.
- **dans le cas où la ligne ne possède pas d'anomalie :**
  - le taux de remboursement corrigé est égal au taux de remboursement initial,
  - la base de remboursement corrigée est égale à la base de remboursement initiale.

Problématique

Les anomalies détectées, les tests et traitements effectués sont cependant spécifiques à cette étude. En effet, il est possible que des anomalies de natures différentes soient détectables pour les bases de données des années à venir. L'objectif de cette partie du mémoire est donc de généraliser l'ensemble de ces traitements pour les prochaines années afin d'identifier de nouvelles anomalies sans pour autant connaitre leurs natures au préalable. Pour cela, il est essentiel de mettre en place un processus qui permettrait d'éviter de nouvelles réalisations de tests et d'analyses, et qui les détecterait automatiquement. De plus, le volume des bases de données en Open Data ne vont cesser de s'accroître au fil des années, ce qui impliquera des analyses plus lourdes et probablement des oublis de prise en considération de certaines observations... Pour répondre à ce besoin, il a donc été choisi de réaliser un outil de détection d'anomalies.

Explication de l'objectif et du fonctionnement de la détection d'anomalies

La détection d'anomalies est un processus généralement utilisé pour lutter contre la fraude, mais aussi pour repérer des comportements atypiques dans un ensemble d'observations, ou bien pour la vérification de la qualité de données. De plus en plus utilisée dans le domaine de l'assurance, cette méthode permet de répondre à un certain nombre de problématiques actuarielles à partir de méthodes d'apprentissage statistiques, de Machine Learning (ML) ou de Scoring.

Il existe effectivement plusieurs méthodes d'apprentissages statistiques :

- **méthode d'apprentissage supervisé :** les données sont labellisées. Une variable cible définie préalablement indique si une observation doit être considérée comme une anomalie ou non ;
- **méthode d'apprentissage non supervisé :** aucune variable cible n'est présente pour ce type de méthode, les observations ne sont pas associées à une variable qui les caractérise.

La méthode la plus fréquemment utilisée en détection d'anomalies est l'apprentissage statistique non supervisé. En effet, l'objectif est de détecter les observations s'éloignant des observations normales (observations aberrantes, observations extrêmes, observations erronées), et dont l'utilisateur ne connaît pas la nature au préalable, c'est-à-dire, il n'a pas la connaissance des étiquettes indiquant si l'observation est une anomalie ou non. De plus, pour que ce processus fonctionne correctement il est essentiel de vérifier les hypothèses suivantes sur les anomalies :

- les anomalies sont rares. Dans le cas de l'étude de ce mémoire, au sein d'une base mensuelle Open Damir, les anomalies représentent en moyenne **21.5%** de la base de données. Le tableau A.2 résume le pourcentage d'anomalies pour chacune des 24 bases utilisées pour notre étude ;
- les anomalies sont nouvelles et différentes les unes des autres. Cette hypothèse est confirmée puisque les anomalies sont d'ores et déjà de natures différentes et cela ne cessera de s'accroître au fil des années.

Dans le cadre de cette étude, l'apprentissage statistique non supervisé sera donc utilisé. Cependant, il sera essentiel de garder connaissance de nos informations sur la variable « ANOMALIE ». Cela permettra de vérifier le bon fonctionnement du modèle et de la bonne détection des anomalies sur notre base de données servant à l'étude. La méthode de Machine Learning « **Isolation Forest** » **(ou iForest)** a été choisie pour l'implémentation de ce processus de détection d'anomalies.

Le fonctionnement de la méthode « iForest »

L'iForest (M. Clarke, 2021) est un algorithme sur les forêts d'isolations. Contrairement aux forêts aléatoires, les forêts d'isolations fonctionnent sur le principe de l'apprentissage non supervisé. Cependant, à l'instar des forêts aléatoires et des arbres de décisions, son principe repose sur la construction d'arbres. Pour un échantillon aléatoire d'observations, l'algorithme va détecter et isoler rapidement les observations atypiques du jeu de données (les anomalies) en fonction d'une variable et d'un seuil de découpage choisi aléatoirement. Quant aux données normales, elles pourront être isolées par l'arbre une fois l'isolation des données atypiques effectuées. Une illustration plus détaillée des étapes réalisées par l'algorithme est fournie ci-après (c.f. schéma 2.5, 2.6, 2.7).

L'étape principale qui est itérée jusqu'à l'isolation de chaque observation est la suivante. L'algorithme choisit aléatoirement une variable, notée X, et un seuil, noté s_1, dont la valeur est comprise entre le minimum et le maximum des valeurs prises par la variable X. Deux branches sont alors créées dans l'arbre :

- une branche qui caractérise les observations ayant une valeur inférieure ou égale à s_1 pour la variable X ;
- une branche qui caractérise les observations ayant une valeur strictement supérieure à s_1 pour la variable X.

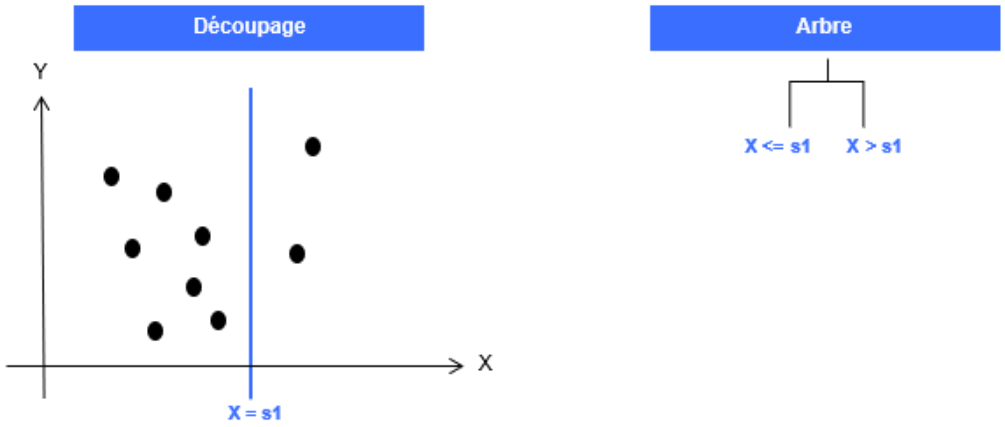


FIGURE 2.5 : Schéma explicatif du fonctionnement de l'iForest (1).

De plus, à chaque fois qu'une observation est isolée du reste des observations, elle n'est plus prise en compte pour la détermination de la valeur du seuil. Finalement, après toutes les itérations effectuées, un arbre est obtenu, composé de plusieurs branches. Cependant, l'iForest est une méthode de forêts d'isolations et est donc constitué de plusieurs arbres. L'algorithme relance les étapes précédentes afin de construire d'autres arbres mais selon des sélections de variables et de seuils différents pour chaque itération. Le graphique 2.6 illustre un arbre obtenu après la réalisation de toutes les itérations.

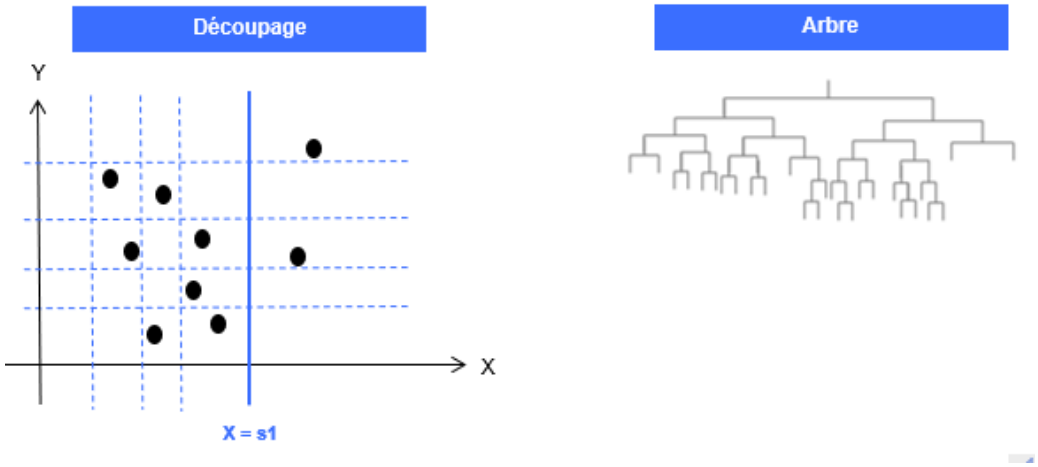


FIGURE 2.6 : Schéma explicatif du fonctionnement de l'iForest (2).

Ainsi, comme expliqué dans le premier paragraphe du fonctionnement de cette méthode de Machine Learning, l'iForest détecte très rapidement, et en premier, les données atypiques.

Cela est identifiable sur l'arbre obtenu sur le graphique 2.7. Effectivement, l'observation identifiée par la couleur bleue a été rapidement détecté. Seules 3 découpes (ou splits) ont été nécessaires pour l'isoler. Pour cette observation, la profondeur de l'arbre est plus faible que pour l'observation identifiée par la couleur verte. La profondeur de l'arbre sert à calculer la valeur finale attribuée à chaque observation, appelée score d'isolation. Ce score est attribué pour chaque observation et est défini comme la moyenne des scores d'isolations pour cette observation sur chaque arbre construit. Le nombre d'arbres utilisés et le nombre de variables ainsi que le seuil de découpage sont donc des paramètres clés pour cet algorithme iForest.

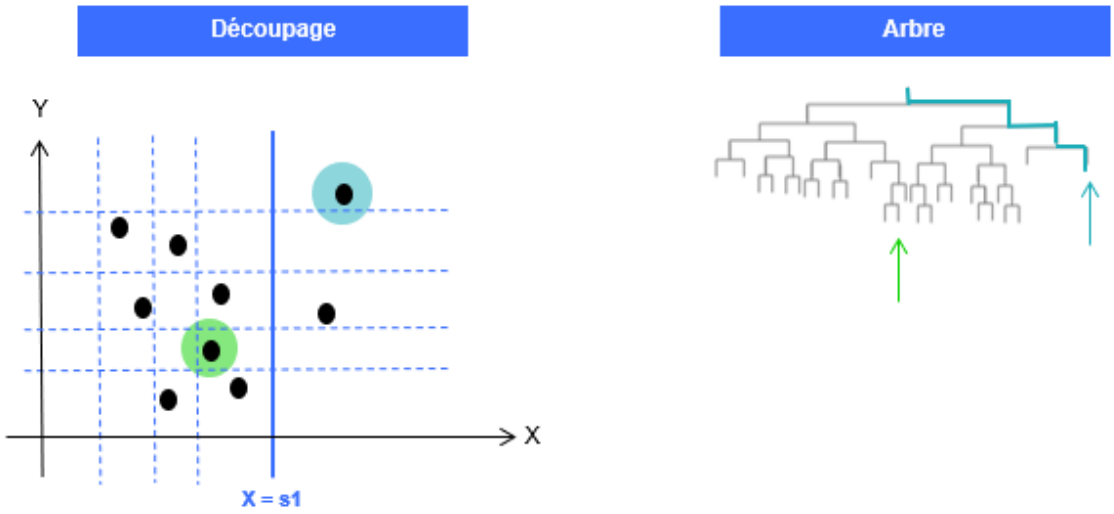


FIGURE 2.7 : Schéma explicatif du fonctionnement de l'iForest (3).

Finalement, lors de la prédiction, l'algorithme va associer une valeur de prédiction selon la valeur du score :

- si le score est positif, la prédiction vaudra +1, et l'observation sera considérée comme non atypique ;
- si le score est strictement négatif, la prédiction vaudra -1, et l'observation sera considérée comme une anomalie.

Le fonctionnement de l'iForest étant à présent expliqué, il est possible de dégager des avantages et des inconvénients sur cette méthode d'apprentissage statistique non supervisé, comme pour toute autres méthodes.

Les avantages et les limites de l'iForest

Avantages : Isolation Forest est une méthode d'apprentissage non supervisé récente et la plus utilisée pour la détection d'anomalies. Il s'agit notamment d'une méthode peu coûteuse en temps d'implémentation et d'exécution. En effet, comme vu précédemment, les forêts d'isolations fonctionnent par l'attribution d'un score selon la réalisation d'un découpage des données à l'aide de la construction d'arbres. Elles n'utilisent donc pas de calculs de distance ou de densité, ce qui diminue considérablement le temps d'exécution. De plus, cet algorithme est peu consommateur en mémoire et peut être exécuté parallèlement sur plusieurs processeurs. Grâce à ces avantages, cet algorithme est performant pour des bases de données volumineuses, ce qui est le cas de cette étude avec l'exploitation des bases de données Open Damir.

Inconvénients et limites : Cependant, l'algorithme iForest peut identifier à tort des observations comme des anomalies ou considérer des observations initialement anormales comme normales. Il s'agit du phénomène appelé « swamping » qui apparaît dès que les observations atypiques sont trop proches

des observations normales. Ce phénomène est principalement identifiable lorsque la détection d'anomalies est utilisée pour s'assurer de la qualité des données ou pour détecter des données erronées. D'autre part, il existe aussi le phénomène appelé « masking », qui apparaît lorsqu'un grand nombre de données atypiques est présent dans le jeu de données entraînant des isolements assez large. Ces deux phénomènes sont principalement dû au volume important d'observations du jeu de données. Afin d'atténuer ces deux phénomènes et comme décrit dans l'explication de l'iForest, l'algorithme doit construire les arbres sur des sous-échantillons de petite taille.

L'iForest est donc performant au niveau du temps d'implémentation pour des bases de données volumineuses telles que les bases de données Open Damir (contenant plus d'une trentaine de millions de lignes). Cependant, la prédiction des anomalies peut être impactée par ce volume important de données. Le paramétrage de l'algorithme doit être effectué en fonction de ces deux limites.

Application et résultats

Comme pour l'intégralité des traitements effectués au sein de cette étude, cet algorithme sera utilisé sur $\mathbb{P}ython$ via l'implémentation disponible dans Scikit-learn (explication de Scikit learn). Ce processus de détection anomalies a été mis en place à l'aide de la base de données mensuelle « A2019_11.csv » qui recense l'ensemble des remboursements effectués par la Sécurité Sociale en novembre 2019. Elle est constituée de 33 794 832 lignes et 55 variables. Le choix d'effectuer ce processus uniquement sur une base mensuelle se justifie par l'augmentation du temps d'exécution du processus par la réalisation d'un hyper paramétrage. En effet, ce temps représente environ 24 heures pour le traitement sur une base de données d'environ une trentaine de millions de lignes. Il faudrait alors environ un mois pour obtenir les résultats sur chacune des 24 bases de données mensuelles Open Damir. Il a donc été considéré que ce traitement est généralisable pour les autres bases mensuelles.

Avant la réalisation du traitement de la détection d'anomalies de ce processus, le nettoyage des données et les filtres choisis pour le périmètre d'étude sont appliqués afin d'établir ce processus uniquement sur le périmètre choisi préalablement (c.f. section 2.1.1). Elle ne comporte donc plus que 12 851 857 lignes et 30 variables (y compris la variable « ANOMALIE »).

Les étapes de la réalisation de ce processus de détection d'anomalies sont les suivantes :

- réduction de la dimension de la base de données via la méthode de Machine Learning LGBM ;
- découpage de la base de données en 3 bases : bases d'apprentissage, de validation et de test ;
- réalisation d'un hyper paramétrage ;
- application de l'iForest avec les paramètres les plus performants déterminés lors de l'hyper paramétrage ;
- obtention des résultats et de la prédiction des anomalies.

Réduction de la dimension de la base de données via la méthode de Machine Learning LGBM. Pour l'application de méthode de Machine Learning, il est important de réduire la dimension des bases de données utilisées. En effet, le fléau de la dimension impacte la prédiction des méthodes de Machine Learning. La réduction de la dimension va donc agir sur deux axes :

- la diminution du temps d'exécution de la méthode de Machine Learning ;

- l'amélioration de la performance de prédiction de la méthode de Machine Learning.

Pour effectuer une réduction de dimension, plusieurs méthodes existent, comme l'Analyse en Composante Principale (ACP). Dans le cadre de l'étude, la méthode de Machine Learning LGBM est choisie. Le Light Gradient Boosting Machine est une méthode de Machine Learning similaire à la méthode XGBoost, très adaptée pour les bases de données volumineuses. Les temps de traitements sont plus rapides et les prédictions sont meilleures avec ce type de méthode. Une explication plus détaillée de la méthode LGBM et XGBoost sera fournie la section 3.2.2. Pour réaliser cet algorithme, il faut découper notre base de données mensuelle en trois bases distinctes :

- deux bases d'apprentissage représentant 60% de la base de données, nommée x_train, contenant l'ensemble des variables sauf la variable « ANOMALIE », et nommée y_train contenant uniquement la variable « ANOMALIE » ;
- deux bases de validations représentant 20% de la base de données, nommée x_validation, contenant l'ensemble des variables sauf la variable « ANOMALIE », et nommée y_validation contenant uniquement la variable « ANOMALIE » ;
- deux bases de validations représentant 20% de la base de données, nommée x_test, contenant l'ensemble des variables sauf la variable « ANOMALIE », et nommée y_test contenant uniquement la variable « ANOMALIE ».

Après le lancement de l'algorithme sur les bases d'apprentissage, le graphique 2.8 fournit l'information des dix variables les plus explicatives de la variable « ANOMALIE » associées à leur score respectif.

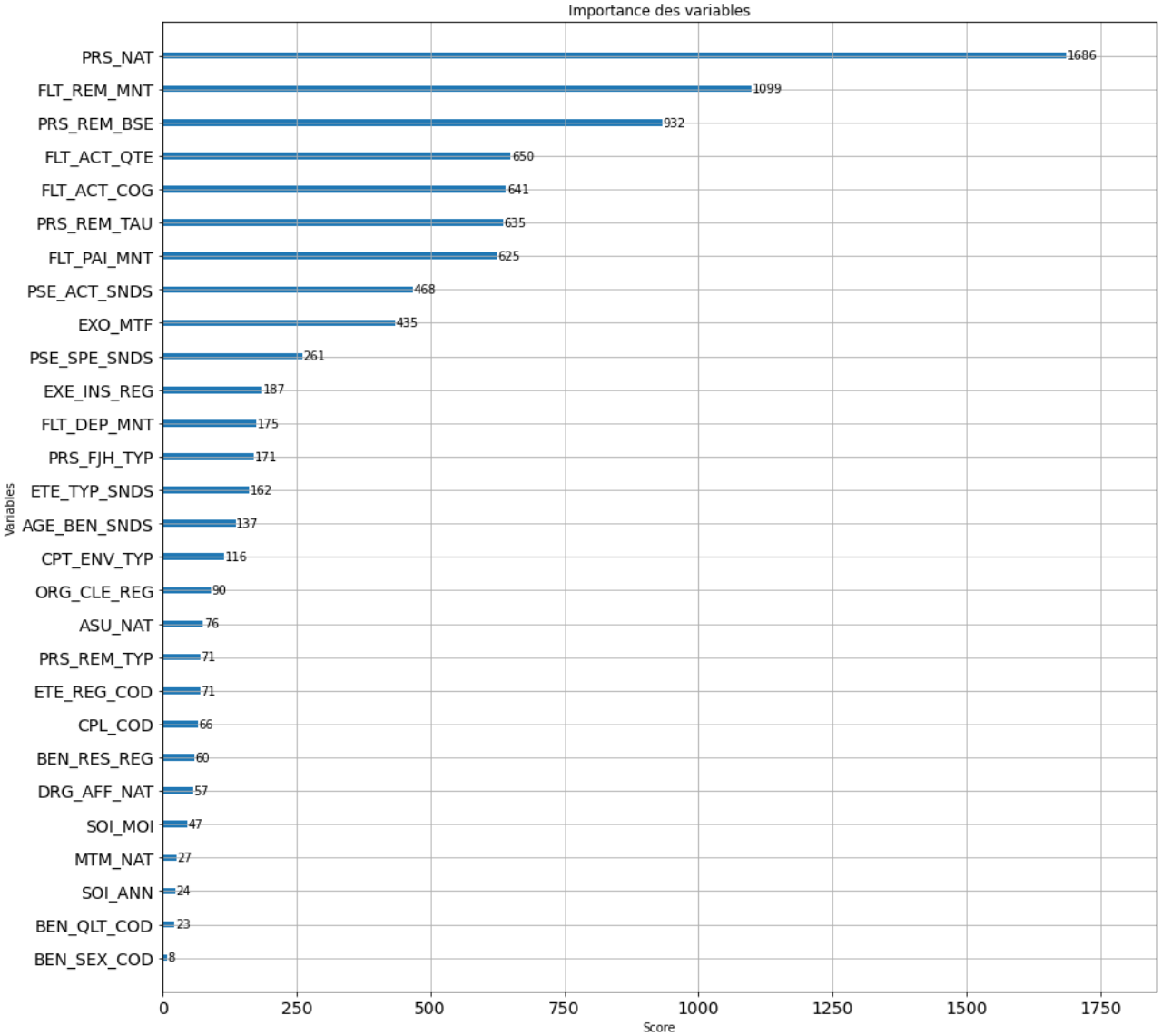


FIGURE 2.8 : Importances des variables explicatives de l'anomalie pour la base A2019_11.csv.

Les variables gardées pour la réalisation du processus de détection d'anomalies sont :

- les variables les plus importantes en terme de score ;
- les indicateurs de montants et de volume. En effet, l'objectif du processus étant de détecter des données erronées sur ces indicateurs, ceux-ci seront gardés.

De plus, seulement 15% des observations seront utilisés pour la réalisation du processus. Cela permettra d'avoir un temps d'exécution raisonnable pour l'hyper paramétrage et d'éviter le phénomène « masking ». Finalement, la base de données mensuelle qui servira à la réalisation de ce processus comportera 1 927 779 lignes et 12 variables résumées dans le tableau 2.5.

TABLE 2.5 : Variables gardées pour le processus de détection d'anomalies.

Variables gardées	Score obtenu via le LGBM
PRS_NAT	1 686
PSE_ACT_SNDS	468
EXO_MTF	435
PSE_SPE_SNDS	261
PRS_REM_BSE	932
FLT_ACT_COG	641
FLT_ACT_QTE	650
FLT_DEP_MNT	175
FLT_PAI_MNT	625
FLT_REM_MNT	1 099
PRS_REM_TAU	635
ANOMALIE	-

Découpage de la base de données en trois bases : bases d'apprentissage, de validation et de test. La base de données traitée est découpée en trois bases :

- base d'apprentissage (nommé « train ») : elle représente 60% des observations. Le modèle est entraîné sur cette base ;
- base de validation : elle représente 20% des observations. La prédiction est effectuée sur cette base, à l'aide du modèle entraîné précédemment ;
- base de test : elle représente 20% des observations. Cette base est utilisée lors de l'exécution du modèle avec les paramètres optimaux, ce qui permet par la suite de conclure sur la robustesse du modèle.

Ce découpage est important afin d'éviter le sur-apprentissage. L'objectif est que le modèle ne soit entraîné que sur un certain nombre de données, puis validé sur une autre partie de la base avec la prédiction. Après l'obtention de ces trois bases, il est important de vérifier que le taux d'anomalies, information indiquée par la variable « ANOMALIE », est stable sur les trois bases. Le tableau 2.6 présente les résultats obtenus.

TABLE 2.6 : Vérification de la stabilité du taux d'anomalies dans chacune des trois bases.

Base	Taux d'anomalies
Train	22.39%
Validation	22.35%
Test	22.46%

Le taux d'anomalies est donc stable sur ces trois bases. A présent, les données sont traitées, le processus de détection d'anomalies peut être lancé avec la réalisation de :

- l'hyper paramétrage sur les bases d'apprentissages et de validation ;

- la création du modèle (servant de processus automatique de détection d'anomalies pour les bases de données des années à venir) avec la base de test.

Réalisation de l'hyper paramétrage. Un hyper paramétrage a été effectué pour choisir les paramètres optimaux afin de réaliser un modèle robuste. Les paramètres optimaux correspondent aux valeurs minimisées du Root Mean Squared Error (RMSE), du Mean Squared Error (MSE) et du critère donnant la Somme des Carrés des Résidus (SCR). En effet, ces indicateurs permettent d'évaluer les écarts entre les valeurs observées et les valeurs prédites. L'objectif est donc de minimiser cet écart. L'explication de ces indicateurs est disponible en annexe B.2.

Plusieurs paramètres sont à renseigner pour l'algorithme de l'Isolation Forest. Un intervalle de valeurs à tester pour chaque paramètre a été choisi pour l'hyper paramétrage. Les tableaux 2.7 et 2.8 décrivent les différents paramètres ainsi que les intervalles choisis.

TABLE 2.7 : Descriptif et valeurs testées pour chaque paramètre de l'algorithme iForest (2).

Nom du paramètre	Descriptif	Valeurs testées
contamination	La proportion de valeurs aberrantes dans l'ensemble de données.	[0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
max_features	Le nombre de caractéristiques/ variable à choisir dans la base X pour former chaque estimateur de base.	[5, 6, 7, 8, 9, 10, 11]
bootstrap	Si ce paramètre est défini à « Vrai », les arbres sont ajustés sur des sous-ensembles aléatoires des données d'apprentissage échantillonnées avec échange. Il s'agit du processus de validation croisée. Si ce paramètre est défini à « Faux », un échantillonnage sans remise est effectué.	Vrai
random_state	Contrôle le caractère aléatoire de la sélection des variables et du seuil pour chaque étape et chaque arbre. Cela permet de reproduire des résultats sur plusieurs appels de fonction.	42

TABLE 2.8 : Descriptif et valeurs testées pour chaque paramètre de l'algorithme iForest (1).

Nom du paramètre	Descriptif	Valeurs testées
n_estimators	Le nombre d'estimateurs de base.	[50, 150, 250, 350, 450]
max_samples	Le nombre d'échantillons à tirer dans les bases X (contenant les variables caractéristiques) pour former chaque estimateur de base.	[50, 150, 250, 350, 450]

Après l'exécution de l'hyper paramétrage, d'une durée d'environ 36 heures, les résultats pour chaque combinaison de ces paramètres avec la valeur du SCR, du MSE, et du RMSE sont obtenus. Un tri par ordre croissant est ensuite effectué. La première ligne du tableau obtenu correspond donc à la valeur des paramètres optimaux pour le processus de détection d'anomalies. Seules les 5 premières lignes des résultats seront données dans le tableau 2.9.

TABLE 2.9 : Résultats de l'hyper paramétrage avec l'algorithme iForest (1).

N° de l'itération	n_estimators	max_samples	contamination	max_features	Score	MSE	RMSE	SCR
0	50	50	0.01	5	0.2750	0.8999	0.9486	346692
5	50	50	0.01	10	0.2555	0.9013	0.9494	347248
2	50	50	0.01	7	0.2653	0.9014	0.9494	347276
1	50	50	0.01	6	0.2718	0.9025	0.9500	347684
986	450	50	0.01	11	0.2630	0.9025	0.9500	347704

La valeur optimale du paramètre contamination est égale à 0.01 ce qui peut être surprenant. En effet, le tableau A.2 indique que le taux d'anomalies au sein de chaque base Open Damir est en moyenne de 21.5%. La valeur optimale pour le paramètre contamination devrait donc se situer à 0.2 ou 0.25. Les autres paramètres optimaux pour ces deux valeurs de contaminations sont donnés dans le tableau 2.10.

TABLE 2.10 : Résultats de l'hyper paramétrage avec l'algorithme iForest (2).

N° de l'itération	n_estimators	max_samples	contamination	max_features	Score	MSE	RMSE	SCR
472	150	450	0.2	8	0.0226	1.2922	1.1367	497836
479	150	450	0.25	8	0.0134	1.4081	1.1866	542496

La prochaine étape est donc de réaliser le processus de détections d'anomalies en exécutant le modèle avec les paramètres optimaux (en gras dans le tableau précédent) sur la base test et d'analyser les performances de ce processus. Un second test sera réalisé en prenant les paramètres du tableau 2.10.

Réalisation du modèle et résultats obtenus. Le modèle le plus robuste étant obtenu, il est possible d'analyser la précision de la prédiction sur la base de test. Le modèle attribue un score à

chacune des observations. Le seuil par défaut est défini à 0. Par conséquent, si le score est strictement négatif, l'observation est considérée comme une anomalie. Dans le cas contraire, l'observation n'est pas considérée comme une anomalie. Le tableau 2.11 présente les résultats obtenus pour les trois valeurs du paramètre contamination avec ce seuil.

TABLE 2.11 : Résultats du processus de détection d'anomalies avec un seuil égal à 0.

Modèle optimal	Contamination = 0.01	Contamination = 0.20	Contamination = 0.25
Valeur du seuil	0	0	0
Taux de bonnes prédictions	**77.54%**	**67.86%**	**65.01%**
Taux de mauvaises prédictions	**22.46%**	**32.14%**	**34.99%**
Faux positifs *(observation considérée comme non atypique alors qu'elle était définie comme une anomalie au préalable par les différents tests)*	21.89%	17.22%	16.16%
Faux négatifs *(observation considérée comme une anomalie alors qu'elle était définie comme non atypique au préalable par les différents tests)*	0.57%	14.92%	18.83%

Le modèle obtenu avec les meilleurs performances est celui lancé avec les paramètres optimaux dont les résultats sont donnés dans la première colonne du tableau récapitulatif. Pourtant, une différence au niveau des taux de faux positifs et faux négatifs est notable pour les trois tests. En effet, une nette amélioration est remarquable lorsque le taux de contamination se rapproche du taux d'anomalies moyen de 0,215. Les anomalies sont mieux captées parmi l'ensemble des observations. Le taux de faux négatifs montre notamment que de nombreuses observations vont être considérées comme atypiques alors qu'elles ne l'étaient pas initialement. Ce scénario est préférable à celui du meilleur modèle obtenu. Un modèle qui capte toutes les anomalies, voire plus (taux de faux positifs moyen, taux de faux positifs moyen) est plus performant qu'un modèle ne captant pas toutes les anomalies (taux de faux positifs élevé, et taux de faux négatifs faibles) puisque l'objectif étant de capter l'intégralité des anomalies de la base de données. Cependant, les performances des deux autres modèles ne sont pas suffisantes. Le premier modèle choisi sera retenu pour les tests suivants.

La robustesse du modèle est donc remise en question face aux résultats obtenus. Cependant, il est possible de modifier la valeur du seuil, qui définit si l'observation est atypique ou non, en fonction de la répartition des scores des observations sur un histogramme. Pour cela, la densité de la variable « score », créé par le modèle, est tracée et visible sur le graphique 2.9.

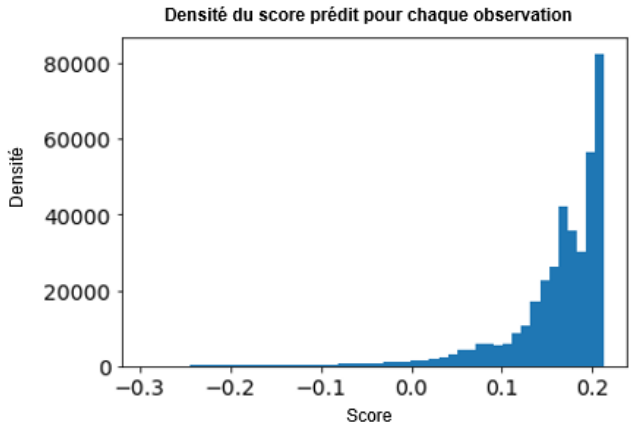


FIGURE 2.9 : Densité du score prédit par le modèle iForest pour chaque observation.

Les anomalies étant supposées rares, elles se situent principalement sur la queue fine de la distribution. Le quantile à 95% de la densité est d'abord choisi pour la valeur du seuil et un test est à nouveau effectué pour conclure de la robustesse du modèle. Avec cette valeur pour le seuil, les anomalies représentent donc 5% d'anomalies dans la base de données. Enfin, un dernier test a été effectué en prenant la valeur du seuil égale au quantile à 78.5%. En effet, les anomalies représenteront 21.5% de la base de données, soit la moyenne déterminée précédemment (c.f. le tableau A.2). Le tableau 2.12 présente les résultats obtenus après modification du seuil, pour les trois valeurs du paramètre contamination choisies.

TABLE 2.12 : Résultats du processus de détection d'anomalies après modification du seuil.

Contamination = 0.01	Seuil à Q95%	Seuil à Q78.5%
Valeur du seuil	-0.1124	-0.0145
Taux de bonnes prédictions	**75.55%**	**65.85%**
Taux de mauvaises prédictions	**24.45%**	**34.15%**
Faux positifs *(observation considérée comme non atypique alors qu'elle était définie comme une anomalie au préalable par les différents tests)*	20.88%	17.48%
Faux négatifs *(observation considérée comme une anomalie alors qu'elle était définie comme non atypique au préalable par les différents tests)*	3.57%	16.67%

Malgré cela, les résultats obtenus ne correspondent pas aux attentes initiales. Malgré les deux méthodes (par défaut et redéfinition du seuil), nous n'obtenons pas une bonne prédiction des anomalies. La méthode n'est donc pas généralisable sur ce grand volume de données. D'autres méthodes pourront être testées par la suite pour réaliser ce processus. Ce sujet n'étant pas principal au mémoire, nous n'irons pas plus loin dans les recherches. Pour les futures années, ces méthodes pourront être testées, ou d'autres analyses manuelles seront donc effectuées.

2.2.4 Agrégation de données externes pour passer outre l'anonymat

Comme vu initialement dans la section 1.2.4, les données de la base Open Damir sont anonymisées. Cela signifie que le nombre de bénéficiaires pour chacune des lignes agrégées est inconnu. Cependant, il est essentiel de posséder cette information afin de déterminer la fréquence, définie comme étant le quotient de la quantité d'actes totale et du nombre d'effectifs couvert, pour chacune des lignes. Le nombre de bénéficiaires n'étant pas disponible dans les bases Open Damir, nous devons les récupérer à partir d'autres bases exogènes.

Pour cela, nous utliserons le site de l'Institut National de la Statistique et des Etudes Economiques (INSEE). Il contient de nombreuses analyses et statistiques sur la population française et son économie. Nous pouvons notamment disposer d'informations sur le nombre de personnes couvertes par la Sécurité Sociale.

En effet, il s'agit d'informations sur les « Bénéficiaires du régime général de l'assurance maladie » publiée chaque année, contenues dans le dossier des données sur les quartiers de la politique de la ville. Les informations sont indiquées selon le sexe, les tranches d'âges, pour les individus bénéficiant de l'Assurance Maladie du régime général, mais aussi ceux bénéficiant du dispositif de la CSS. Elle sont recensées selon plusieurs périmètres (les informations sur les périmètres sont des informations provenant de l'INSEE) :

- les Quartiers Prioritaires de la politique de la Ville (QPV) ;
- les Iris des communes de plus de 10 000 habitants ;
- les communes de plus de plus de 10 000 habitants et/ou contenant au moins un Quartier Prioritaire de la politique de la Ville ;
- les Etablissements Public de Coopération Intercommunale (EPCI) contenant au moins une commune de plus de 10 000 habitants ou un Quartier Prioritaire de la politique de la Ville.

Les informations ne sont donc pas complètes puisque les données des villes et villages de moins de 10 000 habitants n'apparaissent pas au sein de cette base. Ce manque d'information ne permet pas d'utiliser cette base en tant que données exogènes de la base Open Damir. De plus, les données 2019 ne sont pas disponibles en raison de la présence « d'anomalies identifiées par la CNAM ». Cela est dû à la mise en place du nouveau dispositif de la PUMA qui remplace celui de la CMU (c.f. section 1.1.3). « Les statuts de certains bénéficiaires » ont été modifiés et ne sont plus présents au sein de cette base de données de l'INSEE. Il a donc été décidé de ne pas publier les informations. Du fait de l'absence d'informations mais aussi du manque de précision et de fiabilité, cette base n'a donc pas été retenue pour la suite de notre étude.

Après réflexion, il a été décidé de retenir la base de données INSEE donnant des informations sur la population nationale en fonction du sexe, de la région et des tranches d'âges (INSEE, 2021). La base de données avec les grandes classes d'âge n'a pas été choisie puisque elles sont plus large que celles présentes dans la base Open Damir. Le problème de périmètre est donc épargné puisque l'ensemble de la population française est enregistrée au sein de cette base. Cependant, le régime général ne couvre pas l'intégralité des français. Quel pourcentage doit être retenu et appliqué à ces valeurs pour obtenir des informations cohérentes ? Dans les rapports chiffres clés (Direction de la Sécurité Sociale, 2018), (Direction de la Sécurité Sociale, 2019), de nombreux éléments quantitatifs sont communiqués, notamment le pourcentage de personnes couvertes par le régime de général de l'Assurance Maladie Obligatoire pour l'année correspondante à celle du rapport. Le tableau 2.13 présente les pourcentages retenus et extraits de ces rapports.

Table 2.13 : Pourcentage de la population française couverte par le Régimé Général de l'Assurance Maladie Obligatoire en 2018 et 2019.

2018	2019
93%	88%

L'explication de cette diminution est cependant inconnue et ne peut être expliquée sans études complémentaires.

Un traitement de la base INSEE a donc été nécessaire. Il a été effectué sous Excel. En effet, l'objectif de celui-ci est de faire coïncider les tranches d'âges et les régions de la base Open Damir avec celles de la base extraite de l'INSEE. Une fois ces traitements réalisés, un code écrit sous le langage Visual Basic Application (VBA) d'Excel a permis de transformer le format de la base de données INSEE sous un format plus adapté afin de pouvoir la fusionner avec la base de données traitée Open Damir. Chaque ligne indique l'année du soin, du sexe, de la tranche d'âge, de la région et du nombre de bénéficiaires pour ces indicateurs-là. Cet automatisation de tâche permettra un gain de temps pour les études sur la base Open Damir des prochaines années. Un visuel des bases de données créées et modifiées à l'aide de cet outil est disponible en annexes A.3 à A.5. Puis, nous effectuons une jointure entre cette base de l'INSEE modifiée et notre base Open Damir retraitée selon les variables suivantes :

- SOI_ANN (Année du soin),
- AGE_BENEF (Age du bénéficiaire),
- REGION_BENEF (Région de la résidence du bénéficiaire),
- SEXE_BENEF (Sexe du bénéficiaire).

Cette jointure entraîne la suppression des lignes où les modalités de ces variables sont définies comme "Inconnu". La jointure réduit donc le nombre de lignes de notre base d'étude finale.

2.2.5 Bases finales obtenues pour les prochaines études

Le nombre de bénéficiaires est donc fusionné avec la base finale, composée des 24 bases mensuelles retraitées puis agrégées ensembles. Les indicateurs de montants moyens, comme le montant de remboursement obligatoire moyen, le reste à charge moyen ainsi que la fréquence, sont recalculés par ligne. Sur certaines lignes, la quantité d'actes est nulle. Lorsque la quantité d'actes est nulle, les autres indicateurs sont mis à zéro pour éviter toutes anomalies ; les lignes ne sont pas supprimées de manière à en tenir compte dans la tarification.

Le détail du nombre de lignes de chaque base mensuelle après chaque étape de retraitement est donné dans le tableau A.6. Le nombre de lignes a été divisé par 80, ce qui permet de rendre possible les analyses et les calculs des modèles, avec des temps machines réduits. Le nombre de variables a notamment diminué avec la suppression de variables supposées non utiles pour la suite de nos études.

Le nom de certaines variables a été modifié pour les prochains traitements. En effet, certains noms de la base Open Damir ne sont pas assez explicites. Les variables pour lesquelles le nom a été modifié sont les suivantes :

- BEN_SEX_COD a été remplacée par SEXE_BENEF et désigne le sexe du bénéficiaire ;
- AGE_BEN_SNDS a été remplacée par AGE_BENEF et désigne la tranche d'âge du bénéficiaire ;
- BEN_RES_REG a été remplacée par REGION_BENEF et désigne la région de résidence du bénéficiaire ;
- PRS_REM_BSE a été remplacée par BASE_REMB_CORR et désigne la base de remboursement de la Sécurité Sociale ;
- PRS_REM_TAU a été remplacée par TAUX_REMB_CORR et désigne le taux de remboursement d'un soin ;
- FLT_ACT_QTE a été remplacée par QUANTITE_ACTES et désigne le nombre de soins.

D'autres variables calculées ont notamment été créées :

- RAC désigne le reste à charge ;
- RAC_MOY désigne le reste à charge moyen ;
- MNT_REMB_MOY désigne le montant remboursé moyen ;
- DEP_MOY désigne la dépense moyenne ;

- FREQUENCE désigne la fréquence.

Finalement, deux bases de données sont obtenues :

- une base de données Open Damir avec 35 variables et 10 040 857 lignes. Elle nous servira pour les analyses descriptives ainsi que la mise en place de l'outil de reporting santé en section 4.2 ;
- une base de données Open Damir pour la tarification, issue de la base de données finale. Cette base de données est le résultat d'études complémentaires qui seront détaillées dans la partie dédiée à la tarification en chapitre 3.

Un aperçu des données contenues au sein de cette nouvelle base de dix millions de lignes est disponible dans la prochaine partie sous forme de graphiques. Ce sont les statistiques descriptives de la base de données.

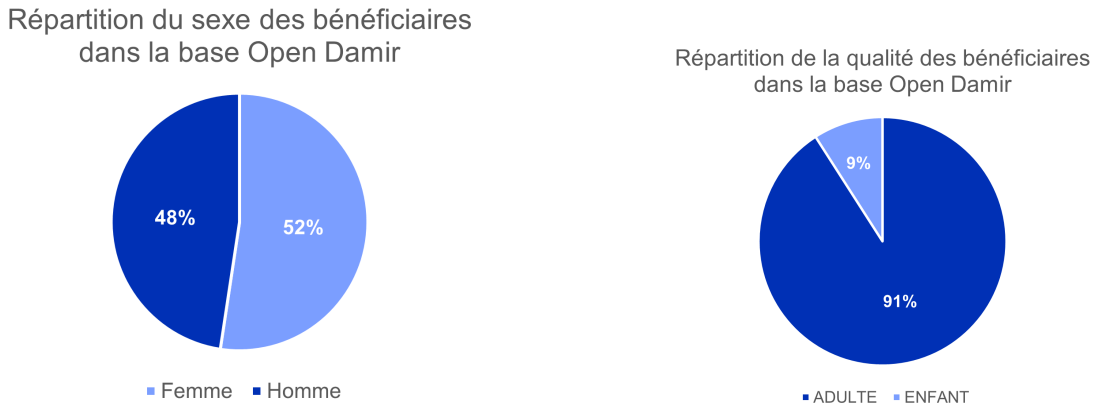
2.3 Statistiques descriptives

Une analyse descriptive de la base de données permet, post-traitement, de connaître la population, les montants et le volume d'actes étudiés. Elle se fait généralement sous la forme de graphiques. Pour cela, nous avons réaliser deux types études : une analyse univariée et une analyse bivariée. Pour rappel, une analyse univariée consiste à regarder le nombre d'occurrences d'une et une seule variable, tandis que l'analyse bivariée permet d'interpréter les tendances d'une variable en fonction d'une autre. Des unités de mesure seront utilisées au sein des graphiques afin de ne pas les surcharger. En effet, la base de données contient 10 040 857 lignes mais le montant des dépenses et des remboursements peuvent représenter des milliards d'euros. La correspondance suivante sera donc utilisée :

- "k" pour les milliers,
- "M" pour les millions,
- "Md" pour les milliards.

2.3.1 Statistiques générales - univariées

Dans un premier temps, les occurrences des variables SEXE_BENEF, QUALITE_BENEF, AGE_BENEF, REGION_BENEF seront étudiées (ce qui correspond respectivement à l'étude des occurences du sexe, de la qualité, de l'âge et de la région des bénéficiaires).



(a) Variable SEXE_BENEF (b) Variable QUALITE_BENEF

FIGURE 2.10 : Analyses univariées

La base Open Damir est constituée à 52% de femmes et 50% d'hommes. Cette proportion est similaire à celle de la population français. Les données sont donc cohérentes. De plus, les bénéficiaires adultes sont majoritaires, ce qui est cohérent avec le fonctionnement de la Sécurité Sociale (c.f. graphiques 2.10).

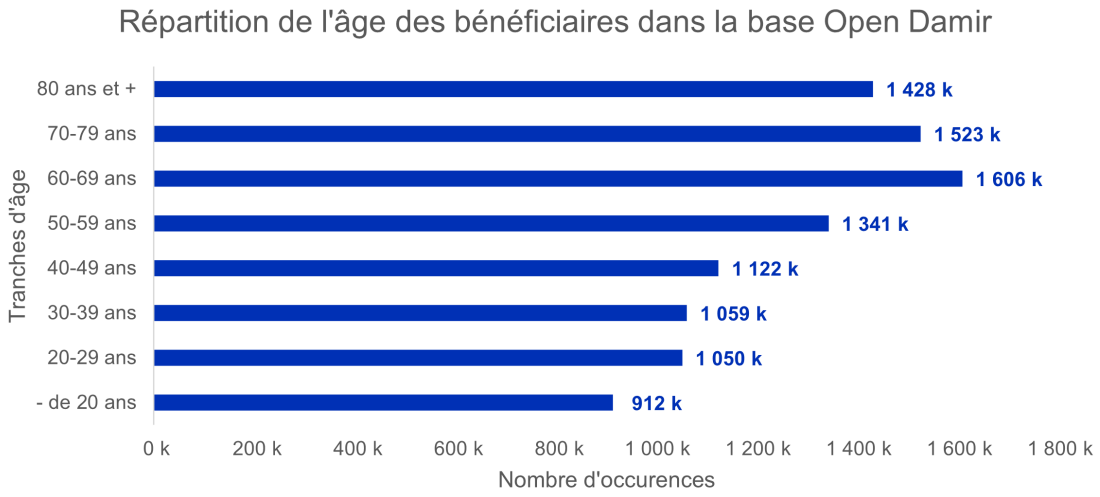


FIGURE 2.11 : Analyse univariée de la variable AGE_BENEF

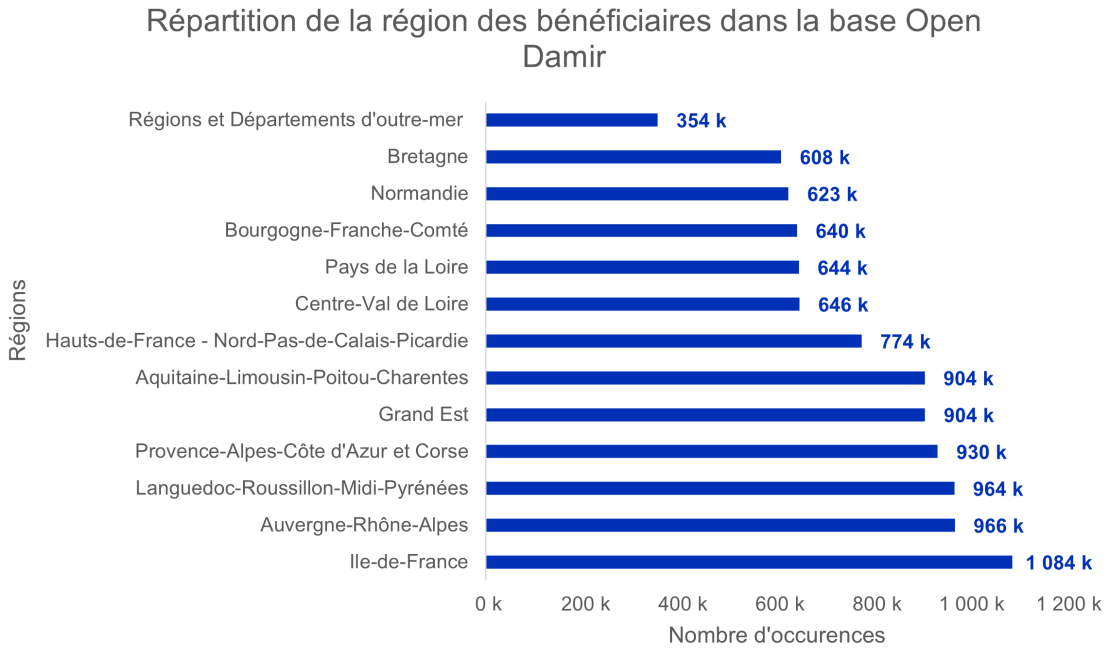


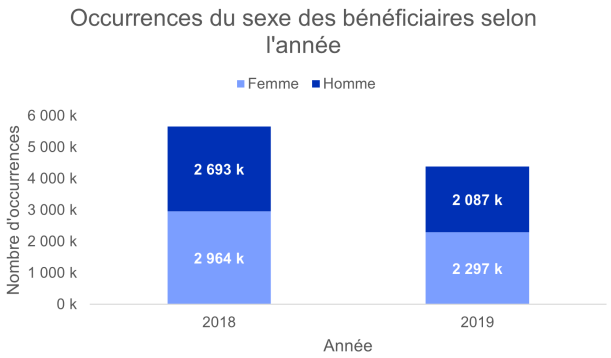
FIGURE 2.12 : Analyse univariée de la variable REGION_BENEF

De plus, les tranches d'âges et les régions avec le plus d'occurrences sont les 60-69 ans et la région d'Ile de France (c.f. les graphiques 2.11 et 2.12). Cela est plutôt cohérent étant donné que pour cette tranche d'âge, la quantité de soins est plus importante, et l'Ile de France est une des régions les plus peuplées de France.

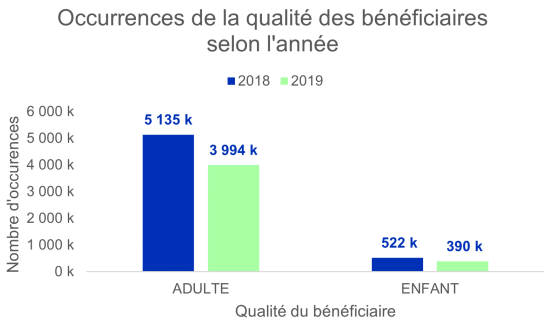
2.3.2 Statistiques bivariées

A présent, les tendances des dépenses, du montant de remboursement total de la Sécurité Sociale, et du nombre de bénéficiaires seront analysées, pour les années 2018 et 2019.

Tout d'abord analysons l'évolution des occurrences pour les variables précédemment étudiées.



(a) Variable SEXE_BENEF



(b) Variable QUALITE_BENEF

FIGURE 2.13 : Analyses selon l'année de soins

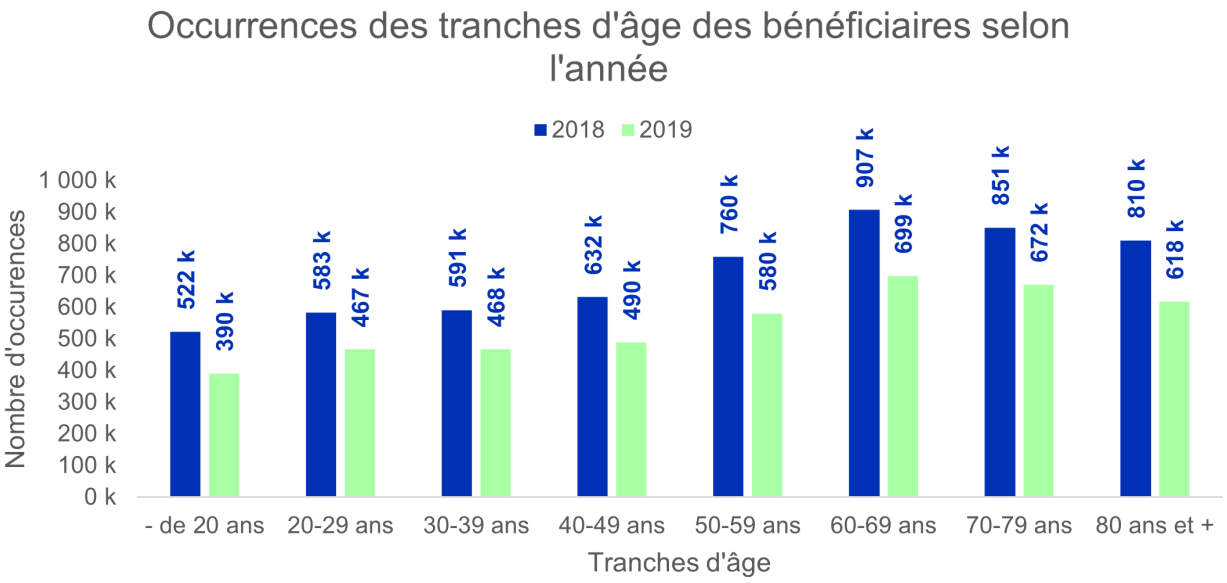


FIGURE 2.14 : Analyse de la variable AGE_BENEF selon l'année de soins

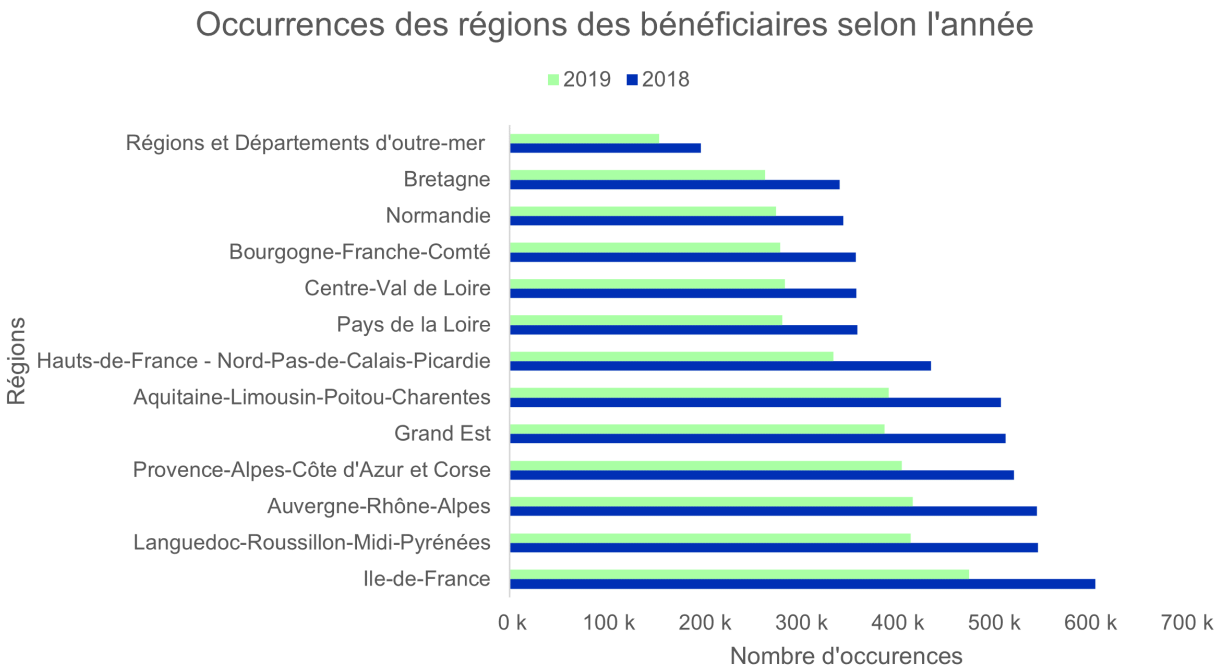


FIGURE 2.15 : Analyse de la variable REGION_BENEF selon l'année de soins

En distinguant par année, une nette diminution des occurrences est observable sur l'année 2019 (c.f. graphique 2.15). En effet, rappelons que les années correspondent aux années de survenance. Or, de nombreuses prestations de survenance 2019 ne sont pas présentes au sein de la base d'étude puisqu'elles ne seront remboursées qu'en 2020. Elles seront donc présentes qu'à partir des bases Open Damir 2020. *A contrario*, la plupart des prestations réalisées en 2018 ont été remboursés en 2019. Cela explique donc cette diminution du nombre d'occurrences pour l'année 2019. Cependant, les tendances sont similaires pour chaque année.

Concernant le nombre de bénéficiaires (c.f. graphique 2.16), une diminution est notamment distinguable. En effet, comme observé lors de l'ajout des données extraites de l'INSEE à notre base d'étude, il s'agit du passage du 93% en 2018 à 88% de personnes couvertes par la Sécurité Sociale en 2019, soit une baisse de 5 points. Enfin, les dépenses et les charges suivent la même tendance sur 2018 et 2019, avec une légère augmentation sur 2019 (c.f. graphique 2.17). En effet, comme expliqué dans les paragraphes précédents, de nombreuses prestations de survenance 2019 sont absentes dans la base d'étude. Afin d'effectuer des comparaisons correctes entre les deux années de survenance, nous devons obligatoirement estimer le montant de ces prestations considérées comme non survenues. Pour cela, des Provisions pour Prestations A Payer (PPAP) sont calculées. Le taux de ces provisions est ensuite intégré dans la base d'études, en multipliant les indicateurs de montant par ce taux selon l'année de survenance et le poste de soins. Une distinction des soins hospitaliers et hors hospitalisation est effectuée. Le calcul des taux de PPAP est détaillé dans la section 4.2.2. De plus, sur le même graphique, nous pouvons remarque que le montant total des dépenses (c'est-à-dire le montant des frais réels des soins réalisés par les bénéficiaires) est plus élevé que le montant total des remboursements de la Sécurité Sociale. Cela s'explique par le fait que les soins ne sont pas remboursés dans leur intégralité. En effet, la majorité des libellés actes pris en charge à 100% par la Sécurité Sociale n'ont pas été pris en compte pour l'étude (c.f. explications à la section 2.2.2). Il est notamment intéressant de remarquer que la charge (remboursement de la Sécurité Sociale) est élevée pour les postes "Soins courants","Pharmacie" et "Hospitalisation" et très peu pour l'optique, l'audiologie et le dentaire (c.f.

graphique 2.20). En effet, la Sécurité Sociale prend peu en charge les dépenses sur ces trois postes. Ils sont majoritairement remboursés par les contrats de complémentaires santé. Pourtant, la consommation restent significatives pour ces trois postes (c.f. graphique 2.19). La réforme 100% permettra une prise en charge plus importante des dépenses sur ces trois postes.

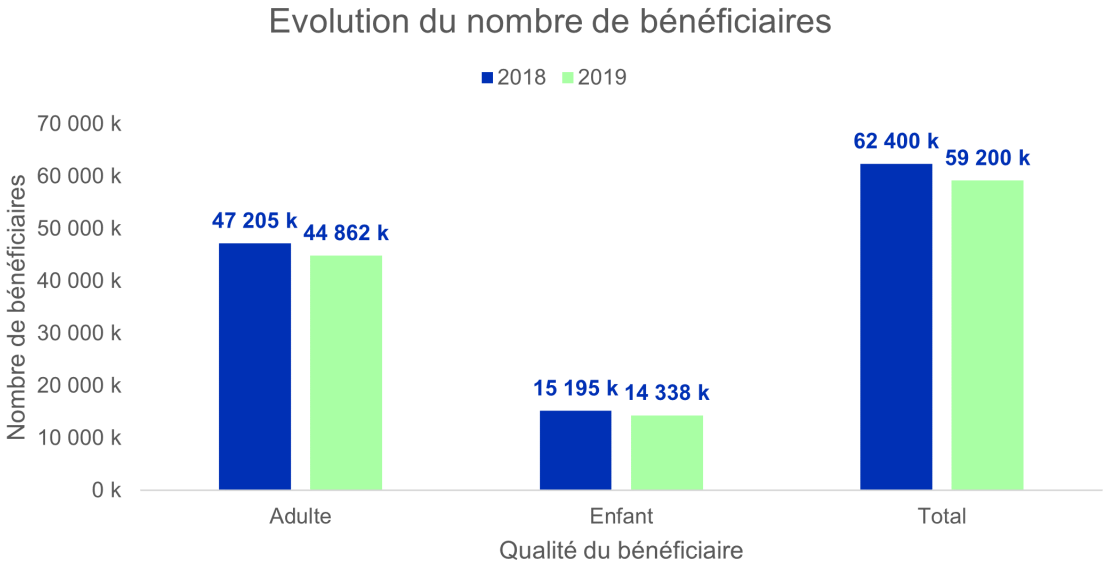


FIGURE 2.16 : Analyse de l'évolution du nombre de bénéficiaires

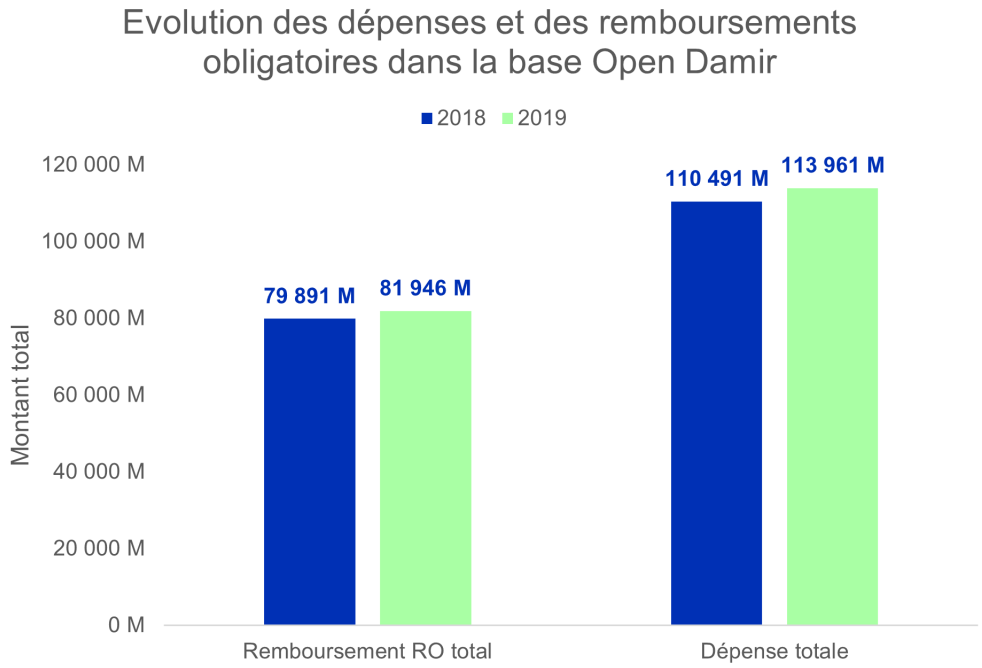


FIGURE 2.17 : Analyse de l'évolution des dépenses et remboursements de la Sécurité Sociale.

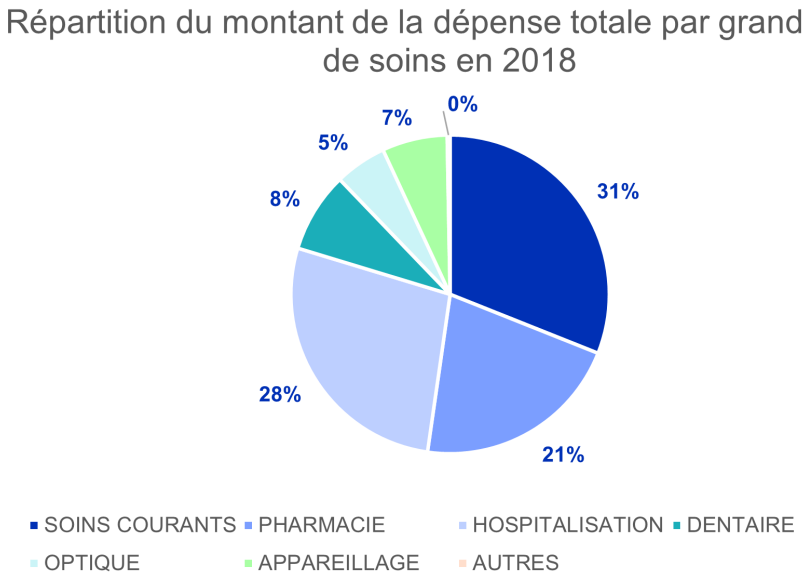


FIGURE 2.18 : Analyse des dépenses par grands postes de soins pour l'année 2018

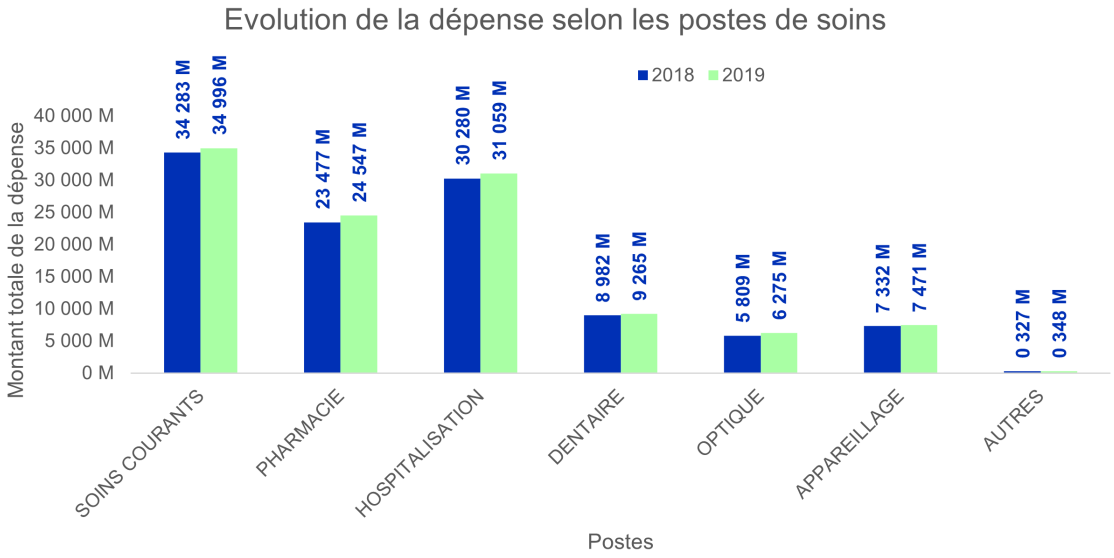


FIGURE 2.19 : Analyse de l'évolution des dépenses selon chaque grand poste de soins.

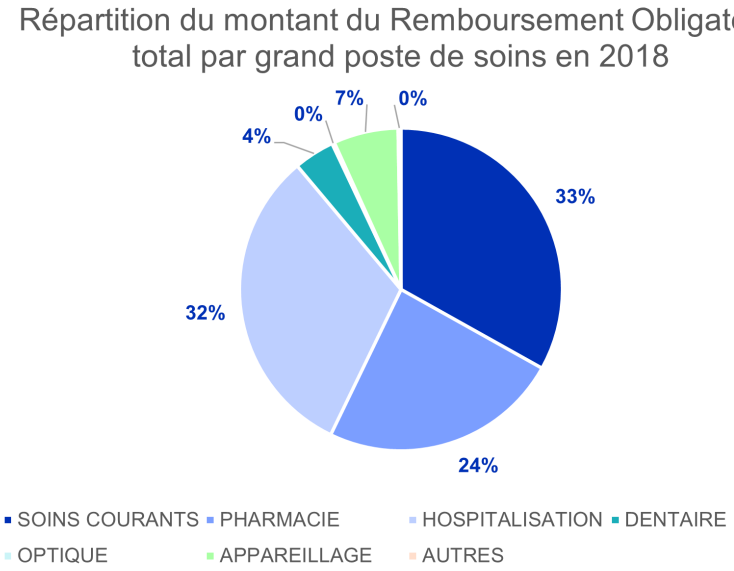


FIGURE 2.20 : Analyse des remboursements de la Sécurité Sociale par grands postes de soins pour l'année 2018

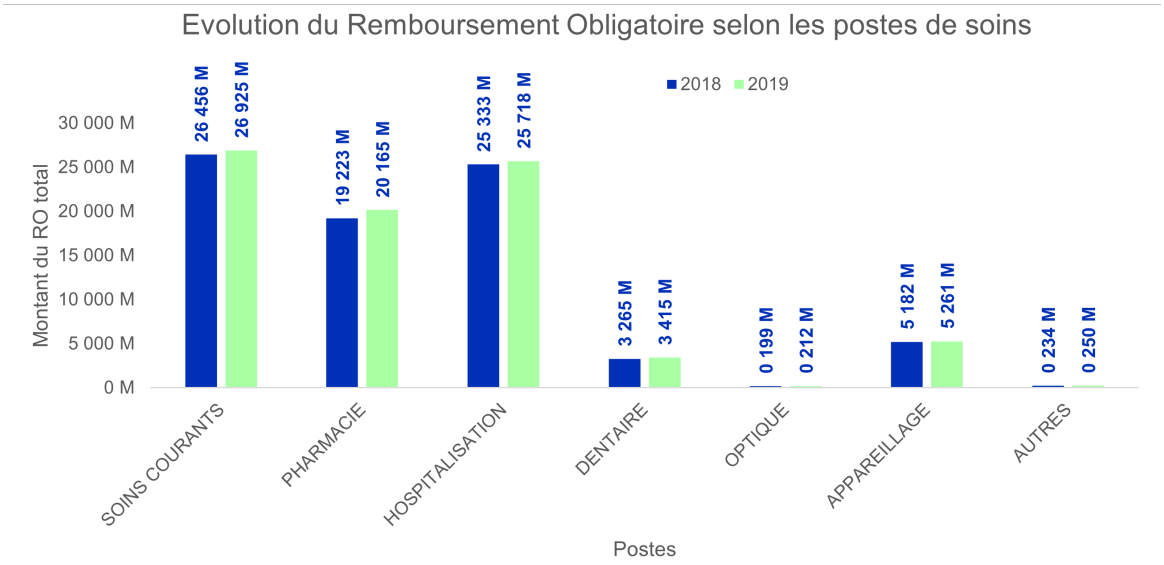


FIGURE 2.21 : Analyse de l'évolution des dépenses selon chaque grand poste de soins.

Chapitre 3

Intégration de la base Open Damir en tarification santé

La base de données utilisée pour l'étude est prête à présent. Le travail de modélisation via un Modèle Linéaire Généralisé, permettant l'obtention d'une tarification de contrat complémentaire santé peut donc débuter. Pour cela, la méthode « Coût moyen × Fréquence » a été choisie. Cependant, il est important de rappeler le principe théorique de cette méthode, les hypothèses à vérifier ainsi que le fonctionnement des Modèles Linéaires Généralisés (MLG).

3.1 Rappels sur la tarification santé

3.1.1 La modélisation « Coût moyen × Fréquence »

Tout d'abord, l'adhérent d'un contrat complémentaire santé paie une prime pure sur une période donnée. Cette prime pure correspond à l'estimation de sa consommation moyenne pour des soins santé sur cette période. La définition de ce montant dans le cadre général et de la modélisation « Coût moyen × Fréquence » est donnée ci-après.

Soit S la variable aléatoire correspondante à la charge totale des prestations de l'adhérent sur une période donnée et π la prime (ou cotisation) pure, déterministe. La charge totale de prestations de l'adhérent est définie par l'équation (3.1)

$$S = \sum_{i=1}^{N} X_i, \tag{3.1}$$

où N est la variable aléatoire à valeurs entières et strictement positives représentant le nombre d'actes santé sur la période considérée et $(X_i)_{i\in\mathbb{N}}$ la suite des prestations individuelles. Les X_i sont positifs et à valeurs réelles. Cette définition de la charge totale des prestations est valide avec l'hypothèse que les X_i sont i.i.d (indépendants et identiquement distribués).

La prime pure se caractérise donc par la formule $\pi = \mathbb{E}(S)$. Dans le cadre de la modélisation « Coût moyen × Fréquence » et d'après les calculs réalisés en annexe B.1, la prime pure est définie selon l'équation (3.2)

$$\mathbb{E}(S) = \mathbb{E}(N) \times \mathbb{E}(X), \tag{3.2}$$

avec :

- $\mathbb{E}(N)$ qui représente la fréquence des prestations ;
- $\mathbb{E}(X)$ qui représente l'estimation du coût moyen de la prestations santé.

L'équation (3.3) ci-dessous définit la variance de la charge totale des prestations S de l'adhérent

$$\mathbb{V}(S) = \mathbb{E}(N)\mathbb{V}(X) + \mathbb{E}(X)^2\mathbb{V}(N). \tag{3.3}$$

La méthode « Coût moyen × Fréquence » va être utilisée dans le cadre de la modélisation GLM, à condition que l'hypothèse d'indépendance soit vérifiée pour la base d'étude.

3.1.2 Les Modèles Linéaires Généralisés, méthode la plus adaptée en tarification santé

En 1972, John Nelder et Robert Wedderbrun développent l'implémentation des Modèles Linéaires Généralisés (MLG), plus communément appelé sous l'intitulé anglais Generalized Linear Models (GLM) (Bel et al., 2016). Cette méthode est une extension du modèle linéaire Gaussien. Pour le traitement des observations, le GLM autorise d'autres lois (conditionnelles) que la loi Gaussienne, appartenant à une famille de lois élargies, qui est la famille exponentielle. De plus, le GLM permet de s'affranchir des postulats établis pour le Modèle Linéaire classique qui sont les suivants :

- la relation linéaire entre la variable d'intérêt et les covariables, c'est-à-dire, $\mathbb{E}[Y_i] = Y_i \times \beta$;
- les observations sont la réalisation d'une variable gaussienne ;
- les Y_i ont la même variance.

Le Modèle Linéaire Généralisé permet donc d'établir une relation non linéaire entre l'espérance de la variable d'intérêt et les variables explicatives en envisageant des observations de nature variée. Cette méthode reste un des outils les plus utilisés dans le domaine de l'assurance santé. En effet, contrairement aux méthodes de Data Science, le GLM dispose de divers avantages. Il permet de modéliser des variables d'intérêt, qu'elles soient réelles ou entières, mais surtout de quantifier l'importance des variables explicatives X_j par des coefficients explicites. C'est pourquoi cette méthode a été choisie dans le cadre de la tarification réalisée sur les bases Open Damir.

Plus précisément, l'objectif du Modèle Linéaire Généralisé consiste à expliquer une variable aléatoire $Y \in \mathbb{R}_n$ en fonction d'un ensemble de p variables explicatives $\mathcal{X} = (\mathcal{X}_1, ..., \mathcal{X}_p)$ pouvant être quantitatives ou qualitatives, selon la formule (3.4) suivante

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta X_{3i} + \epsilon_i, \tag{3.4}$$

avec :

- X qui correspond à la matrice des variables explicatives, et X_{ji} représentant la donnée de la variable j pour l'individu i ;
- β qui correspond à la matrice des coefficients des variables explicatives et β_j représentant le coefficient associé à la variable j ;

- ϵ_i qui correspond au résidu de l'individu i, c'est-à-dire la différence entre la valeur observée et la valeur obtenue par le modèle pour cet individu.

Ces variables explicatives X_j doivent être cependant indépendantes deux à deux Le GLM repose finalement sur trois composantes qui vont être détaillées par la suite :

- la composante aléatoire qui correspond à la variable d'intérêt Y ;
- la composante déterministe qui correspond au prédicteur η ;
- la fonction de lien qui établit cette relation non linéaire entre les deux composantes précédentes.

La composante déterministe

Pour chaque observation $y_i \in [1, n]$, nous disposons d'un p-uplet de vecteurs x_i pour $i \in [1, p]$ qui sont les variables explicatives du modèle. La composante déterministe correspond au prédicteur défini par

$$\eta_i = \sum_{j=1}^{p} X_{ji}\beta_j,$$

où

- le vecteur $\beta = (\beta_1, ..., \beta n)$ contient les coefficients de régression à estimer ;
- le vecteur $\mathcal{X} = (X_j)_{j \in \mathbb{R}_p}$ contient les variables X_j de nature quantitatives ou qualitatives. Cependant, il peut également s'agir de combinaisons de variables quantitatives initiales, telles que $X_k = X_i \times X_j$ pour tout $i, j \in [1, n]^2$ ou X_j^m, avec m une puissance strictement positive, afin d'introduire une représentation polynomiale au sein du modèle.

La composante aléatoire

La composante aléatoire Y est la variable d'intérêt. On note $\mathcal{Y} = (Y_1, ..., Y_n)$ où les densités de $Y_i \in [1, n]$ appartiennent à une famille de lois spécifiques qui est la famille exponentielle naturelle. Si l'appartenance à cette famille n'est pas vérifiée, le GLM ne peut pas être utilisé. Le choix de la loi de probabilité pour les variables aléatoires Y_i au sein de la famille exponentielle naturelle est guidé par la nature du problème et les données à disposition pour l'étude.

Par définition, la famille exponentielle naturelle est une famille de lois de probabilités contenant des lois usuelles et spécifiques comme la loi normale, la loi binomiale, mais aussi la loi de Poisson, la loi Gamma, etc. Ces lois s'écrivent toutes sous la forme exponentielle afin d'unifier les résultats. La variable aléatoire Y a une densité de probabilité, notée f_Y. Cette densité appartient à la famille exponentielle naturelle et s'écrit selon la formule (3.5) suivante

$$f_Y(y) = \exp\left(\frac{y\theta - b(\theta)}{\gamma(\phi)} + c(y, \phi)\right), \tag{3.5}$$

où :

- γ est une fonction définie sur $\mathbb{R}$ et est non nulle ;
- c est une fonction connue, définie sur $\mathbb{R}^*$ et dérivable ;
- b est une fonction connue, définie sur $\mathbb{R}$, trois fois dérivable et sa dérivée première est inversible (c'est-à-dire que sa dérivée première existe) ;
- $\theta \in \mathbb{R}$ est le paramètre naturel de la loi, appelé aussi paramètre canonique ;
- $\phi \in \mathbb{R}$ est le paramètre de nuisance ou de de dispersion.

Si f_Y appartient à la famille exponentielle, alors les égalités suivantes existent

$$\mathbb{E}_\theta[Y] = b'(\theta) = \mu,$$
$$\mathbb{V}_\theta[Y] = b''(\theta)\gamma(\phi).$$

La fonction de lien

L'espérance de la variable d'intérêt est liée aux variables explicatives par une fonction g inversible, appelée fonction de lien, telle que

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

Cette égalité peut être réécrite telle que

$$\mathbb{E}[Y_i] = g^{-1}(x_i\beta) = b'(\theta_i) = \mu.$$

Par définition, la fonction de lien lie notamment le prédicteur linéaire η à la moyenne μ_i. La fonction de lien généralement choisie est la fonction de lien canonique. Elle permet de transformer l'espérance $\mathbb{E}(Y)$ en paramètre naturel, c'est-à-dire

$$g = b'^{-1},$$

soit

$$\theta_i = b'^{-1} \times g^{-1}(x_i\beta) = x_i\beta.$$

D'autres fonctions de lien ne respectent pas cette égalité mais peuvent être choisie comme fonction de lien tant qu'elles correspondent à une bijection de l'espace de $\mathbb{E}(Y)$ dans $\mathbb{R}$. Cependant, en pratique et ce pour des raisons théoriques, la fonction de lien choisie par défaut est la fonction de lien canonique car elle assure la convergence de l'algorithme de Newton Raphson. Cet algorithme permet de déterminer l'estimateur de maximum de vraisemblance par des procédures itératives d'optimisation et d'approcher informatiquement le zéro d'une fonction (la valeur pour laquelle la dérivée de la fonction s'annule). C'est pourquoi la fonction de lien canonique est généralement utilisée dans le cadre du GLM et le sera pour la suite de l'étude. Le tableau 3.1 résume les fonctions de liens des lois de probabilité généralement appliquées pour la variable aléatoire Y, le nom qui lui est attribuée, ainsi que le modèle pour lequel elles sont généralement utilisées.

TABLE 3.1 : Lois et fonctions de lien couramment utilisées pour la réalisation d'un GLM.

Loi de probabilité Y	Fonction de lien	Nom du lien canonique	Pour la modélisation ...
Binomiale	$g(\mu) = logit(\mu) = ln(\frac{\mu}{1-\mu})$	logit	Modèles de conversion / résiliation (0 si l'assuré reste, 1 si l'assuré résilie, par exemple)
Normale (ou Gaussienne)	$g(\mu) = \mu$	identité	Coût moyen
Gamma	$g(\mu) = -\frac{1}{\mu}$	réciproque (ou inverse)	Coût moyen
Inverse gaussienne	$g(\mu) = -\frac{1}{2\mu^2}$	-	Coût moyen
Poisson	$g(\mu) = ln(\mu)$	log	Fréquence / Nombre de prestations
Binomiale négative	$g(\mu) = ln(\frac{\alpha\mu}{1+\alpha\mu})$ (avec α le paramètre d'une loi Gamma)	-	Fréquence / Nombre de prestations

3.1.3 Etapes pour la réalisation d'un GLM appliqué à l'étude

Pour établir une tarification santé, la méthode « Coût moyen × Fréquence » par le Modèle Linéaire Généralisé doit s'effectuer pour chaque libellé brochure. En effet, la consommation des actes santé, et donc les remboursements, ont une tendance différente selon les libellés brochures. La modélisation ne peut donc pas être similaire. Pour rappel, 52 libellés brochure sont présents dans la base d'étude (c.f. section 2.2.2). Après sélection des observations concernées par le libellé brochure, les étapes suivantes sont effectuées pour chacun de ces libellés brochures :

- variables significatives pour la variable d'intérêt ;
- vérification de l'indépendance entre les variables explicatives choisies ;
- choix de la loi ajustée à la variable d'intérêt Y ;
- vérification par le biais de graphiques (densités observée et théoriques, P-P plot, Q-Q plot, etc.) ;
- choix de la fonction de lien ;
- modélisation et validation du modèle par l'analyse des résidus ;
- interprétation des coefficients obtenus par le GLM.

Ces étapes sont effectuées pour la modélisation du coût moyen et de la fréquence et plus particulièrement :

- la dépense moyenne d'une prestation santé, donnée par la variable calculée DEPENSE_MOY ;
- le nombre total de prestations, donné par la variable FLT_ACT_QTE, renommée en QUANTITE_ACTES pour la tarification.

Cependant, certains libellés brochures de notre table de tarification, ne sont pas présents au sein de la base. En effet, ce sont dix postes pour lesquels la prise en charge de la Sécurité Sociale est nulle. Au total, ce sont donc $42 \times 2 = 84$ modèles qui vont être réalisés pour la construction de la tarification santé. Ceux-ci seront modélisés sur la base traitée, sur laquelle les valeurs négatives des montants de frais réels et des quantités d'actes sont supprimées. Elles correspondant à des régularisations de la Sécurité Sociale. Les montants négatifs représentent –114 M € sur les 217 546 M € de frais réels totaux, soit 0,05% du total. Enlever les montants négatifs aura donc un impact négligeable sur notre étude. Enfin, il nous faudra passer des montants de frais réels modélisés au remboursement complémentaire moyen pour la tarification de ce dernier.

Les éléments théoriques ont été donnés. Les étapes de construction de la tarification santé sur la base Open Damir peuvent donc être effectuées.

3.2 Construction d'un tarif en santé à partir de la base Open Damir

Dans un premier temps, pour réaliser une tarification santé à l'aide de la méthode des Modèles Linéaires Généralisés, des études préliminaires sont nécessaires. En effet, pour qu'un modèle soit robuste, il ne suffit pas de lancer l'algorithme sur la base de données brute. De nombreux phénomènes, déjà observés dans l'implémentation de l'algorithme iForest, peuvent impacter négativement la robustesse du modèle, comme le surapprentissage ou le fléau de la dimension. Enfin, avant d'appliquer un modèle « Coût moyen × Fréquence », il est nécessaire de vérifier l'hypothèse d'indépendance, décrite en section 3.1.1. Différents tests et analyses sont donc réalisés pour améliorer d'autant plus la qualité des résultats :

- test d'indépendance pour pouvoir appliquer la méthode « Coût moyen × Fréquence » (c.f. section 3.2.1) ;
- sélection des variables pertinentes pour la modélisation (c.f. section 3.2.2) ;
- identification des lois ajustées aux variables d'intérêts (c.f. section 3.2.3) ;
- application du GLM et analyse des résultats (résidus, indicateurs de performances) (c.f. section 3.2.4).

Pour cela, diverses méthodes seront utilisées pour chacune de ses étapes, afin de vérifier la pertinence et l'adéquation des résultats obtenus. Ces méthodes seront appliquées sur la base de données finale construite pour la réalisation de la tarification santé Damir. Cette base de données est composée de 10 040 857 lignes et de 35 variables, dont 13 variables quantitatives et 22 variables qualitatives.

3.2.1 Test d'indépendance - Coût moyen et fréquence

Premièrement, l'utilisation d'un modèle « Coût moyen × Fréquence » n'est possible qu'après vérification de l'hypothèse d'indépendance de ces deux variables. Pour ce test, deux indicateurs ont été choisis :

- le coefficient de corrélation de Pearson, nommé le R de Pearson ;
- le coefficient de corrélation de Spearman, nommé le Rhô de Spearman.

Explications théoriques des coefficients de Pearson et de Spearman.

Ces deux coefficients de corrélation reposent sur le même principe. Ils permettent de mesurer la corrélation entre deux variables continues ou ordinales. Cette corrélation se détermine par la valeur prise par ces coefficients. Les coefficients de Pearson et de Spearman prennent des valeurs comprises entre -1 et 1. Généralement, les interprétations associées sont les suivantes :

- si le coefficient est égal à 0, cela signifie qu'il n'y a aucune relation entre les deux variables. Ces variables et leurs variations sont donc indépendantes entre elles ;
- si le coefficient est égal à +1, cela signifie qu'il existe une relation positive entre les deux variables. Les deux variables varient dans le même sens ;
- si le coefficient est égal à -1, cela signifie qu'il existe une relation négative entre les deux variables. Les deux variables varient en sens opposé.

Par exemple, les graphiques 3.1 (créés à partir de la libraire *Matplotlib* de $\mathbb{P}ython$) représentent la corrélation entre deux variables choisies parmi celles disponibles dans la base de données traitée. Ces variables ont été choisies après une étude des coefficients de corrélation de Pearson préalable. Le comportement des observations sur un axe en deux dimensions va dépendre de la valeur du coefficient de corrélation.

Le premier graphique montre qu'il existe une corrélation linéaire et positive entre les variables BASE_REMB_CORR et FLT_REM_MNT. En effet, si la valeur de la première variable augmente, la valeur de la seconde augmente aussi. En revanche, sur le second graphique, cette corrélation linéaire est moins importante. Il est toujours possible d'observer cette droite linéaire entre les deux variables. Cependant, de nombreuses observations sont visibles et éloignées de cette droite. Le coefficient est donc moins important. Enfin, aucune linéarité est observée sur le troisième et dernier graphique. Il n'y aucune tendance particulière pour ces deux variables. Le coefficient de corrélation est donc bien nul.

Les formules (3.6) et (3.7) définies pour le calcul des coefficients de corrélation ne sont pas identiques pour le coefficient de Pearson et de Spearman. En effet, le coefficient de Pearson est défini comme (Lemakistatheux, 2013)

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{Cov(XY)}{\sigma_X \sigma_Y}, \tag{3.6}$$

avec :

- X_i et Y_i qui désignent respectivement les valeurs prises par la variable X et Y pour l'individu i ;
- $\overline{X}$ et $\overline{Y}$ qui désignent respectivement la moyenne de la variable X et Y ;

- $Cov(XY)$ qui déisgne la covariance des variables X et Y ;
- σ_X et σ_Y qui désignent les écarts-types des variables X et Y.

Alors que le coefficient de Spearman est défini par

$$r_s(R^1, R^2) = \frac{\sum_{i=1}^{n}(R_i^1 - \overline{R^1})(R_i^2 - \overline{R^2})}{\sqrt{\sum_{i=1}^{n}(R_i^1 - \overline{R^1})^2 \sum_{i=1}^{n}(R_i^2 - \overline{R^2})^2}} = \frac{S_{R^1} + S_{R^2} - \sum_{i=1}^{n}(R_i^1 - R_i^2)}{2\sqrt{S_{R^1} S_{R^2}}}. \tag{3.7}$$

En notant g_1, g_2 le nombre de groupes d'ex-aequos pour $R^1 R^2$, nous avons

$$S_{R^1} = \frac{n(n^2-1) - \sum_{k_1=1}^{g_1}(t_{k_1}^3 - t_{k_1})}{12},$$
$$S_{R^2} = \frac{n(n^2-1) - \sum_{k_1=1}^{g_2}(t_{k_2}^3 - t_{k_2})}{12},$$

avec t_k le nombre d'ex-aequos au sein du groupe k.

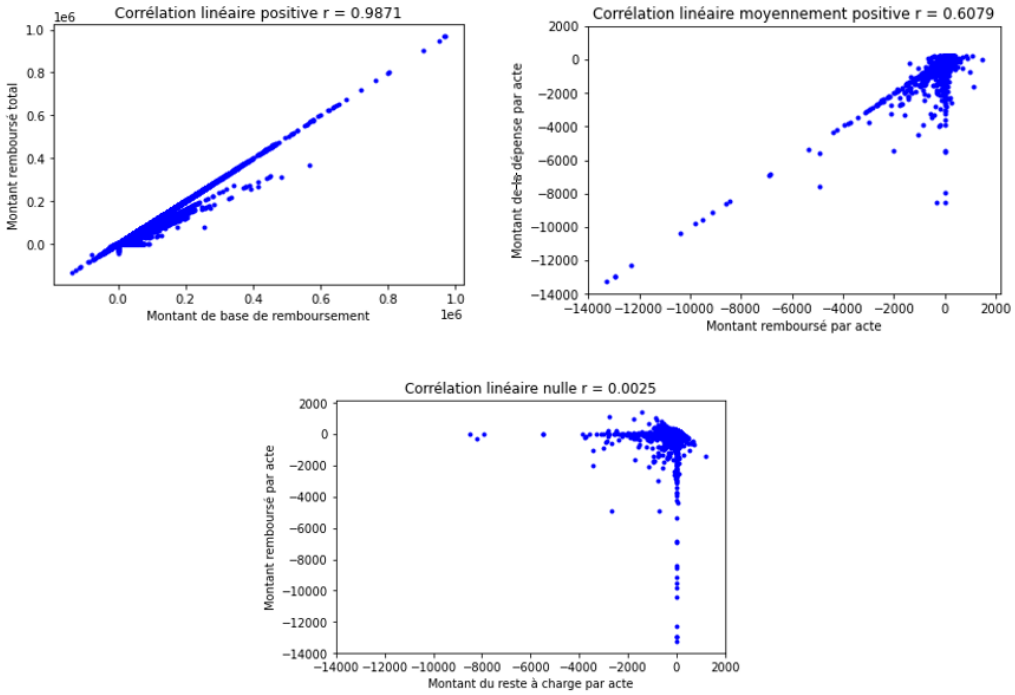


FIGURE 3.1 : Exemple de représentation des corrélations entre deux variables.

La différence entre ces deux calculs implique une différence dans l'intérêt au niveau de l'utilisation de ces coefficients. Le coefficient de Pearson reflète la relation linéaire entre deux variables continues ou ordinales, tandis que le coefficient de Spearman reflète les relations linéaires et monotones pour les deux mêmes types de variables que celui de Pearson. Le schéma explicatif 3.2 permet de comprendre la différence entre ces deux interprétations (Lemakistatheux, 2013).

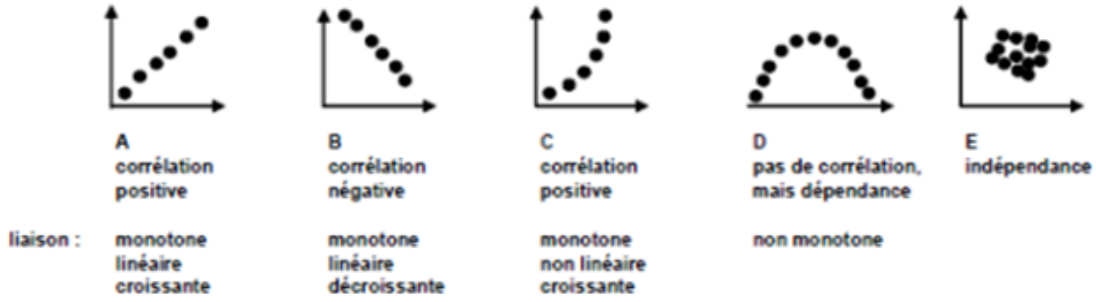


Figure 3.2 : Représentation de l'interprétation des corrélations de Spearman sur un graphique.

Enfin, pour déterminer des coefficients de corrélations corrects, il est essentiel de prendre en compte deux facteurs qui impactent ces résultats, que sont les valeurs extrêmes et le volume de données. En effet :

- les données extrêmes influencent les coefficients de corrélation. La présence d'une observation très différente des autres peut modifier la valeur du coefficient de corrélation de façon significative. Il est donc préférable d'éliminer les valeurs extrêmes de la base de données afin d'obtenir un coefficient qui reflète au mieux la corrélation entre deux variables ;
- lorsque l'indépendance de deux variables continues (ou ordinales) doit être testée sur une base de données volumineuse, il sera préférable de se fier au coefficient de Pearson, contrairement au coefficient de Spearman qui est préféré pour des échantillons de plus petite taille.

Application et résultats

Les coefficients de Pearson et de Spearman ont donc été utilisés pour étudier l'indépendance entre les variables choisies pour la modélisation : la fréquence (variable FREQUENCE) et le coût moyen (variable DEPENSE_MOY). Les coefficients de corrélations ont été déterminés pour chacun des libellés brochure de la table définie dans la section 2.2.2, avec l'élimination au préalable des valeurs extrêmes. En effet, un GLM est réalisé pour chaque libellé brochure dans le cadre de la construction d'une tarification santé.

Ensuite, il est nécessaire de définir un seuil de corrélation au-delà duquel les deux variables seront jugées corrélées. Dans le cadre de cette étude, ce seuil est fixé à 0.25. En effet, d'après les études sur ce sujet, la non corrélation est généralement observée en-dessous d'un seuil égal à 0.5. Cependant, nous avons préféré choisir un seuil plus prudent et égal à 0.25. Cela signifie que :

- si le coefficient est supérieur ou égal à 0.25 (et inférieur ou égal à 1), il existe une corrélation positive entre les deux variables ;
- si le coefficient est compris entre -0.25 et 0.25, les deux variables seront donc considérées comme indépendantes ;

- si le coefficient est inférieur ou égal à -0.25 (et supérieur ou égal à -1), il existe une corrélation négative entre les deux variables.

Les tableaux A.8, A.9, A.10, disponibles en annexe, résument l'ensemble des coefficients de Pearson et de Spearman obtenus pour chaque libellé brochure. D'après les remarques faites précédemment et étant donnée la volumétrie importante de la base Open Damir, nous baserons nos analyses uniquement sur le coefficient de Pearson qui est plus adapté aux échantillons de grande taille. D'après les résultats obtenus pour ce coefficient, la fréquence et le coût moyen sont bien indépendants pour chaque libellé brochure. De plus, nous remarquons que la volumétrie de nos données impacte les résultats du coefficient de Spearman pour lequel l'indépendance n'est pas vérifiée pour certains libellés brochures. Dans le cas où le coefficient de Spearman serait choisi pour la validation de l'hypothèse d'indépendance, il faudrait appliquer une modélisation « Prime pure » pour les libellés pour lesquels l'indépendance n'est pas vérifiée.

Dans ce cas, au vu des résultats obtenus, il sera donc possible d'utiliser le modèle « Coût moyen × Fréquence » pour chaque libellé brochure. Lors de la réalisation du GLM, chaque libellé acte sera donc modélisé par le biais de deux modèles :

- un modèle pour la quantité d'actes, associé à un poids correspondant à l'exposition des bénéficiaires ;
- un modèle pour le coût moyen.

L'hypothèse essentielle étant vérifiée, les autres analyses peuvent être appréhendées.

3.2.2 Sélection des variables pour la tarification

Lors de la réalisation d'un Modèle Linéaire Généralisé, une sélection des variables est indispensable pour améliorer les performances du modèle. Pour cela, des études de corrélations doivent être réalisées. Plusieurs tests de modèles doivent également être effectués afin de ne retenir que les variables les plus significatives. La base de données utilisée pour la tarification comportant 35 variables, ce traitement peut donc s'avérer long. C'est pourquoi une méthode de Machine Learning, appelé LightGBM, va être implémenté afin de retenir les variables les plus significatives pour chacune des variables d'intérêt, modélisées par la suite via un Modèle Linéaire Généralisé. Ces traitements seront toujours réalisés sous $\mathbb{P}ython$.

Utilisation d'une méthode de Machine Learning : LightGBM

La sélection de variables est une étape importante dans la construction de Modèles Linéaires Généralisés ou d'autres modèles. En effet, comme vu en section 2.2.3, le fléau de la dimension est un phénomène à prendre en compte. Le nombre de variables vient impacter les résultats des modèles. Si de nombreuses variables sont choisies, alors le modèle peut surapprendre par rapport aux données. En revanche, il faut un certain nombre de variables minimum afin que la tendance soit captée correctement par le modèle. Plusieurs méthodes statistiques existent afin de contourner ce fléau et sélectionner un nombre suffisant de variables :

- l'Analyse en Composante Principale (ACP),
- la méthode stepwise ou backward,

- d'autres méthodes orientées Machine Learning.

Dans le cadre de l'étude, l'objectif est de traiter et de manipuler un nombre important de données à partir de méthodes innovantes qui acceptent ce volume. Pour la sélection de variables, la méthode de Machine Learning LGBM a été choisie. Mais en quoi consiste cette méthode ?

Explications théoriques de la méthode LGBM Le Light GBM (ou LGBM) est une méthode de Gradient Boosting fonctionnant sur le principe d'apprentissage sur des arbres. A la différence des méthodes de Gradient Boosting, le LGBM construit les arbres de manière verticale. En effet, l'algorithme choisit une feuille enregistrant la plus grande perte au niveau de l'indicateur de performance, et continue cette construction à partir de cette feuille. Le schéma 3.3 compare une construction horizontale des autres méthodes de Gradient Boosting et la construction verticale du LGBM.

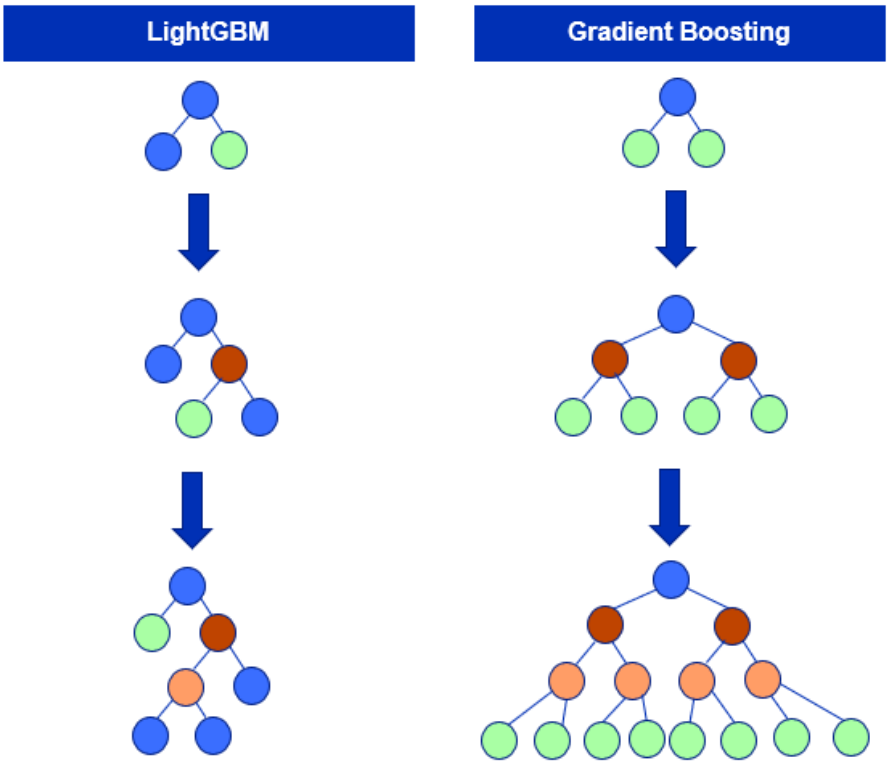


FIGURE 3.3 : Comparaison de la construction des arbres pour la méthode LGBM et des autres méthodes de Gradient Boosting.

Des explications complémentaires sur le fonctionnement du Gradient Boosting sont explicitées par A. SAINI (2021). Mais pourquoi avoir fait le choix d'utiliser la méthode LightGBM et non la méthode XGboost qui est une des méthodes de Gradient Boosting la plus connnue, nommé le XGBoost ? Cela est dû au volume de données sur lequel la méthode de Machine Learning va être appliquée. En effet, sur des ensembles de données volumineux, le LGBM a pour avantage d'être plus rapide et plus précis (ou de précision identique) que le XGBoost. De plus, le LGBM a la capacité d'apprentissage sur plusieurs processeurs en parallèle, ce qui demande moins de mémoire pour fonctionner. Après application de ces deux méthodes sur l'ensemble de données d'étude, les performances obtenues sont formelles, que ce soit pour les méthodes appliquées pour la variable d'intérêt de la quantité d'actes ou pour la dépense moyenne (c.f. le tableau 3.2).

TABLE 3.2 : Performances des méthodes de Machine Learning XGBoost et LightGBM.

Dépense moyenne			
	Score - MSE	Score - RMSE	Temps d'exécution
XGBoost	66.68	8.17	00:35:54
LightGBM	46.66	6.83	00:01:36
Quantité d'actes			
	Score - MSE	Score - RMSE	Temps d'exécution
XGBoost	9917.50	99.59	00:34:28
LightGBM	10000.35	100.00	00:01:21

Le LightGBM prédit mieux ou de manière similaire que le XGBoost et est largement plus rapide (1 minute contre 34 minutes pour le XGBoost). Le LightGBM est donc utile pour un nombre de données important, et non pour une petite base de données, au risque qu'il surapprenne les données. C'est pourquoi, dans le cadre de l'étude sur la base de données d'étude finale, comportant 10 040 857 lignes, il a été choisi d'utiliser le LGBM, afin que les deux critères, robustesse et rapidité, soient réunies pour la sélection des variables. Les résultats obtenus pour le XGBoost seront tout de même présentés, afin de pouvoir comparer la sélection de variables importante effectuée par les deux méthodes de Machine Learning.

Application des méthodes LightGBM et XGBoost et présentation des résultats. Les deux méthodes XGBoost et LGBM, implémentées sur $\mathbb{P}ython$, dans le package Scikit-Learn (J. BROWNLEE, 2021), sont donc appliquées à l'ensemble des données de l'étude, afin de sélectionner les variables importantes, d'une part pour la dépense moyenne et d'autre part pour la quantité d'actes, qui sont les deux variables à modéliser via un GLM. Avant l'exécution de tout algorithme, des retraitements sont nécessaires. En effet, lors de la modélisation GLM, un seuil est appliqué pour l'ensemble des valeurs prises par les variables d'intérêt. Pour être en concordance avec la modélisation GLM, les traitements suivants sont donc effectués :

- pour la variable cible DEPENSE_MOY (représentant la dépense moyenne et dans le GLM, le coût moyen), seules les dépenses moyennes strictement positives sont prises en comptes. De plus, les valeurs extrêmes de cette variable sont supprimées. Seules les observations attritionnelles sont gardées ;
- pour la variable cible QUANTITE_ACTES (représentant la quantité des actes et dans le GLM la fréquence), seules les quantités d'actes positives et nulles sont prises en comptes. De plus, les valeurs extrêmes de cette variable sont supprimées. Seules les observations attritionnelles sont gardées ;

De plus, pour l'implémentation de la méthode, les variables corrélées aux variables d'intérêt sont enlevées. Ces corrélations sont données dans le tableau 3.5. Enfin, la méthode de Machine Learning LGBM est implémentée avec les paramètres et leurs valeurs associées, résumés dans le tableau 3.3.

TABLE 3.3 : Paramètres utilisés pour l'implémentation de la méthode LightGBM.

Nom du paramètre	Descriptif	Valeurs choisies
objective	**Paramètre le plus important.** Il spécifie l'application du modèle, qu'il s'agisse d'un problème de régression ou de classification. *Par défaut : modèle de régression*	regression (modèle de régression)
metric	**Paramètre important.** Il définit la métrique utilisée pour l'évaluation de l'erreur de prédiction. *MSE : erreur quadratique moyenne*	mean_squared_error (MSE)
boosting	Il spécifie le type d'algorithme à exécuter.	bd (arbre de décision Gradient Boosting traditionnel)
num_leaves	Il définit le nombre de feuilles dans l'arbre complet. *Par défaut: 31*	31
learning_rate	Il détermine l'impact de chaque arbre sur le résultat final. *Valeurs récurrentes : 0,1, 0,001, 0,003…*	0.05
num_boost_round	Il définit le nombre d'itérations du boosting à réaliser. *Généralement supérieur à 100.*	500
evals_result	Il s'agit d'un dictionnaire utilisé pour stocker la valeur de la métrique prise pour chaque itération	Dictionnaire vide

Une fois les retraitements effectués et les valeurs des paramètres déterminées, l'exécution des méthodes peut être lancée. Plusieurs étapes et réflexions composent ce traitement :

- les deux méthodes sont exécutées pour chacune des variables d'intérêt afin de comparer les résultats obtenus. Un score est attribué à chaque variable selon l'information qu'elles apportent à la variable d'intérêt en question ;
- les résultats issus de l'exécution sont analysés. Les variables explicatives à utiliser pour la réalisation du GLM sont sélectionnées à partir des résultats obtenus par la méthode LGBM. Le choix est notamment effectué selon l'objectif final de l'implémentation du GLM ;
- pour les variables sélectionnées, une analyse de corrélation est réalisée. Si deux variables importantes sont corrélées entre elles, alors la variable au score le moins élevé est éliminée. Les résultats obtenus en section 3.2.2 sur les corrélations vont donc être utiles pour ces analyses.

Les graphiques 3.4 et 3.5 présentent les 20 variables les plus importantes selon leur score, pour chaque méthode de Gradient Boosting testée et pour les deux variables d'intérêt.

FIGURE 3.4 : Graphiques des 20 variables les plus importantes pour la quantité d'actes.

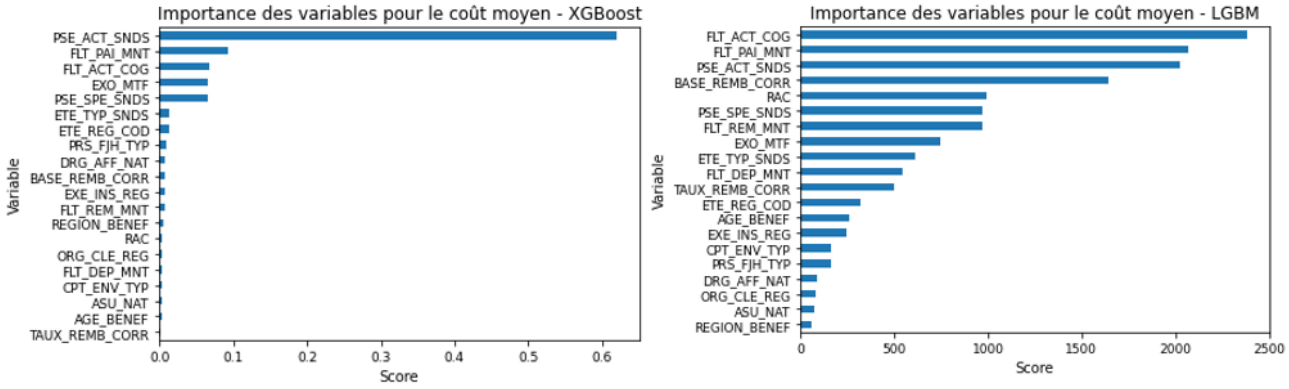


FIGURE 3.5 : Graphiques des 20 variables les plus importantes pour la dépense moyenne.

L'ordre d'importance des 20 premières variables est quasi-similaire entre les deux méthodes. Seuls le score et la position de certaines variables changent. La méthode retenue pour la suite des travaux est la méthode LGBM. Le graphique 3.6 montre l'importance en décroissance des variables, ainsi que l'importance cumulée, pour la modélisation de la quantité des actes, obtenues par la méthode LGBM.

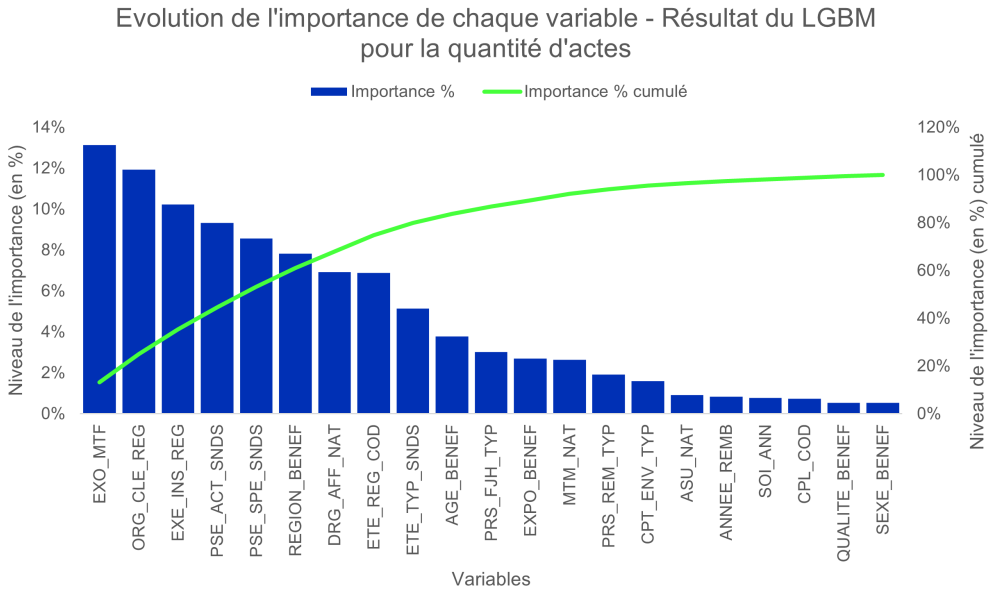


FIGURE 3.6 : Evolution de l'importance de chaque variable obtenue par le LGBM sur la quantité d'actes.

Les premières variables possèdent un niveau d'importance plus significatif que les autres sur la variable d'intérêt de la quantité des actes. L'ajout des variables suivantes apporterait donc peu d'informations significatives pour la variable cible. En effet, la courbe verte montre qu'après la variable AGE_BENEF (donnant les informations sur l'âge du bénéficiaire), le niveau de l'importance cumulée se stabilise. Les variables à sélectionner pour la modélisation GLM doivent donc être limitée à ces premières variables. Finalement, le choix des variables se fera selon une des deux hypothèses suivantes :

- seules les variables qualitatives dont la modalité peut être comparée et utilisée *a posteriori* pour la tarification sont utilisées ;
- seules les variables les plus importantes (jusqu'à celle représentant l'âge du bénéficiaire) sont utilisées.

Pour réaliser la tarification santé dans le cadre de cette étude, la première hypothèse sera retenue. Il faut effectivement sélectionner des variables pour lesquelles des points de comparaison sont possibles avec les caractéristiques d'un assuré. Selon les résultats obtenus précédemment par la méthode LGBM (pour les deux variables d'intérêt), les variables les plus importantes respectant la première hypothèse sont donc les suivantes :

- AGE_BENEF (Age du bénéficiaire),
- REGION_BENEF (Région du bénéficiaire),
- QUALITE_BENEF (Qualité du bénéficiaire),
- SOI_ANN (Année de survenance du soin).

La variable SOI_ANN, qui correspond à l'année de survenance du soin, n'est *a priori* pas significative. Cependant, il est essentiel de l'intégrer à la modélisation GLM, afin que l'inflation entre les deux années 2018 et 2019 soit captée par le modèle pour qu'il soit correctement calibré aux données.

Enfin, il est important de s'assurer que les variables explicatives choisies pour la modélisation GLM ne soient pas dépendantes entre elles. Il est donc indispensable de vérifier cette indépendance. Dans ce cas, seule une variable est choisie parmi les deux variables corrélées, et a le score le plus important parmi les deux. Un lien est donc réalisé avec les dépendances obtenues par le test de V de Cramer, disponible en section 3.2.2, afin de déterminer les variables qui sont éventuellement à supprimer. La qualité de bénéficiaire est dépendante de l'âge du bénéficiaire et son score est moins élevé. La variable QUALITE_BENEF sera donc supprimée. Finalement, après l'étude de corrélation des variables précédentes, les variables choisies pour modéliser la quantité d'actes et la dépense moyenne sont les suivantes :

- **AGE_BENEF** (Age du bénéficiaire),
- **REGION_BENEF** (Région du bénéficiaire),
- **SOI_ANN** (Année de survenance du soin).

La variable **EXPO_BENEF**, qui correspond à l'exposition des bénéficiaires, sera notamment intégré dans la modélisation de la quantité d'actes en tant que poids. De plus, dans le cadre de l'étude, aucune information sur le niveau de gamme du produit souscrit par l'assuré en complément n'est disponible puisqu'il s'agit des remboursements de l'Assurance Maladie Obligatoire et non pas d'organismes complémentaires de santé. Des coefficients d'ajustements seront donc appliqués à la tarification finale.

La partie suivante va détailler les étapes de l'étude des corrélations, préliminaires et intermédiaires à la sélection des variables importantes.

Analyse de corrélation

Dans le cadre de la mise en place du processus de sélection des variables, détaillée précédemment, pour la modélisation GLM des deux variables d'intérêt, deux études de corrélations et de dépendances ont été réalisées, et selon deux types de méthodes :

- étude de corrélation des variables quantitatives, de la base de données via le coefficient de corrélation linéaire, le R de Pearson, théoriquement détaillé en section 3.2.1. Cette étude de corrélation permet de déterminer les variables auxquelles sont corrélées les deux variables d'intérêt, la quantité d'actes et la dépense moyenne ;
- étude de dépendance des variables qualitatives de la base de données via le test de V de Cramer (c.f. les explications données ci-après). Cette étude de dépendance permet de vérifier que les variables explicatives sélectionnées pour la modélisation GLM ne sont pas dépendantes entre elles.

Les valeurs aberrantes des variables d'intérêt QUANTITE_ACTES et DEPENSE_MOY sont éliminées de la base de données d'étude utilisée pour les études de corrélation et de dépendance. En effet, comme vu en section 2.2.3, les coefficients de corrélation peuvent être modifiés en présence de valeurs extrêmes, et peuvent ne pas refléter la réelle corrélation entre deux variables.

Explication théorique du test de V de Cramer Pour analyser des variables qualitatives, le test du Chi-Deux est souvent utilisé. Ce test s'effectue sur un tableau de contingence, représentant la distribution des effectifs des deux variables. Cependant, le résultat de cette statistique de test indique seulement la significativité de la relation entre les deux variables qualitatives, c'est-à-dire si deux variables sont indépendantes ou liées entre-elles. Pour connaitre l'intensité de cette relation, le V de Cramer doit être choisi (Fun Mooc - Grenoble Alpes, 2021). Il utilise les résultats donnés par le test du Chi-Deux, et a pour valeur un coefficient variant de 0 à 1. Cette mesure d'association est donc facilement interprétable. Pour cela, des seuils pour chaque niveau de force de relation doivent être définis. Dans le cadre de l'étude, les seuils définis dans le tableau 3.4 seront choisis.

Table 3.4 : Définition des seuils de dépendance pour le test de V de Cramer.

Valeur du V de Cramer	Type de corrélation	
0	Indépendance	Indépendance
Entre 0.05 et 0.1	Dépendance très faible	
Entre 0.1 et 0.2	Dépendance faible	
Entre 0.2 et 0.4	Dépendance modérée	
Entre 0.4 et 0.8	Dépendance forte	
Entre 0.8 et 1	Dépendance possible	

Pour un V de Cramer inférieur à 0,2, les deux variables étudiées seront donc supposées indépendantes. La dépendance entre deux variables est observée lorsque le V de Cramer est égal à 1. Mais qu'en est-il de la définition de cette mesure d'association ? Comment les valeurs du V de Cramer, détaillées ci-dessus, sont-elles calculées ? Le V de Cramer est défini par

$$V = \sqrt{\frac{\mathcal{X}^2}{\mathcal{X}^2_{max}}} = \sqrt{\frac{\mathcal{X}^2}{n \times (\min(l,c) - 1)}} = \sqrt{\frac{Chi - Deux}{Taille\ de\ l'echantillon \times ddl}},$$

avec n le nombre d'observations, l et c respectivement le nombre de modalités de la première et la seconde variable. $\mathcal{X}^2$ représente le test de Chi-Deux qui détermine la différence entre les résultats théoriques à l'aide des fréquences espérées et des résultats réellement observés à l'aide des fréquences observées. Plus cette différence est marquée, plus la valeur du Chi-Deux est élevée et donc la relation entre les deux variables est significative, c'est-à-dire que les variables sont indépendantes. En effet, le test du Chi-Deux est défini par

$$\mathcal{X}^2 = \sum_{\mathcal{C}} \frac{(|Fréquence\ observée\ -\ Fréquence\ esperée| - 0.5)^2}{Fréquence\ esperée},$$

avec $\mathcal{C}$ qui désigne l'ensemble des cellules du tableau de contingence.

Contrairement au test du Chi-Deux, le test de V de Cramer permet d'éliminer l'impact que peuvent avoir des bases de données volumineuses sur le test du Chi-Deux, par la présence du Chi-Deux maximal. Il reste donc stable si la base de données et le nombre de modalités par variable augmentent simultanément.

Application et analyse des résultats. Les résultats obtenus pour les deux études de corrélation et de dépendance vont donc être présentés.

Dans un premier temps, ce sont 13 variables quantitatives dont deux variables d'intérêt, qui sont concernées par l'étude de corrélation via le coefficient de corrélation linéaire, le R de Pearson. Après implémentation, les coefficients de Pearson obtenus pour chaque variable quantitative sont affichés sur la carte de chaleur 3.7 (ou heatmap).

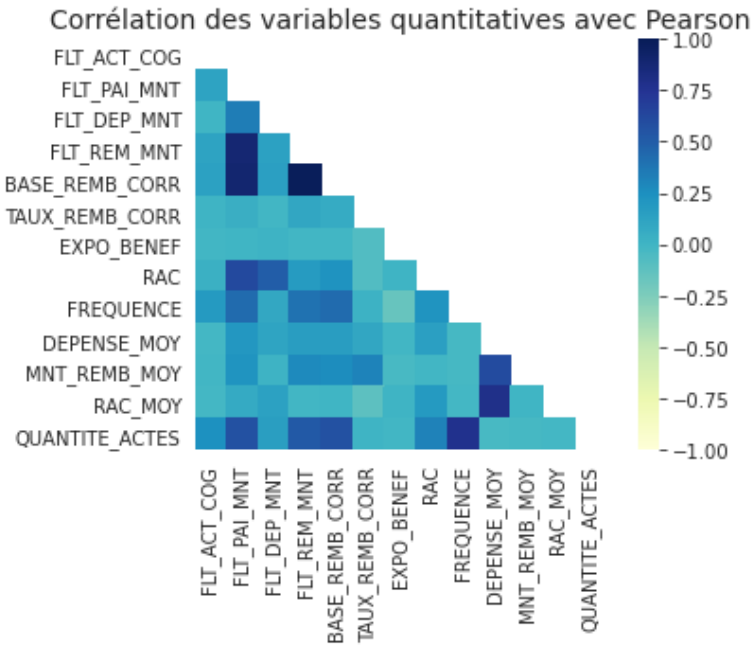


FIGURE 3.7 : Heatmap des coefficients de Pearson pour les 13 variables quantitatives.

D'après la définition des seuils en section 3.2.1, les variables sont supposées liées entre-elles lorsque la valeur du coefficient de corrélation est supérieure à 0.25). Le tableau 3.5 liste les variables quantitatives corrélées aux variables d'intérêt, la dépense moyenne et la quantité d'actes.

TABLE 3.5 : Liste des variables corrélées aux variables d'intérêt.

Dépense moyenne	
Variable	**Corrélation**
MNT_REMB_MOY	0.61
RAC_MOY	0.79
Quantité d'actes	
Variable	**Corrélation**
FLT_PAI_MNT	0.57
FLT_REM_MNT	0.53
BASE_REMB_CORR	0.57
RAC	0.32
FREQUENCE	0.78

Ces variables ne sont donc pas prises en compte dans le processus de sélection de variables. Dans un second temps, ce sont les quatre variables qualitatives, sélectionnées après les analyses des résultats obtenus lors du processus de sélection de variables via le LGBM, qui sont concernées par cette étude de dépendance selon le V de Cramer. Pour cela, les variables catégorielles sont transformées et encodées, ce qui signifie que chaque modalité d'une variable est associée à un entier, variant de 0 aux nombres de modalité de la variable. Après implémentation, les coefficients de V de Cramer obtenus entre chaque variable catégorielle, sont affichés sur la carte de chaleur 3.8 (ou heatmap).

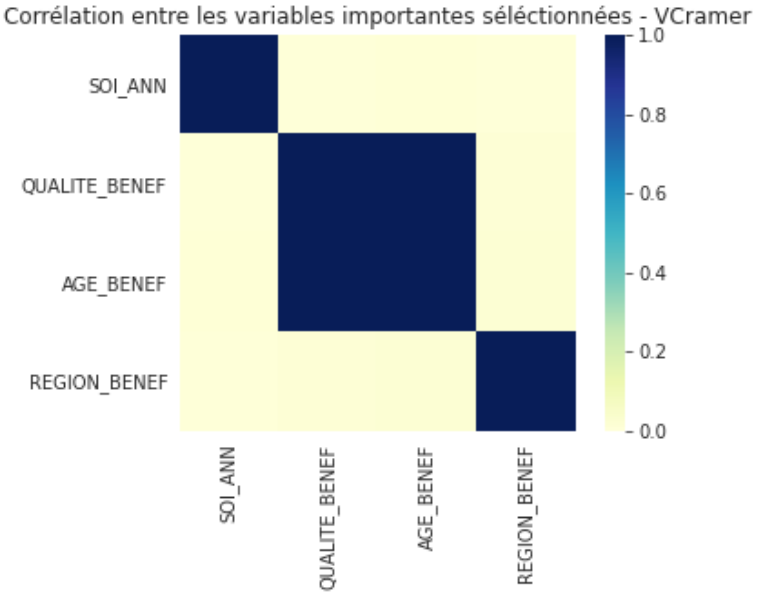


FIGURE 3.8 : Heatmap des coefficients de V de Cramer pour les quatre variables catégorielles choisies.

D'après le graphique 3.8, une dépendance est identifiable entre la qualité et l'âge du bénéficiaire. En effet, le coefficient de V de Cramer vaut 1. Une des deux variables ne doit donc pas être prise en compte pour la modélisation GLM. Comme vu précédemment, la variable AGE_BENEF sera gardée, car son niveau d'importance pour chaque variable d'intérêt est plus élevé que pour la variable QUALITE_BENEF, et notamment plus importante pour l'élaboration de tarification.

Les trois premières étapes pré-réalisation du GLM étant effectuées, l'étude de la quatrième étape peut donc débuter.

3.2.3 Détermination des lois ajustées aux variables d'intérêt

Cette partie consiste à déterminer la loi de probabilité qui ajuste le mieux chacune des deux variables intérêts à modéliser. Les différentes lois possibles correspondent aux lois détaillées dans le tableau 3.1 de la section 3.1. L'implémentation pour la détermination de ces lois s'effectuera via le langage *R*. Pour cela, plusieurs méthodes de détection, de visualisation et de confirmation du choix des lois seront utilisées et présentées dans cette partie.

Détection automatique des lois sur *R*

Deux packages à disposition sur le *R*-Cran vont être utilisés pour la détermination des lois d'ajustements :

- le package « *vcd* » pour la variable d'intérêt « Quantité d'actes » (Meyer D., Zeileis A. et Hornik K., 2021) ;
- le package « *fitdstrplus* » pour la variable d'intérêt « Dépense moyenne » (Delignette-Muller M-L. et Dutang C., 2014).

Tout d'abord, le package « *vcd* » est utilisé pour ajuster des lois à la quantité d'actes de la base d'étude pour chacun des libellés brochures. Il a été implémenté en 2015 par David Meyer, Michael Friendly, Kurt Hornik et bien d'autres créateurs. Ce package met à disposition diverses fonctions, dont la fonction « *goodfit* » qui sera utilisée ci-après afin d'ajuster une distribution discrète, c'est-à-dire correspondant à des données de comptage, à la variable d'intérêt « Quantité d'actes » pour les tests de qualité d'ajustement. Pour une loi donnée, entre la loi de Poisson ou la loi Binomiale négative, la sortie de cette fonction est constituée du nom de la loi, des paramètres qui permettent d'ajuster au mieux cette distribution aux données, mais aussi les données observées et les données prédites à partir de cette loi. Ces sorties sont obtenues selon la méthode du maximum de vraisemblance ou du test du Chi-Deux. Dans le cadre de l'étude, le maximum de vraisemblance a été choisi. Grâce à ces informations, il est possible de calculer un indicateur de qualité d'ajustement, le RMSE, pour les deux lois possibles. Le RMSE minimal déterminera alors la loi qui s'ajuste le mieux aux données. Cette fonction permet la prise de décision au niveau du choix de la loi de façon plus automatique. Il est notamment possible de tracer les résultats de la fonction « *goodfit* ». Ces graphiques sont tout de même utiles pour une vérification *a posteriori*, pour confirmer ou modifier les résultats obtenus, si souhaité.

Dans un second temps, le package « *fitdstrplus* » est utilisé pour ajuster des lois à la dépense moyenne de la base d'étude pour chacun des libellés brochures. Ce package a été implémenté par Delignette-Muller, Pouillot, Denis et Dutang en 2014 dans l'objectif de rendre l'ajustement de lois aux données plus rapide et automatique. En effet, il s'agit d'une étape incontournable en statistiques et avant une modélisation. Trouver la distribution la plus ajustée pour modéliser une variable d'intérêt ainsi que les paramètres associés, constitue une étape longue et itérative, notamment en tarification santé où de nombreuses modélisations (84 modélisations dans le cadre de ce mémoire) sont effectuées. Cette étape nécessite d'essayer plusieurs lois, plusieurs paramètres puis de comparer les résultats obtenus pour les différents modèles afin d'évaluer la qualité de l'ajustement et acter le choix de la loi la plus adaptée. Ce package met à disposition cette expertise à travers diverses fonctions disponibles. Ces fonctions permettent de déterminer la loi qui s'ajuste le mieux aux données. Par exemple, la première fonction utilisée est la fonction « *fitdstr* » qui estime les paramètres de distribution en maximisant la fonction de vraisemblance. Le retour de cette fonction est composé de plusieurs éléments

comme la loi choisie, les paramètres associés, le Critère d'Information d'Akaike (AIC en anglais) et le Critère d'Information Bayésien (BIC en anglais) évalués pour chacune des lois théoriques, etc. Tous ces éléments permettent la prise de décision de manière plus automatique de la loi la plus ajustée. Ce package comporte notamment des fonctions qui permettent d'afficher différents types de graphiques. L'utilité de ces graphiques pour une vérification *a posteriori* sera détaillée dans les paragraphes suivants.

Résultats obtenus pour la variable d'intérêt « Quantité d'actes »

Les deux lois de probabilités choisies et possibles pour la variable d'intérêt Y, qui représentent la quantité d'actes, sont la loi de Poisson et la loi Binomiale négative. Théoriquement, pour des valeurs de paramètres quelconques, les densités des deux lois sont respectivement celles représentées sur le graphique 3.9.

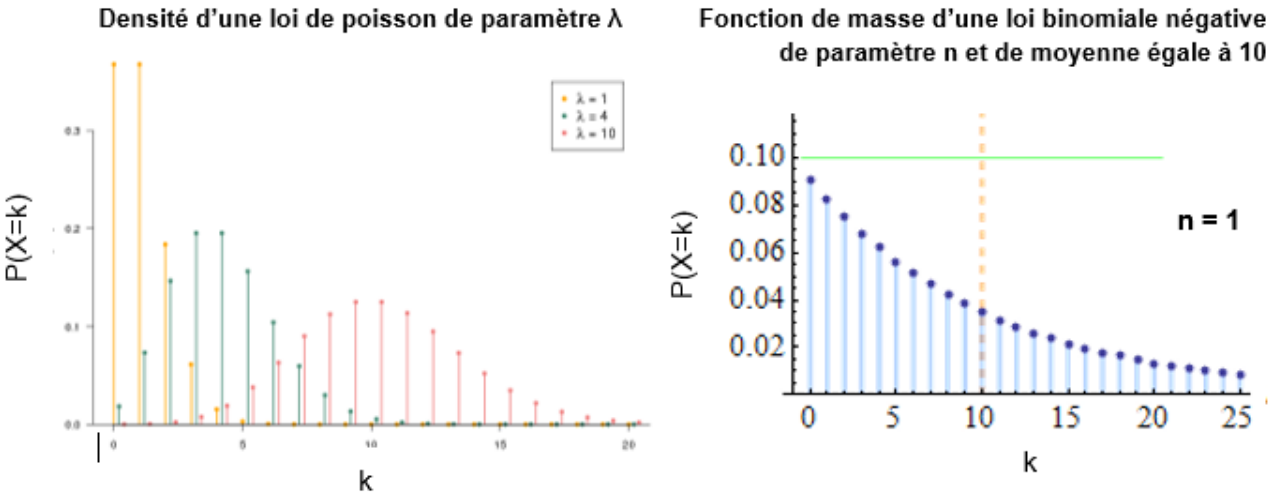


FIGURE 3.9 : Graphique des densités des lois de Poisson et Binomiale négative.

Après application de la fonction « *goodfit* » et l'analyse du RMSE pour l'ensemble des libellés brochures, la loi Binomiale négative est majoritairement sélectionnée pour la variable d'intérêt « Quantités d'actes », sauf pour le libellé « Soins à l'étranger » pour lequel la loi de Poisson s'ajuste mieux à la variable d'intérêt. Les résultats obtenus, lors de l'utilisation de la fonction « *goodfit* », indiquent, pour chacun des libellés brochures, la loi choisie ainsi que les paramètres associés les plus adaptés en fonction des observations de la base. Ces résultats sont par la suite vérifiés graphiquement. Un histogramme de la répartition de la variable d'intérêt ainsi que le graphique extrait de la fonction « *goodfit* » sont tracés. Les résultats obtenus seront présentés seulement pour deux libellés brochures :

- les « actes techniques médicaux » pour la loi binomiale négative ;
- les « soins à l'étranger » pour la loi de Poisson.

Les graphiques extraits par la fonction « goodfit » (M. FRIENDLY, 2000), testés avec les deux lois pour un même libellé brochure, permettent de vérifier que la loi choisie s'ajuste correctement aux données, ou mieux qu'avec l'autre loi. Les graphiques 3.10 et 3.11 représentent les résultats obtenus pour ces deux libellés.

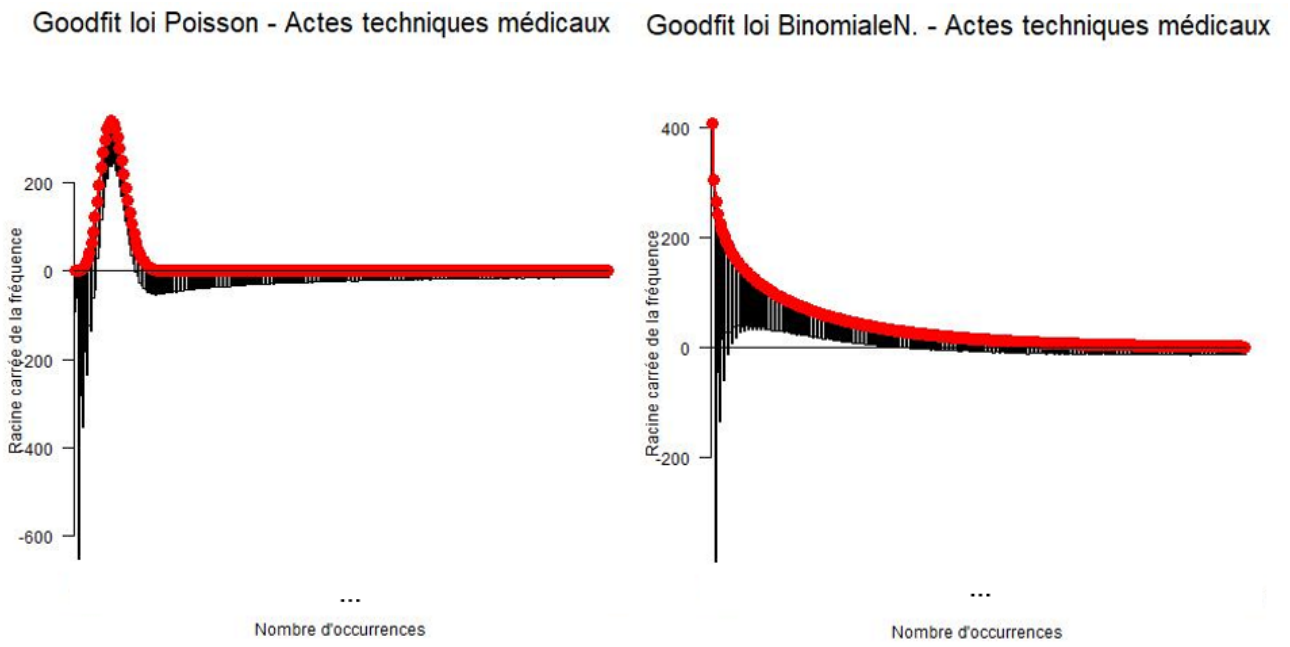


FIGURE 3.10 : Graphiques extraits de la fonction « *goodfit* » pour le libellé « Actes techniques médicaux » .

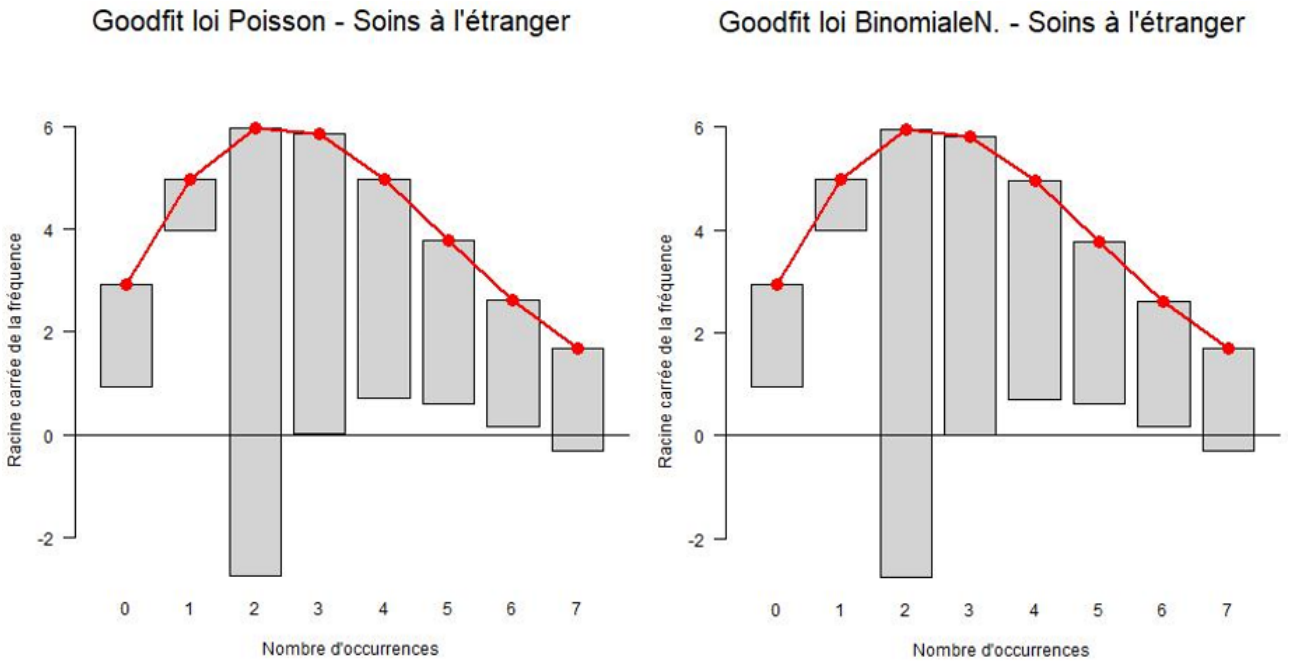


FIGURE 3.11 : Graphiques extraits de la fonction « *goodfit* » pour le libellé « Soins à l'étranger » .

Pour le libellé « Actes techniques médicaux », la loi Binomiale négative est effectivement plus adaptée aux observations. En effet, les écarts entre l'axe des abscisses et l'extrémité des bâtons sont moins importants sur le graphique « *goodfit* » de cette loi par rapport à ceux obtenus pour la loi de Poisson. Cependant, les graphiques de la fonction « *goodfit* » se révèlent non pertinents pour le libellé « Soins à l'étranger ». En effet, aucune différence n'est identifiable sur les deux graphiques. Le

graphique 3.12 de la répartition de la quantité des actes pour les libellés « Actes techniques médicaux » et « Soins à l'étranger » va donc permettre de confirmer le choix initial.

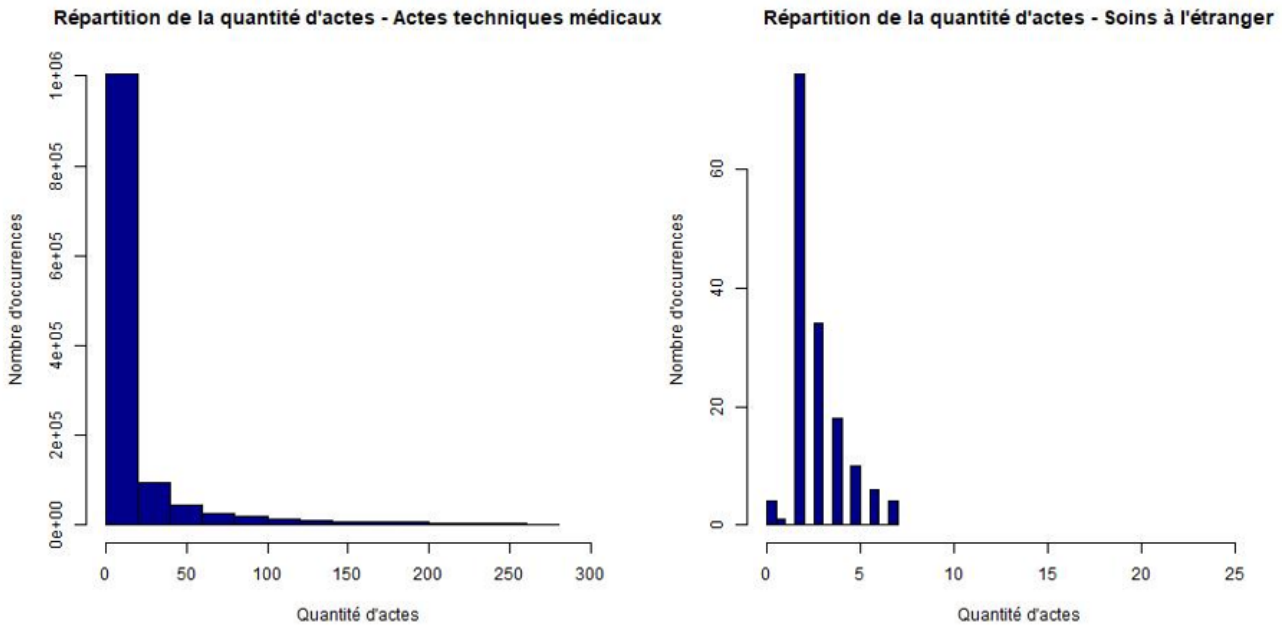


FIGURE 3.12 : Répartition de la quantité d'actes pour les libellés « Actes techniques médicaux » et « Soins à l'étranger ».

Pour les « actes techniques médicaux », la répartition est similaire à la densité d'une loi Binomiale négative. Cette loi sera donc gardée pour ce libellé. Pour les « soins à l'étranger », il s'agit effectivement de la densité d'une loi de Poisson. Cette loi sera donc gardée pour ce libellé. Les lois choisies pour l'intégralité des libellés brochures et la variable d'intérêt « Quantités d'actes » sont données en annexes A.11 à A.13.

Résultats obtenus pour la variable d'intérêt « Dépense moyenne »

La démarche précédente est répétée afin d'identifier les lois de probabilités pour la variable d'intérêt « Dépense moyenne » et pour chacun des libellés brochures. Cependant, d'autres fonctions seront utilisées pour cette identification. Elles sont issues du package « *fitdstrplus* » disponible sur *R*.

Les trois lois de probabilités choisies et possibles pour la variable d'intérêt Y, qui représentent la dépense moyenne sont la loi de Weibull, la loi Lognormale et la loi Gamma. Le choix de la loi est effectué selon la valeur de l'AIC et du BIC. La loi choisie correspond aux valeurs de l'AIC et BIC minimales entre les trois lois. A l'aide du package « fitdstrplus », et pour chaque libellé brochure, plusieurs graphiques sont tracés afin de vérifier que le résultat du choix de la loi est correct :

- la densité observée de la dépense moyenne,
- la fonction de densité cumulée observée de la dépense moyenne,
- P-P plot (Probabilités observées en fonction des probabilités théoriques),
- Q-Q plot (Quantiles observés en fonction des quantiles théoriques).

Pour chacun de ces graphiques, la courbe observée est confrontée aux courbes théoriques des trois lois possibles. La loi qui s'ajuste au mieux à la dépense moyenne observée est celle pour laquelle la courbe théorique dans chacun des graphiques est la plus proche de la courbe empirique. De plus, les résultats obtenus seront présentés pour les quatre libellés brochures uniquement :

- les « Analyses médicales » pour la loi Weibull,
- les « Actes techniques médicaux » pour la loi Lognormale,
- le libellé « Grand appareillage » pour la loi Gamma,
- les « Consultations de généralistes ».

Après application des méthodes précédemment décrites, deux des quatre graphiques sont présentés pour ces quatre libellés brochures.

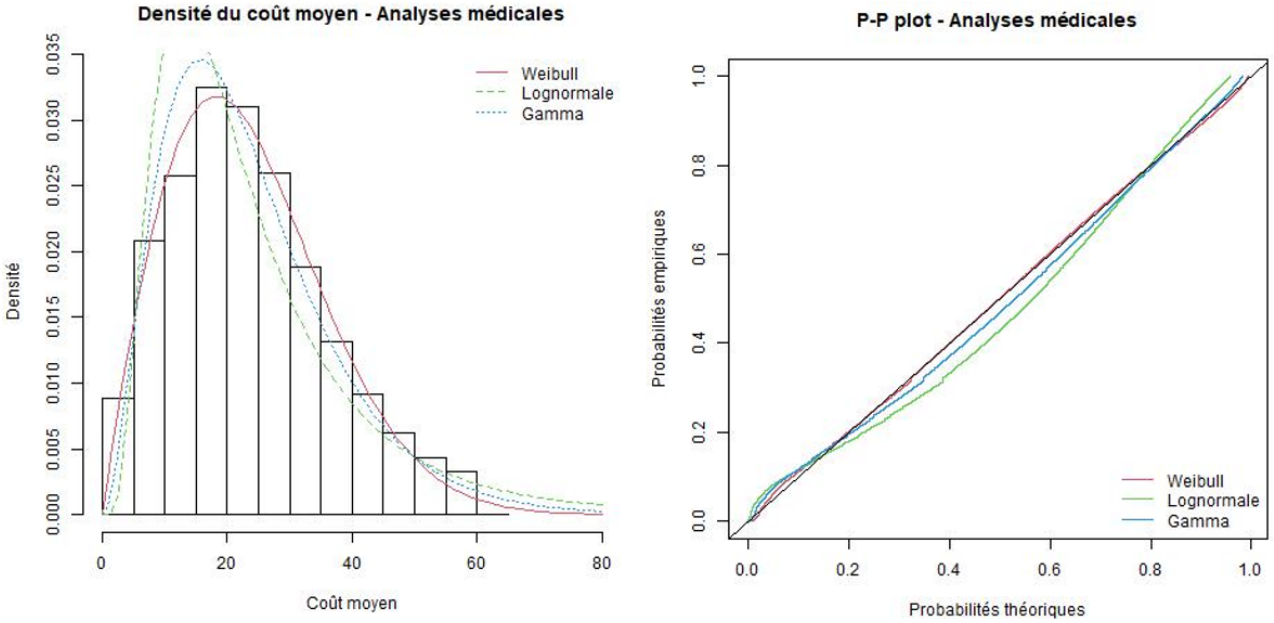


FIGURE 3.13 : Densité et P-P plot de la dépense moyenne pour le libellé « Analyses médicales ».

Pour les analyses médicales, la loi Weibull est confirmée d'après les densités observées et théoriques et d'après le graphique P-P plot 3.13. La courbe théorique de la loi Weibull respecte la densité observée et est totalement superposée à la ligne théorique pour le graphique P-P plot 3.13.

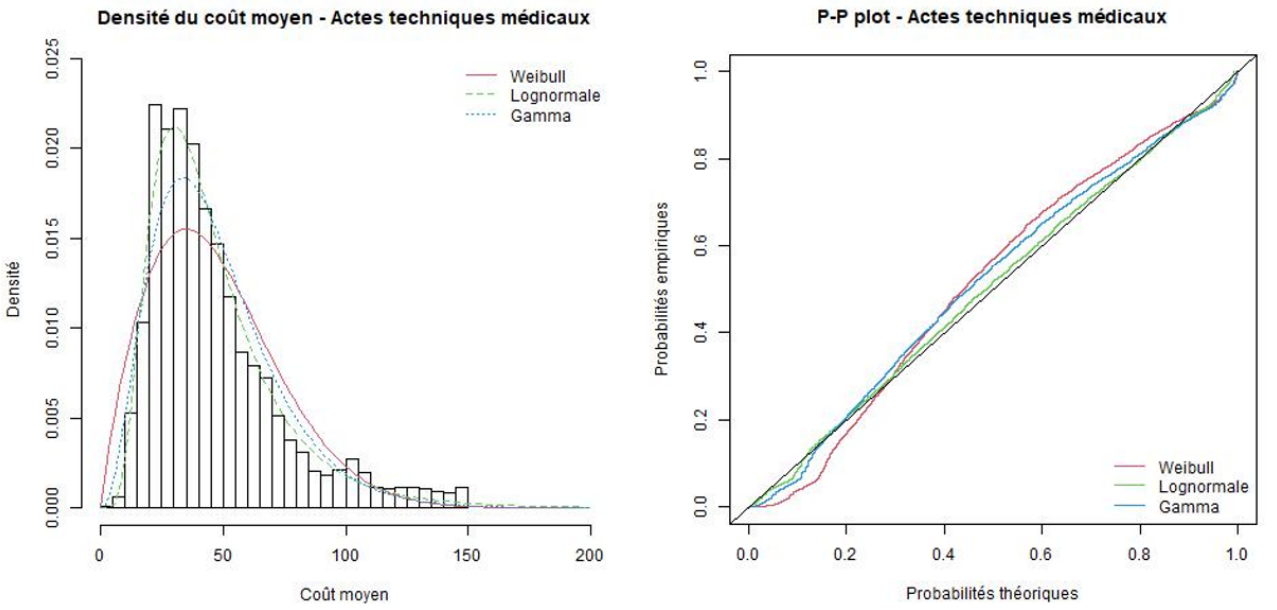


FIGURE 3.14 : Densité et P-P plot de la dépense moyenne pour le libellé « Actes techniques médicaux ».

Pour les actes techniques médicaux, la loi Lognormale s'ajuste effectivement mieux que les autres lois. Le sommet de la densité ainsi que la queue de la distribution est similaire à la densité de la loi Lognormale. De plus, la courbe théorique de la loi Lognormale est plus proche que les autres courbes théoriques à la courbe empirique sur le graphique P-P plot 3.14.

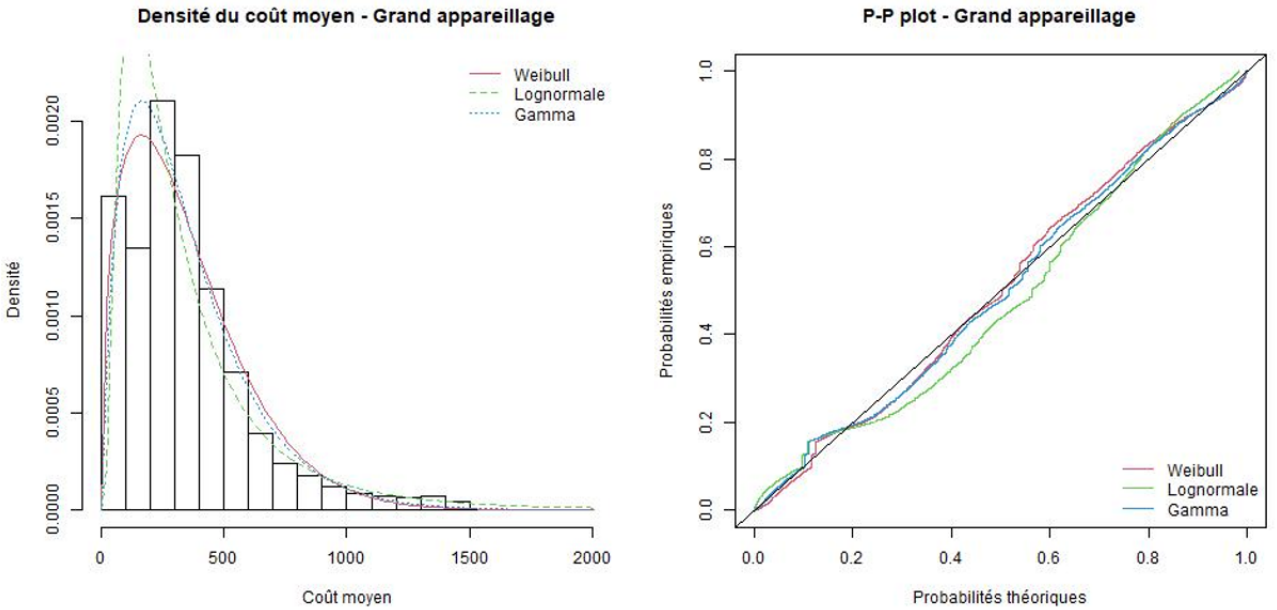


FIGURE 3.15 : Densité et P-P plot de la dépense moyenne pour le libellé « Grand appareillage ».

Enfin, pour les grands appareillages, la loi Gamma s'ajuste effectivement mieux que les autres lois. Les sommets des densités théorique et observée sont au même niveau. De plus, la courbe théorique de la loi Gamma est plus proche que les autres courbes théoriques à la courbe empirique sur le graphique

P-P plot 3.15.

Ces graphiques doivent nécessairement être visualisés afin de confirmer ou non la loi choisie. Cette nécessité provient du fait que pour certains libellés, la loi n'est tout de même pas adaptée après l'application de ces méthodes. En effet, pour le libellé brochure « Consultations de généralistes », ce phénomène est visible et détectable avec le tracé des graphiques 3.16.

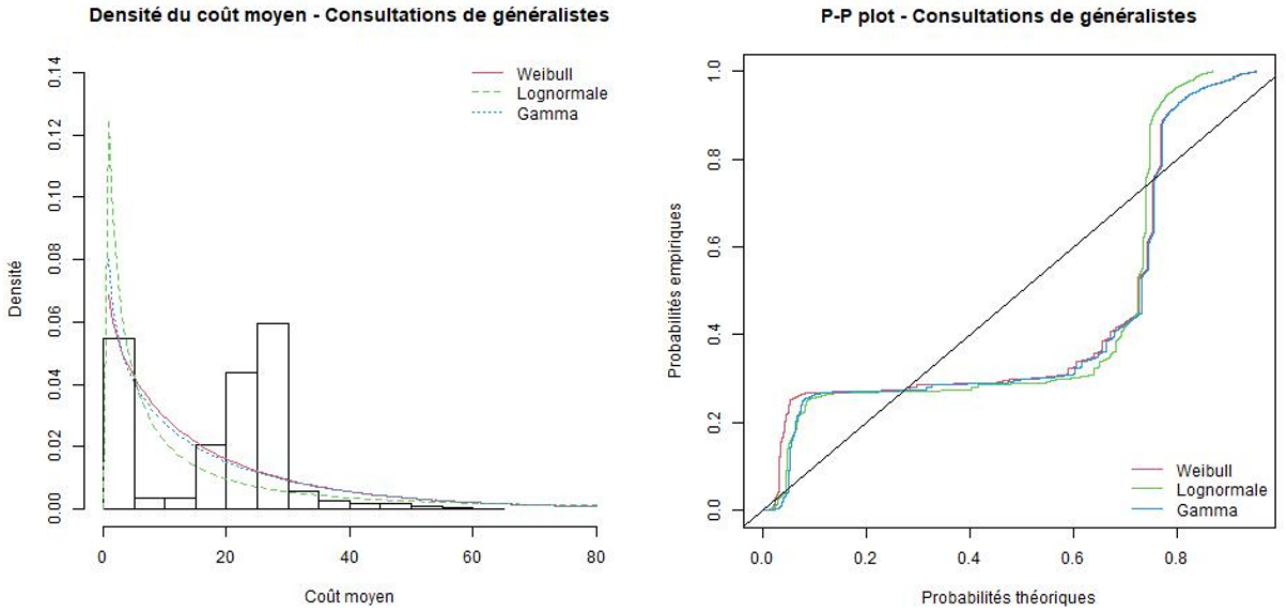


FIGURE 3.16 : Densité et P-P plot de la dépense moyenne pour le libellé « Consultations de généralistes ».

Sur le graphique de la densité, deux pics sont distinguables. Aucune des trois lois possibles n'est adaptée pour cette répartition de la dépense moyenne. De plus, le graphique P-P plot indique un grand écart entre l'empirique et le théorique. Dans ce cas-là, la dépense moyenne sera modélisée en deux temps, avec un découpage au niveau des deux pics présents sur le graphique de densité.

En conclusion, un de ces graphiques doit être au minimum tracé pour effectuer ces vérifications et adapter la modélisation. Les lois choisies pour l'intégralité des libellés brochures et la variable d'intérêt « Dépense moyenne » sont données en annexes A.11 à A.13.

Remarque : Les algorithmes choisissent la loi qui s'ajuste au mieux sur les données parmi les lois sélectionnées. Nénanmoins, ce n'est peut-être pas la meilleure loi par rapport à d'autres existantes. D'autres lois pourraient mieux s'ajuster avec la quantité d'actes ou bien la dépense moyenne. Pour cela, un processus automatique peut être mis en place. Ce processus testera toutes les lois existantes et déterminera la loi la mieux ajustée. Cependant, ce processus doit être exécuté pour chaque libellé brochure et pour les deux variables d'intérêt, soit 82. Le temps d'exécution serait donc très long au vu de la volumétrie de la base de données de l'étude. Il est donc impossible de le réaliser dans le cadre de ce mémoire. Il sera implémenté en entreprise pour améliorer cette étape.

3.2.4 Réalisation du modèle sur la tarification santé

Cette partie a pour objectif de présenter les différentes étapes effectuées pour la tarification santé avec les bases Open Damir. Pour la construction de cette tarification, les principaux points traités sont :

- la construction du GLM pour l'obtention des coefficients GLM qui serviront à la tarification ;
- l'analyse des résultats obtenus par les GLM : ces analyses permettront de valider le modèle et d'évaluer sa robustesse et sa fiabilité. La significativité des variables et le choix des combinaisons de variables quantitatives seront notamment abordés ;
- l'interprétation des coefficients GLM obtenus.

Deux modélisations seront effectuées, d'une part pour la dépense moyenne et d'autre part pour la quantité d'actes, afin d'obtenir la tarification de la prime pure selon la méthode « Coût moyen x Fréquence » (c.f. section 3.1.1). De plus, ces modélisations seront effectuées pour l'ensemble des libellés brochures. La démarche choisie et les résultats obtenus pour ces modélisations seront uniquement explicités pour le libellé brochure « Actes d'anesthésie ». Enfin, pour établir une tarification de contrats de complémentaires santé, quelques traitements supplémentaires seront nécessaires pour passer d'une modélisation de la dépense moyenne (correspondant au montant moyen des frais réels pour un acte santé) au remboursement complémentaire moyen. Ce point sera abordé en dernière partie de la modélisation GLM de la dépense moyenne.

Etude de la modélisation de la dépense moyenne

Nous rappelons que les variables utilisées pour la modélisation de la dépense moyenne sont les suivantes :

- **Variable d'intérêt :** DEPENSE_MOY (Dépense moyenne) ;
- **Variables explicatives quantitatives :** AGE_BENEF (Age du bénéficiaire) ;
- **Variables explicatives catégorielles :** REGION_BENEF (Région de la résidence du bénéficiaire) et SOI_ANN (Année de survenance du soin).

Présentation des choix effectués pour les actes d'anesthésie

La loi choisie, qui ajuste au mieux la dépense moyenne des « Actes anesthésie », est la loi Log normale. En effet, il s'agit d'une loi de distribution dans $\mathbb{R}_+$ et les dépenses moyennes sont positifs. De plus, la fonction de densité et de la densité cumulée sont confondues avec la fonction de densité théorique de la loi Log normale (c.f. graphique 3.17).

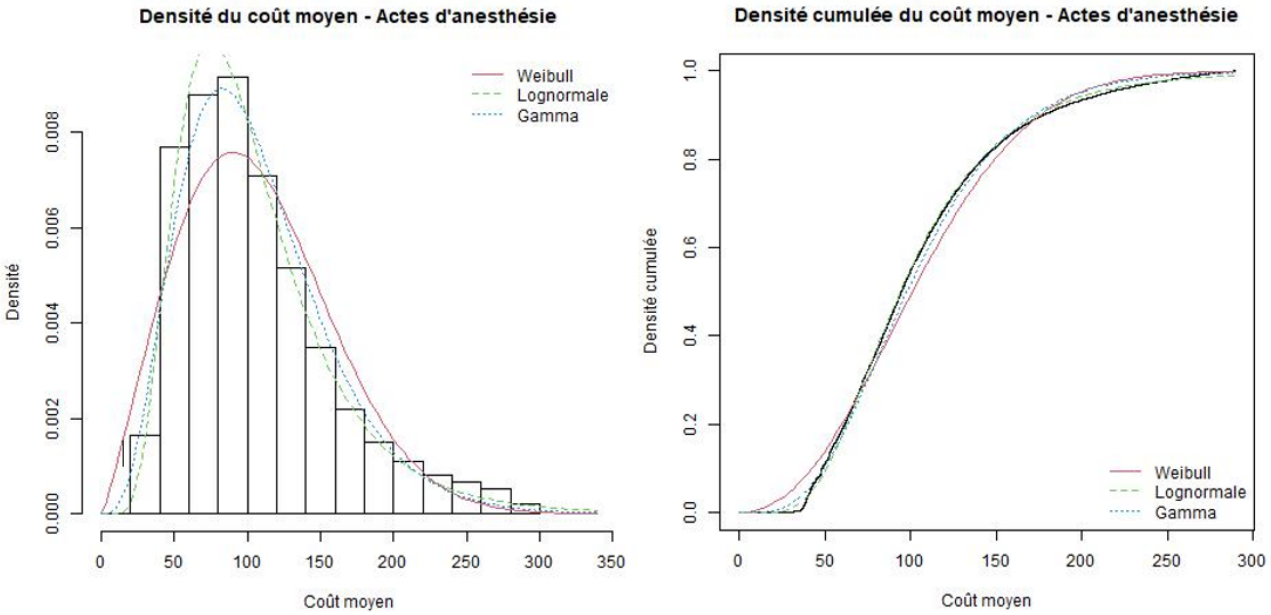


FIGURE 3.17 : Graphiques de densité de la dépense moyenne pour le libellé « Actes d'anesthésie ».

Il est notamment important de rappeler que la loi Log normale n'appartient pas à la famille exponentielle. Nous pouvons tout de même réaliser la modélisation avec cette loi en utilisant la loi normale et en prenant le logarithme de la variable d'intérêt. La fonction de lien utilisée dans ce cas est donc la fonction d'identité telle que : $g(\mu_i) = \mu_i = \mathbb{E}(ln(Y_i))$ Où $ln(Y_i)$ est une loi normale de paramètres μ_i et σ_i. Nous obtenons donc l'égalité suivante pour la modélisation de la dépense moyenne : $\mathbb{E}(Y_i) = e^{\beta_0+\beta_1 Age_i+\cdots+\frac{{\sigma_i}^2}{2}}$, où Y_i représente la variable d'intérêt DEPENSE_MOY.

L'objectif de cette tarification est de modéliser le plus précisément possible la dépense moyenne, sur une base d'apprentissage représentant 80% de la base d'étude filtrée sur le libellé brochure « Actes d'anesthésie ». Dans un premier temps, nous devons déterminer les combinaisons possibles des variables explicatives quantitatives qui rendraient le modèle GLM plus robuste. Pour cela, la seule variable explicative quantitative utilisée pour le modèle, AGE_BENEF, sera élevée à la puissance i pour i allant de 1 à 6 maximum. La méthode que nous utiliserons pour déterminer cette combinaison est la méthode forward. Cette méthode repose sur le principe d'ajout de variables explicatives jusqu'à obtenir un modèle robuste. L'arrêt de l'algorithme apparaît généralement lorsque l'ajout d'une variable explicative n'améliore plus significativement le critère AIC. D'autres méthodes existent comme la méthode backward, qui consiste à sélectionner l'ensemble des variables explicatives puis de les retirer une à une pour chaque modèle), ou la méthode forward-backward, qui ajoute et élimine des variables explicatives jusqu'à obtenir un modèle robuste. Ces méthodes ne seront pas utilisées dans le cadre de ce mémoire, respectivement par choix, mais aussi pour le temps d'exécution plus long. Les bases de données sur lesquelles le GLM est appliqué sont effectivement volumineuses (entre 100 000 et 2 millions de lignes). Cependant, ces méthodes pourront être implémentées dans un cadre d'amélioration du processus de tarification santé.

Les combinaisons et variables explicatives présentes pour le modèle n°i sont : AGE_BENEF^1, AGE_BENEF^2, ..., AGE_BENEF^i, REGION_BENEF, SOI_ANN. Pour la modélisation de la dépense moyenne avec le libellé brochure « Actes d'anesthésie », les valeurs des AIC obtenus pour chacun des modèles exécutés sont présentés dans le tableau 3.6.

TABLE 3.6 : Evaluation de la robustesse pour chaque modèle testé sur la variable d'intérêt DEPENSE_MOY.

Modèle	AIC	Déviance
Modèle n°6	77535.69	12832.84
Modèle n°3	77551.10	12837.48
Modèle n°4	77553.03	12837.46
Modèle n°5	77554.64	12837.38
Modèle n°2	77563.91	12840.69
Modèle n°1	78239.80	12988.55

Nous modéliserons donc la dépense moyenne pour les actes d'anesthésie selon le modèle n° 6. L'égalité évaluant la dépense moyenne par la modélisation GLM devient

$$\mathbb{E}(Y_i) = e^{\beta_0+\beta_1 Age_i+\beta_2 {Age_i}^2+\beta_3 {Age_i}^3+...+\beta_6 {Age_i}^6+\beta_7 1_{Region\ =\ 24}+...+\frac{{\sigma_i}^2}{2}},$$

où Y_i représente toujours la variable d'intérêt DEPENSE_MOY.

Analyse de la significativité des variables

Un premier GLM est exécuté avec l'ensemble des variables explicatives et combinaisons de la variable AGE_BENEF. Certaines variables pourront être éliminées ou retraitées par la suite selon leur significativité sur la variable d'intérêt DEPENSE_MOY. En effet, plusieurs informations sur le modèle exécuté sont obtenues comme le critère AIC, les valeurs prédites, mais aussi les coefficients GLM et les tests de significativité de chacune des variables explicatives renseignées au sein du GLM. Elles sont ensuite analysées afin d'identifier les retraitements ou suppression nécessaires.

Une variable est désignée comme significative lorsque la p-value qui lui est associée est inférieure à 0.1. Pour le test de significativité d'une variable explicative, l'hypothèse nulle s'écrit : $H_0 : \beta_i = 0$. Cette hypothèse signifie que le coefficient associé à la variable explicative en question est nul et donc cette variable n'a aucun impact sur la modélisation de la variable d'intérêt. L'hypothèse nulle est donc rejetée lorsque la p-value est généralement inférieure à 0.05. Dans le cadre des GLM implémentés sur *R*, la significativité est observée lorsque la p-value est inférieur à 0.1. Nous retiendrons cependant le seuil habituellement appliqué sur la p-value pour le rejet de l'hypothèse nulle. D'après les éléments d'informations donnés et les résultats présentés dans le tableau 3.7, nous remarquons que certaines variables explicatives ne sont pas significatives. Ce sont intégralement des variables concernant la région de résidence du bénéficiaire.

TABLE 3.7 : Coefficients GLM extraits du premier modèle réalisé pour la modélisation de la dépense moyenne.

Variables	Coefficient GLM estimé	P-value	Significativité de la variable
(Intercept)	5.27E+00	1.27E-212	***
I(AGE_BENEF)	-1.60E-01	3.60E-06	***
I(AGE_BENEF^2)	1.12E-02	3.88E-06	***
I(AGE_BENEF^3)	-3.69E-04	3.90E-06	***
I(AGE_BENEF^4)	6.22E-06	3.95E-06	***
I(AGE_BENEF^5)	-5.18E-08	4.28E-06	***
I(AGE_BENEF^6)	1.68E-10	4.74E-06	***
REGION_BENEF5	-1.03E-01	5.25E-18	***
REGION_BENEF11	-5.71E-02	4.30E-11	***
REGION_BENEF24	-2.20E-02	2.58E-02	*
REGION_BENEF27	-9.55E-03	2.98E-01	
REGION_BENEF28	-5.82E-02	3.82E-09	***
REGION_BENEF32	-5.80E-02	1.51E-10	***
REGION_BENEF52	-1.19E-01	9.48E-34	***
REGION_BENEF53	-1.34E-01	5.10E-40	***
REGION_BENEF75	-1.86E-02	3.52E-02	*
REGION_BENEF76	-7.04E-02	1.76E-16	***
REGION_BENEF84	3.92E-02	5.91E-06	***
REGION_BENEF93	-5.45E-02	3.06E-10	***
SOI_ANN2019	8.38E-02	1.54E-102	***

Des regroupements de variables sont donc réalisés selon deux hypothèses prises en compte :

- les coefficients GLM obtenus pour les variables regroupées sont proches ;
- les régions regroupées sont géographiquement proches, ou possèdent des caractéristiques communes (littoral, frontière terrestre, capitale, etc.).

Les modalités disponibles pour la variable REGION_BENEF sont présentées dans le tableau 3.8.

TABLE 3.8 : Modalités de la variable REGION_BENEF.

REGION_BENEF	Nom de la région associée
5	Régions et Départements d'outre-mer
11	Ile-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France - Nord-Pas-de-Calais-Picardie
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Aquitaine-Limousin-Poitou-Charentes
76	Languedoc-Roussillon-Midi-Pyrénées
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur et Corse

D'après le nouveau découpage géographique des régions présentés sur la carte 3.18, les coefficients GLM obtenus dans le tableau 3.7 et les hypothèses exposées précédemment, nous décidons de regrouper les régions suivantes ensemble :

- **Premier regroupement (Nord-Est) :** REGION_BENEF27 et REGION_BENEF44. La nouvelle variable issue de ce regroupement se nommera REGION_BENEF2744 ;
- **Deuxième regroupement (Sud/Sud-Ouest) :** REGION_BENEF75, REGION_BENEF76, REGION_BENEF93. La nouvelle variable issue de ce regroupement se nommera REGION_BENEF757693 ;
- **Troisième regroupement (Nord-Ouest) :** REGION_BENEF24, REGION_BENEF28, REGION_BENEF32, REGION_BENEF52, REGION_BENEF53. La nouvelle variable issue de ce regroupement se nommera REGION_BENEF2428325253.

FIGURE 3.18 : Carte des régions de France.

Une fois ces regroupements réalisés, le modèle GLM est exécuté une seconde fois. Les nouveaux résultats obtenus pour les coefficients GLM et tests de significativité sont satisfaisants. L'ensemble des variables explicatives sont significatives. (c.f. le tableau 3.9).

TABLE 3.9 : Coefficients GLM extraits du modèle final réalisé pour la modélisation de la dépense moyenne.

Variables	Coefficient GLM estimé	P-value	Significativité de la variable
(Intercept)	5.19E+00	8.81E-206	***
I(AGE_BENEF)	-1.58E-01	4.65E-06	***
I(AGE_BENEF^2)	1.11E-02	5.07E-06	***
I(AGE_BENEF^3)	-3.64E-04	5.18E-06	***
I(AGE_BENEF^4)	6.15E-06	5.28E-06	***
I(AGE_BENEF^5)	-5.12E-08	5.73E-06	***
I(AGE_BENEF^6)	1.66E-10	6.33E-06	***
REGION_BENEF11	1.91E-02	7.17E-03	**
REGION_BENEF2744	7.21E-02	7.35E-37	***
REGION_BENEF5	-2.67E-02	1.36E-02	*
REGION_BENEF757693	2.73E-02	3.86E-08	***
REGION_BENEF84	1.15E-01	2.10E-59	***
SOI_ANN2019	8.34E-02	3.64E-101	***

Les régions contenues dans une variable regroupée sont donc associées au même coefficient GLM puisqu'elles sont exposées au même risque. Afin de valider ce modèle GLM, nous devons analyser les résidus.

Analyse des résidus - Validation du modèle

Un résidu brut est défini par

$$\hat{\epsilon}_i = Y_i - \hat{\mu}_i,$$

avec Y_i les valeurs observées et $\hat{\mu}_i$ les valeurs prédites. Il définit l'écart entre la valeur observée et la valeur prédite par le modèle pour la variable d'intérêt. Pour la modélisation de la dépense moyenne, les résidus étudiés sont les **résidus de déviance** et les **résidus de Pearson**. Les résidus de déviance sont définis par

$$r_i^D = sign(\hat{\epsilon}_i) \times \sqrt{d_i},$$

avec d_i la contribution de l'observation i à la déviance telle que : $D = \sum_{i=1}^{n} d_i$. La déviance sera définie à la section 3.2.4, tandis que les résidus de Pearson sont définis par

$$r_i^P = \frac{\hat{\epsilon}_i}{\sqrt{Var(\hat{\mu}_i)}}.$$

Les quatre hypothèses qui doivent être vérifiées pour les résidus d'un Modèle Linéaire Généralisé sont :

- **Hypothèse 1 :** Les résidus sont distribués selon une loi normale de moyenne nulle ;
- **Hypothèse 2 :** Les résidus sont distribués avec une variance constante ;
- **Hypothèse 3 :** Linéarité respectée.

Ces hypothèses vont être vérifiées graphiquement. Seuls les résultats pour les résidus de Pearson seront présentés. Cette analyse par observation des résidus permet notamment de détecter des valeurs aberrantes, c'est-à-dire des points extrêmes, qui pourraient détériorer la qualité du modèle. Vérifions, grâce aux graphiques des résidus de Pearson, si les hypothèses précédentes sont vérifiées pour le modèle effectué.

Le graphique 3.19, nommé le Q-Q plot, permet d'évaluer l'hypothèse 1, c'est-à-dire, la normalité des résidus. Si les résidus se situent le long de la droite pointillée alors l'hypothèse de normalité est vérifiée et acceptée. A l'inverse, s'ils s'en écartent, alors nous rejetons l'hypothèse de normalité. Dans le cas de notre étude et de ce modèle, les points sont majoritairement proches de la droite. L'hypothèse 1 est donc bien vérifiée. Cependant, trois points sont identifiés comme valeurs aberrantes. Les supprimer engendrerait alors une amélioration de la qualité du modèle.

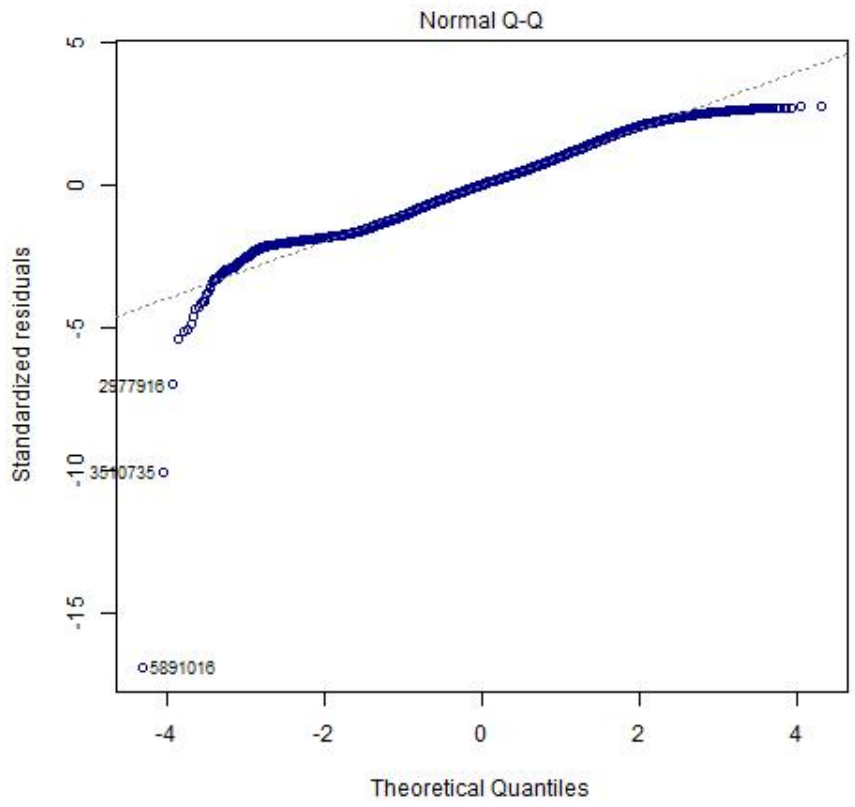


FIGURE 3.19 : Graphique de vérification de l'hypothèse de normalité des résidus de Pearson.

Le graphique 3.20 permet de vérifier l'hypothèse 2, c'est-à-dire l'homogénéité des résidus. Ce graphique représente les valeurs prédites de la variable d'intérêt par le modèle en fonction de la racine carrée des résidus standardisés. Si les points visibles sur le graphique sont répartis de façon homogène sur le graphique, et qu'aucune tendance n'apparait clairement, alors l'hypothèse est acceptée. La courbe violette, présente sur le graphique, est alors généralement plate dans le cadre d'acceptation de l'hypothèse. Dans le cadre de notre étude et de ce modèle, aucune tendance n'apparait. Les points sont plutôt homogènes. L'hypothèse 2 est donc bien vérifiée. De plus, les valeurs aberrantes détectées précédemment sont de nouveaux présentes. Nous remarquons qu'elles ne reflètent pas le même comportement que les autres points, et sont très éloignés du groupe homogène de points.

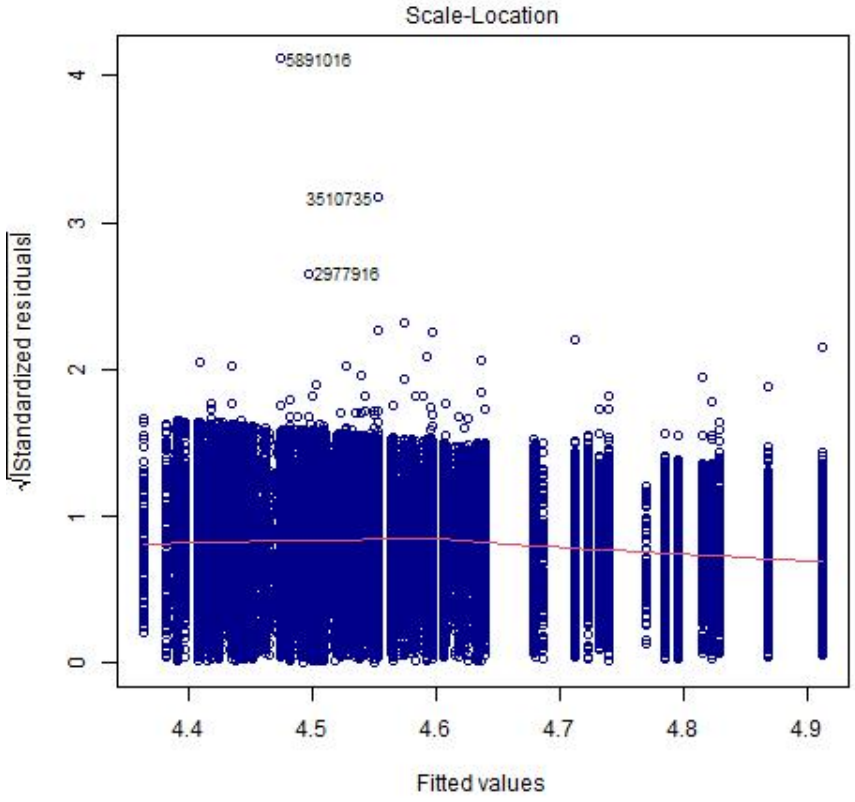


FIGURE 3.20 : Graphique de vérification de l'hypothèse d'homogénéité des résidus de Pearson.

Enfin, le graphique 3.21 permet la vérification de l'hypothèse 3 qui est l'hypothèse de linéarité. Le graphique représente les valeurs prédites de la variable d'intérêt par le modèle en fonction des résidus, non standardisés cette fois-ci. L'hypothèse est vérifiée si les points du graphique sont uniformément répartis autour de 0. La courbe violette doit donc être significativement plate et autour de l'axe du 0. Dans le cadre de notre étude, les résidus sont effectivement bien centrés autour de 0. L'hypothèse 3 est donc bien vérifiée. Encore une fois, ces valeurs aberrantes sont présentes. Elles seront supprimées afin d'obtenir un modèle performant. De plus, le graphique 3.22 de vérification de l'hypothèse 3 avec les résidus de Déviance fournit un résultat identique. Les valeurs aberrantes ne sont cependant pas mises en avant. C'est pour cette raison que les analyses sont principalement concentrées autour des résidus de Pearson.

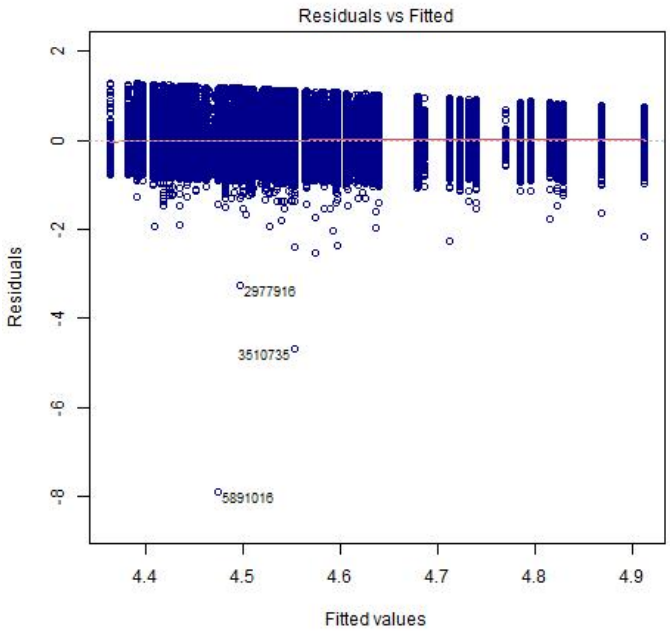


FIGURE 3.21 : Graphique de vérification de l'hypothèse de linéarité des résidus de Pearson.

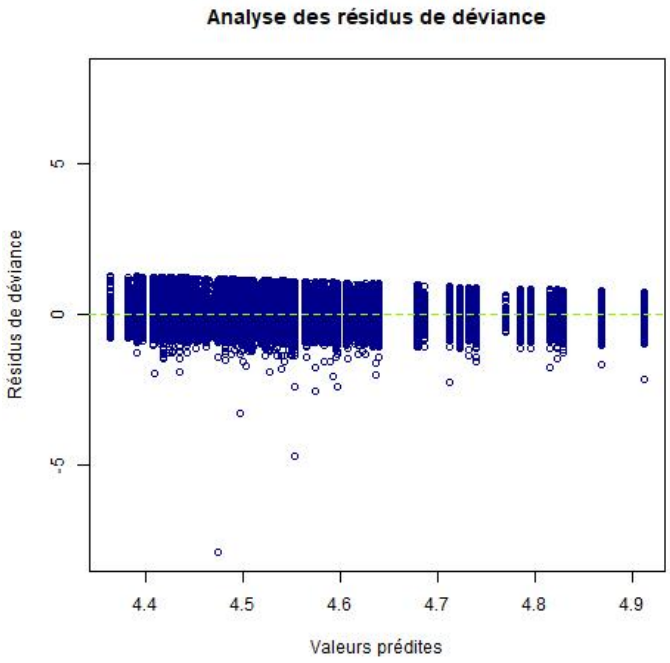


FIGURE 3.22 : Graphique de vérification de l'hypothèse de linéarité des résidus de Déviance.

Toutes les hypothèses sont donc acceptées avec le modèle GLM implémenté pour les actes d'anesthésie. La validation du modèle est donc effectuée. Vérifions à présent la performance de ce modèle.

Performance du modèle

La performance du modèle, c'est-à-dire la qualité de l'ajustement, s'évalue en fonction de la déviance. La déviance est définie par

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

Si la déviance est positive et très petite, alors le modèle est de bonne qualité. De plus, pour mesurer l'amélioration de la qualité de l'ajustement du modèle lors de l'ajout de variables explicatives, nous devons analyser la différence obtenue entre la déviance nulle et la déviance résiduelle. La déviance nulle représente la déviance obtenue dans le cadre du modèle sans variable. Ce modèle évalue un seul paramètre puisqu'il est exécuté sans aucune variables explicatives. La déviance résiduelle est la déviance obtenue pour le modèle implémenté précédemment. La déviance nulle sera notée D_0, et la déviance résiduelle D_X. L'amélioration de la qualité de l'ajustement est donc évaluée par le taux, défini tel que

$$\tau = \frac{D_X - D_0}{D_0}.$$

Les résultats obtenus pour ce modèle, présentés au sein du tableau 3.10, sont alors satisfaisants.

TABLE 3.10 : Résultats de la déviance résiduelle du modèle GLM pour la dépense moyenne.

Déviance du modèle GLM	Taux calculé
12871.46	-5.46%

Le taux obtenu est de - 5.46%. En termes de déviance, cela signifie le modèle implémenté est amélioré de 5% par rapport au modèle sans variable. Cette amélioration est donc non négligeable. Le modèle étant validé, nous pouvons à présent interpréter les coefficients obtenus (c.f. tableau 3.9).

Interprétation des coefficients GLM obtenus

Le modèle GLM réalisé étant robuste et vérifiant les hypothèses sur les résidus, nous pouvons donc interpréter les coefficients GLM obtenus et fournis dans le tableau 3.9. Pour rappel, dans le cas de la loi Log normale, un modèle linéaire sur le logarithme de la variable d'intérêt est effectué.

Tout d'abord, pour un modèle linéaire, le signe du coefficient obtenu pour une variable explicative donnée, notée X, indique l'effet de cette variable sur la variable d'intérêt Y. En effet, toutes choses égales par ailleurs, un coefficient négatif pour une variable explicative quantitative signifie que l'augmentation de la valeur prise par X d'une unité fera diminuer Y de X unités. Inversement, toutes choses égales par ailleurs, pour un coefficient positif, augmenter la valeur prise par X d'une unité entraînera l'augmentation de X unités la valeur de Y. Cependant, le signe du coefficient ne donne pas d'indication sur l'intensité de l'effet de la variable explicative sur la variable d'intérêt. Cette intensité dépend de la fonction de lien choisie, et donc de la formule définissant $\mathbb{E}(Y)$ avec Y la variable d'intérêt. Dans notre cas, le modèle linéaire utilisé est $\ln(Y) \sim X$. Toutes choses égales par ailleurs, augmenter la valeur de X d'une unité entraînera l'augmentation ou la diminution de $X\%$ de la valeur de Y.

Par exemple, pour un assuré homme de 40 ans, résidant en Ile-de-France, le montant de la dépense moyenne estimée en 2018 s'élève à

$$\mathbb{E}(Y) = e^{5.19 - 0.16 \times 40 + 0.01 \times 40^2 + ... + 0.02} = 84.20 \; euros.$$

Cependant, il s'agit du montant de la dépense réelle du soin et non du remboursement effectué par l'organisme complémentaire. Nous devons appliquer la formule de passage détaillé dans la section suivante et en annexe B.3.

Passage des frais réels au remboursement complémentaire

Pour rappel, le montant de la prise en charge de la prestation santé est rarement égal au montant de la dépense, sauf dans le cadre de la CSS ou de personnes en état d'ALD. Le calcul du montant remboursé par un organisme complémentaire santé s'effectue donc en fonction de la base et du taux de remboursement, ainsi que d'éventuels planchers ou plafonds imposés. La modélisation précédente a été effectuée sur la dépense moyenne. Pour retrouver le montant remboursé, le calcul suivant est effectué

$$Y_{RC} = \min(\max(Y - \tau_{RSS} \times BRSS, 0), (\tau_{RC} - \tau_{RSS}) \times BRSS),$$

où :

- Y désigne la dépense moyenne d'un soin ;
- $BRSS$ désigne la base de remboursement de la Sécurité Sociale ;
- τ_{RSS} désigne le taux de remboursement défini par la Sécurité Sociale pour un libellé donné ;
- τ_{RC} désigne le taux de remboursement de la complémentaire santé (y compris le remboursement de la Sécurité Sociale).

Dans le cadre des remboursements de la Sécurité Sociale, le taux de remboursement des actes d'anesthésie s'élève majoritairement à 80%. Quant aux consultations d'anesthésiste, elles sont remboursées à hauteur de 70% de la base de remboursement. Pour un contrat complémentaire santé, le taux de remboursement est défini selon les garanties souscrites par l'adhérent. La base de remboursement des actes d'anesthésie qui sera retenue pour les calculs suivants est définie comme la base de remboursement moyenne pour cet acte sur la base de donnée utilisée pour la tarification. Après calcul, la base de remboursement s'élève à 92.63 €.

Le calcul du remboursement complémentaire pour les actes d'anesthésie sera donc effectué lors de la tarification d'une garantie d'un contrat complémentaire santé. L'équation (3.8) donne le montant de la prise en charge des actes d'anesthésie pour un contrat complémentaire santé garantissant un remboursement à hauteur de τ_{RC}% de la base de remboursement (remboursement y compris celui de la sécurité sociale)

$$\mathbb{E}(Y_{RC}) = \mathbb{E}(\min(\max(Y - \tau_{RSS} \times BRSS, 0), (\tau_{RC} - \tau_{RSS}) \times BRSS)). \tag{3.8}$$

Après avoir détaillé le calcul (c.f. B.3), la prise en charge est calculée grâce à l'équation suivante

$$\begin{aligned}\mathbb{E}(Y_{RC}) = e^{\mu+\frac{\sigma^2}{2}}(\phi(\frac{\ln(R_{Mut}) - (\mu+\sigma^2)}{\sigma}) - \phi(\frac{\ln(R_{SS}) - (\mu+\sigma^2)}{\sigma}))\\ - R_{SS} \times (F_{\mu,\sigma}(R_{Mut}) - F_{\mu,\sigma}(R_{SS})) + (R_{Mut} - R_{SS}) \times (1 - F_{\mu,\sigma}(R_{Mut})),\end{aligned}$$

avec :

- R_{Mut} désigne le remboursement complémentaire (hors remboursement de la Sécurité Sociale) ;
- R_{SS} désigne le remboursement de la Sécurité Sociale ;
- $F_{\mu,\sigma}$ désigne la fonction de répartition de la loi Log normale de paramètres μ et σ ;
- ϕ désigne la fonction de répartition de la loi normale centrée réduite.

En effet, comme expliqué précédemment, l'objectif de ce mémoire est d'intégrer une tarification nationale à la tarification santé actuelle d'Actélior. Pour cela, la démarche du processus entre les deux tarifications doit donc être similaire afin de pouvoir effectuer des comparaisons. Les travaux effectués dans le cadre du mémoire de Chery (2015), créateur de l'outil de tarification santé d'Actélior, ont donc été repris. Le calcul précédent est extrait de ce mémoire. Les calculs intermédiaires seront exposés de nouveau en annexe B.3.

Etude de la modélisation de la fréquence

La seconde étape de la méthode « Coût moyen × Fréquence » est la modélisation de la fréquence. Les étapes de cette modélisation sont similaires à celle de la dépense moyenne. Cependant, pour cette modélisation, nous n'utiliserons pas la variable FREQUENCE mais la variable QUANTITE_ACTES en ajoutant la variable de l'exposition des bénéficiaires EXPO_BENEF en offset. En effet, dans le cadre d'un GLM, lorsque la variable d'intérêt dépend linéairement d'une autre variable, nous devons ajouter cette dernière en tant que variable offset, permettant alors de "tarer" le modèle. Les variables utilisées pour la modélisation de la quantité d'actes sont donc les suivantes :

- **Variable d'intérêt :** QUANTITE_ACTES (Nombre d'actes réalisés) ;
- **Variables explicatives quantitatives :** AGE_BENEF (Age du bénéficiaire) ;
- **Variables explicatives catégorielles :** REGION_BENEF (Région de la résidence du bénéficiaire) et SOI_ANN (Année de survenance du soin) ;
- **Variable offset :** ln(EXPO_BENEF) (Logarithme de l'exposition des bénéficiaires).

Les résultats de cette modélisation sont présentés ci-après.

Présentation des choix effectués pour les actes d'anesthésie

La loi choisie, qui ajuste au mieux la quantité des « Actes anesthésie », est la loi **Binomiale Négative**. En effet, après application de la fonction goodfit et l'analyse du RMSE pour ce libellé brochure, la loi Binomiale négative est la loi s'ajustant le mieux à la variable d'intérêt. Cette décision se vérifie sur le graphique 3.23.

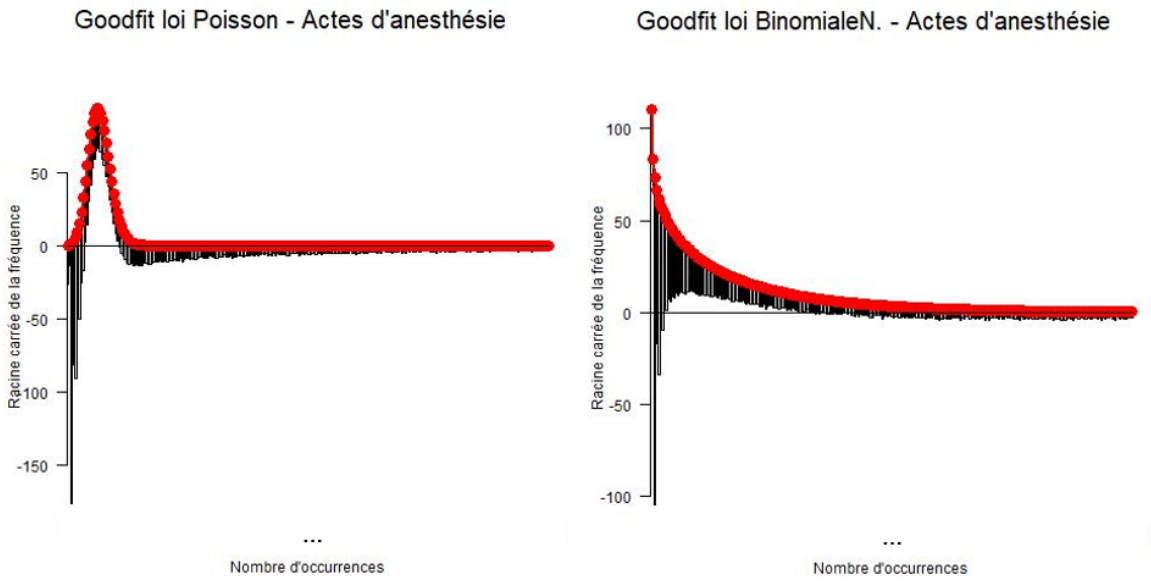


FIGURE 3.23 : Graphiques Goodfit de la quantité d'actes pour le libellé « Actes d'anesthésie ».

Nous pouvons remarquer que la loi Binomiale négative est effectivement plus adaptée aux observations. En effet, les écarts entre l'axe des abscisses et l'extrémité des bâtons sont moins importants sur le graphique goodfit de cette loi par rapport à ceux obtenus pour la loi de Poisson. Le graphique 3.24 de la répartition de la quantité des actes pour le libellé « Actes d'anesthésie » permet de confirmer cette hypothèse. Nous pouvons observer une forte décroissance du nombre d'occurrences pour des fréquences faibles d'actes d'anesthésie. Cette décroissance est moins marquée pour des fréquences plus élevées. Cela traduit donc un phénomène de sur-dispersion. Effectivement, la moyenne de l'échantillon est de 15.95 et sa variance est de 1162.26. Dans le cas de sur-dispersion, la loi de Poisson n'est pas optimale.

Cependant, il est tout de même important de noter que la loi Binomiale négative ne s'ajuste pas parfaitement aux données. D'autres lois, que nous n'étudions pas dans le cadre de la modélisation de la fréquence au sein de ce mémoire, pourraient être utilisées. Pour la suite des traitements, nous utiliserons la loi Binomiale négative, en gardant tout de même à l'esprit qu'il ne s'agit pas de la loi la plus adaptée, pour l'analyse des futurs résultats.

La fonction de lien utilisée pour la loi Binomiale négative est donnée dans le tableau 3.1. Nous obtenons donc l'égalité suivante pour la modélisation de la quantité d'actes

$$g\left(\mu_i\right) = \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) = \mathbb{E}(Y_i),$$

où α est le paramètre d'une loi Gamma et Y_i représente la variable d'intérêt QUANTITE_ACTES. Pour déterminer le paramètre alpha, nous utilisons la méthode des moments. En effet, pour une variable aléatoire X suivant une loi Gamma de paramètres α et β, l'espérance et la variance associée sont définies par

$$\mathbb{E}(X) = \alpha \times \beta,$$
$$\mathbb{V}(X) = \alpha \times \beta^2.$$

La valeur du paramètre α s'obtient donc par la formule suivante : $\alpha = \frac{\mathbb{E}(X)^2}{\mathbb{V}(X)}$. L'espérance et la variance

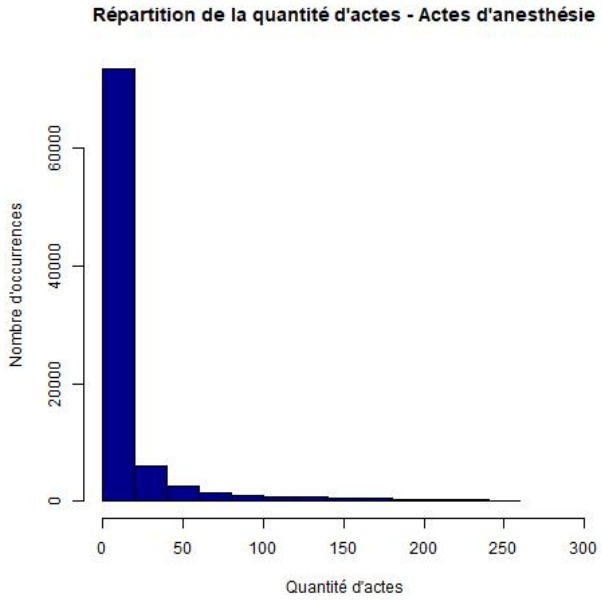


FIGURE 3.24 : Histogramme de la quantité d'actes pour le libellé « Actes d'anesthésie ».

empiriques ont déjà été donc déterminées précédemment. La valeur de α vaut alors

$$\alpha = \frac{(15.94976)^2}{1162.259} = 0.2189.$$

Nous obtenons donc l'égalité(3.9) pour la modélisation de la fréquence

$$\frac{\mu_i}{expo_i} = \frac{\mathbb{E}(Y_i)}{expo_i} = \frac{1}{\alpha \times expo_i} \times \frac{e^{\beta_0+\beta_1 Age_i+\beta_2 {Age_i}^2+\beta_3 {Age_i}^3+...+\beta_6 {Age_i}^6+...+\beta_{20}1_{REGION\_BENEF\ =\ 1132}}}{1-e^{\beta_0+\beta_1 Age_i+\beta_2 {Age_i}^2+\beta_3 {Age_i}^3+...+\beta_6 {Age_i}^6+...+\beta_{20}1_{REGION\_BENEF\ =\ 1132}}}, \tag{3.9}$$

où Y_i représente la variable d'intérêt QUANTITE_ACTES et $expo_i$ représente la durée d'exposition pour un bénéficiaire i.

A présent, nous réalisons le même processus que celui effectué pour la modélisation de la dépense moyenne : la détermination de la combinaison des variables explicatives qui rend le modèle plus robuste. Pour rappel, les combinaisons et variables explicatives présentes pour le modèle sont : AGE_BENEF^1, AGE_BENEF^2, ..., AGE_BENEF^i, $REGION\_BENEF$, SOI_ANN. Pour la modélisation de la quantité d'actes avec le libellé brochure « Actes d'anesthésie », les valeurs du critère AIC obtenus pour chacun des modèles exécutés sont présentés dans le tableau 3.11. Nous modéliserons donc la quantité des actes d'anesthésie selon le modèle n°6.

TABLE 3.11 : Evaluation de la robustesse pour chaque modèle testé sur la variable d'intérêt QUANTITE_ACTES.

Modèle	AIC	Déviance
Modèle n°6	434722.07	68000.85
Modèle n°5	434725.25	68001.70
Modèle n°4	434732.52	68003.18
Modèle n°3	435497.87	68131.99
Modèle n°2	435724.19	68170.21
Modèle n°1	435819.28	68189.84

Analyse de la significativité des variables

Un premier GLM est exécuté avec l'ensemble des variables explicatives et combinaisons de la variable AGE_BENEF. Certaines variables pourront être éliminées ou regroupées par la suite selon leur significativité sur la variable d'intérêt QUANTITE_ACTES. Pour identifier ces éventuels regroupements, nous analysons, dans le tableau 3.12, les informations extraites du GLM.

TABLE 3.12 : Coefficients GLM extraits du premier modèle réalisé pour la modélisation de la fréquence.

Variables	Coefficient GLM estimé	P-value	Significativité de la variable
(Intercept)	-1.29E+01	5.99E-159	***
I(AGE_BENEF)	3.88E-01	7.40E-05	***
I(AGE_BENEF^2)	-1.98E-02	4.02E-03	**
I(AGE_BENEF^3)	5.30E-04	1.93E-02	*
I(AGE_BENEF^4)	-8.19E-06	3.26E-02	*
I(AGE_BENEF^5)	6.88E-08	3.16E-02	*
I(AGE_BENEF^6)	-2.38E-10	2.30E-02	*
REGION_BENEF5	1.03E+00	5.96E-204	***
REGION_BENEF11	-7.11E-01	1.38E-183	***
REGION_BENEF24	6.77E-01	1.47E-126	***
REGION_BENEF27	5.30E-01	5.77E-92	***
REGION_BENEF28	4.71E-01	5.65E-63	***
REGION_BENEF32	-5.47E-02	3.44E-02	*
REGION_BENEF52	4.25E-01	1.47E-51	***
REGION_BENEF53	3.39E-01	1.72E-31	***
REGION_BENEF75	-1.68E-01	2.46E-11	***
REGION_BENEF76	-1.27E-01	2.47E-07	***
REGION_BENEF84	-4.54E-01	4.99E-75	***
REGION_BENEF93	-7.85E-02	1.55E-03	**
SOI_ANN2019	1.27E-01	2.43E-30	***

La déviance résiduelle et le critère AIC obtenus pour ce modèle ont pour valeur respectivement 68 001 et 434 722 (c.f. tableau 3.14). De plus, l'ensemble des variables explicatives sont significatives. Pour rappel, dans le cadre des GLM implémentés sur *R*, la significativité est observée lorsque la p-value est inférieur à 0.1. Nous retiendrons cependant le seuil de 0.05, habituellement appliqué pour le rejet

de l'hypothèse nulle. Cependant, deux variables pourraient être impactées par un regroupement : REGION_BENEF11, REGION_BENEF93. Elles correspondent respectivement aux régions Ile-de-France et Provence-Alpes-Côte-d'Azur-et-Corse. Comme pour la dépense moyenne, les regroupements sont effectués entre variables de coefficients GLM et de localisation géographique proches. En s'appuyant à nouveau sur la carte des régions de France (c.f. graphique 3.18) et sur les différentes modalités de la variable région (c.f. tableau 3.8), nous effectuons deux regroupements, afin de voir si la qualité du modèle est améliorée. :

- **Premier regroupement (Nord) :** REGION_BENEF11 et REGION_BENEF32. La nouvelle variable issue de ce regroupement se nommera REGION_BENEF1132 ;
- **Deuxième regroupement (Sud/Sud-Ouest) :** REGION_BENEF75, REGION_BENEF76, REGION_BENEF93. La nouvelle variable issue de ce regroupement se nommera REGION_BENEF757693.

Une fois ces regroupements réalisés, le modèle GLM est exécuté une seconde fois. Les nouveaux résultats obtenus et présentés au sein du tableau 3.13 pour les coefficients GLM et tests de significativité sont satisfaisants. L'ensemble des variables explicatives sont significatives.

TABLE 3.13 : Coefficients GLM extraits du second modèle réalisé pour la modélisation de la fréquence.

Variables	Coefficient GLM estimé	P-value	Significativité de la variable
(Intercept)	-1.29E+01	8.30E-160	***
I(AGE_BENEF)	3.81E-01	1.06E-04	***
I(AGE_BENEF^2)	-1.94E-02	4.98E-03	**
I(AGE_BENEF^3)	5.19E-04	2.26E-02	*
I(AGE_BENEF^4)	-8.01E-06	3.72E-02	*
I(AGE_BENEF^5)	6.74E-08	3.60E-02	*
I(AGE_BENEF^6)	-2.33E-10	2.64E-02	*
REGION_BENEF1132	-2.40E-01	4.29E-47	***
REGION_BENEF24	7.99E-01	2.30E-228	***
REGION_BENEF27	6.53E-01	2.11E-190	***
REGION_BENEF28	5.94E-01	1.20E-128	***
REGION_BENEF44	1.23E-01	1.23E-09	***
REGION_BENEF5	1.15E+00	2.99E-303	***
REGION_BENEF52	5.48E-01	6.39E-110	***
REGION_BENEF53	4.62E-01	1.31E-72	***
REGION_BENEF84	-3.31E-01	9.36E-58	***
SOI_ANN2019	1.28E-01	2.37E-30	***

Cependant, la valeur du critère AIC et de la déviance résiduelle sont plus élevées (c.f. tableau 3.14). Les regroupements n'améliore donc pas la qualité du modèle. Le premier modèle GLM est donc validé.

Afin de valider ce modèle GLM, nous devons analyser les résidus.

TABLE 3.14 : Déviance et critère AIC obtenus pour les deux modélisations GLM de la fréquence.

Modèle	Déviance	AIC
Première exécution	68000.85	434722
Seconde execution	68107.17	435361

Analyse des résidus - Validation du modèle

Une fois le modèle obtenu, nous vérifions les hypothèses que les résidus doivent respectés. Pour cela, deux types de résidus sont analysés pour la modélisation de la fréquence :

- les Crunched résidus (ou crunched residuals) (LAMON, 2019) ;
- les résidus d'Anscombe (NDIBI OTAKANA, 2017).

Pour rappel, nous avons précedemment défini les termes suivants :

- μ_i désigne la valeur prédite pour une observation i ;
- Y_i désigne la valeur observée pour la variable d'intérêt de la quantité d'actes pour une observation i.

Premièrement, donnons les définitions de ces deux types de résidus. Les crunched résidus sont construits par la création de 500, 2500 ou 10 000 groupes d'observations. Ces groupes sont formés en triant dans un premier temps les valeurs ajustées par ordre croissant. Dans notre cas, 416 groupes ont été formés. Aucune fonction de calcul automatique des crunched résidus n'est disponible sur *R*. Ils sont donc déterminés manuellement par

$$r_i^C = \frac{\sum_i (Y_i - \hat{\mu}_i)}{\sqrt{\mathbb{V}(\sum_i \hat{\mu}_i)}}.$$

Après traitement, nous obtenons le graphique 3.25 des crunched résidus de notre modèle de fréquence. Les résultats obtenus sont satisfaisants. Les résidus sont effectivement bien centrés autour de la valeur zéro et sont répartis de façon homogène. Les hypothèses des résidus sont finalement vérifiées.

Quant aux résidus d'Anscombe, ce sont des résidus de déviance approximés par une transformation de la variable d'intérêt de manière à ce qu'elle se rapproche le plus possible d'une loi normale centrée réduite. Ces résidus sont définis par

$$r_i^A = \frac{h(Y_i) - h(\hat{\mu}_i)}{\sqrt{h(\hat{\mu}_i)}}.$$

Par exemple, si la variable d'intérêt Y suit une loi de Poisson, la fonction h transformant la loi de Poisson en une loi centrée réduite est donnée par : $h(a) = a^{2/3}$. Ce qui donne

$$r_i^A = \frac{3}{2} \frac{Y_i^{\frac{2}{3}} - \hat{\mu}_i^{\frac{2}{3}}}{\hat{\mu}_i^{\frac{2}{3}}}.$$

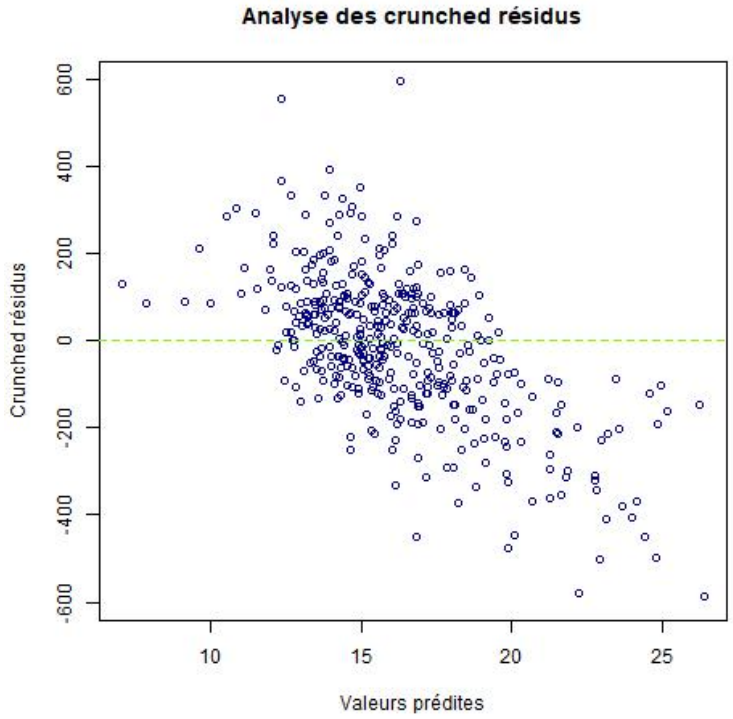


FIGURE 3.25 : Analyse des crunched résidus pour la première modélisation de la fréquence.

De plus, les résidus d'Anscombe permettent d'atténuer les limites des résidus standards et sont plus facilement interprétables et exprimables que les résidus de déviance. Cependant, nous ne les étudierons pas pour le libellé « Actes d'anesthésie » puisque nous avons choisie la loi Binomiale négative pour la modélisation. La transformation à appliquer est donc différente.

Performance du modèle

La performance du modèle, c'est-à-dire la qualité de l'ajustement, s'évalue en fonction de la déviance. La déviance a déjà été définie lors de l'analyse de la qualité de l'ajustement pour la modélisation de la dépense moyenne en section 3.2.4. Les résultats obtenus pour le modèle et présentés au sein du tableau 3.12 de la quantité d'actes d'anesthésie sont satisfaisants.

TABLE 3.15 : Résultats de la déviance résiduelle du modèle GLM pour la fréquence.

Déviance du modèle GLM	Taux calculé
68000.85	-13.87%

Le taux obtenu est de – 13.87%. En termes de déviance, cela signifie le modèle implémenté est amélioré de 14% par rapport au modèle sans variable. Cette amélioration est donc non négligeable. Le modèle étant validé, nous pouvons à présent déterminer la fréquence, puis la prime pure des actes d'anesthésie (c.f. tableau 3.12).

3.2.5 Tarification finale Damir

Calculons à présent la prime pure obtenue par les modélisations précédentes pour les actes d'anesthésie. Pour cela, nous avons choisi un assuré homme de 40 ans habitant en région d'Ile-de-France.

Obtention de la prime pure modélisée pour les actes d'anesthésie

Dans un premier temps, nous devons calculer le remboursement complémentaire moyen à partir des éléments donnés en section 3.2.4 et de la dépense moyenne estimée pour cet assuré (c.f. section 3.2.4). Pour cela, la base de remboursement de la Sécurité Sociale $BRSS$, le taux de remboursement de la Sécurité Scoiale $\tau_R SS$ et le taux de remboursement de la complémentaire santé pour les actes d'anesthésie doivent être définis.

Tout d'abord, la base de remboursement des actes d'anesthésie est déterminée en effectuant la moyenne sur les bases de remboursement par acte. En effet, initialement, la base de remboursement est un montant agrégé et correspond donc au montant de plusieurs soins regroupés. La base de remboursement moyenne obtenue pour un acte d'anesthésie est de 92.63 €.
De plus, comme expliqué en section 3.2.4, le taux de remboursement de la Sécurité Sociale pour les actes d'anesthésie est de 80%. Enfin, pour le remboursement complémentaire, nous choisirons une garantie remboursant à hauteur de 150% de la $BRSS$ les actes d'anesthésie.

Nous obtenons donc les valeurs des remboursements suivants :

- $R_{SS} = 80\% \times 92.63 = 74.10$ €,
- $R_{Mut} = 150\% \times 92.63 = 138.94$ €.

Pour appliquer la formule définie à la section 3.2.4, il nous faut également les paramètres de la loi Log Normale utilisées pour la modélisation, ainsi que la valeur prise par sa fonction de répartition. Ces paramètres sont :

- $\mu = \beta_0 + \beta_1 \times 40 + \beta_2 \times 40^2 + \cdots + \beta_{11} \times 1_{REGION\_BENEF11} = 4.4332$,
- $\sigma = 0.4663$,
- $F_{\mu,\sigma}(R_{SS}) = 0.3920$,
- $F_{\mu,\sigma}(R_{Mut}) = 0.8586$.

Nous pouvons à présent calculer le montant estimé du remboursement complémentaire moyen pour un assuré homme de 40 ans résidant en Ile-de-France. En utilisant l'équation **??**, le remboursement complémentaire annuel moyen pour les actes d'anesthésie est de

$$\mathbb{E}(Y_{RC}) = 21.42 \; euros.$$

Enfin, nous devons aussi calculer la fréquence des actes d'anesthésie puisque nous sommes dans le cadre de la méthode « Coût moyen × Fréquence ». Pour rappel, nous avons fait l'hypothèse que l'exposition de chaque personne dans la base Open Damir est égale à 1. Après avoir modélisé la quantité des actes, nous obtenons la fréquence par le biais de l'équation 3.9 et nous obtenons

$$E(Y_{Freq}) = 0.000109 \; fois.$$

La prime pure annuelle pour les actes d'anesthésie s'élève donc à

$$\mathbb{E}(RC) = \mathbb{E}(Y_{RC}) \times \mathbb{E}(Y_{Freq}) = 21.42 \times 0.000109 = 0.00233 \; euros.$$

Ce qui donne un montant mensuel de 0.00019 €. Ce montant est inférieur au montant obtenu par l'outil de tarification Actélior qui est de 0.001 €, mais reste cohérent. En effet, par avis d'expert, le coût des actes d'anesthésie est très peu élevé pour les acteurs de l'assurance. Le prochain chapitre, le chapitre 4, abordera la méthode utilisée pour obtenir une tarification liant celle d'Actélior et de la base Open Damir.

Application de coefficients d'ajustement

La tarification Damir n'intègre pas les notions de niveaux de garantie au sein de la modélisation. En effet, cette distinction n'est pas disponible pour l'Assurance Maladie Obligatoire. Les niveaux de garantie d'un contrat complémentaire santé indiquent le niveau de remboursement choisi. Par exemple, la prise en charge des dépenses de soin par l'organisme assureur, pour un produit « Haut de gamme », sera plus élevée que pour un produit « Bas de gamme ». Pour l'Assurance Maladie Obligatoire, les taux de remboursement sont identiques pour l'ensemble des assurés, sauf en cas d'adhésion à la CSS ou pour une personne en état d'ALD. Dans ce cas, les variables indiquant si la personne bénéficie des aides de la CSS ou de l'ALD pourraient servir de variables explicatives. Ces personnes n'étant pas intégrées dans le périmètre de l'étude, aucune différence de niveau n'est donc présente au sein de la base de données Open Damir traitée. Pour intégrer, cette notion de niveaux de garantie, un coefficient d'ajustement sera appliqué à la prime pure obtenue.

De plus, des frais de gestion ou de prestations devraient être pris en compte. Cependant, nous intégrerons ces frais dans un second temps. Les coefficients d'ajustement pour les niveaux de garanties et les coefficients d'ajustement ralatifs aux frais seront appliqués une fois le tarif de la prime pure obtenu.

Chapitre 4

Outils pour optimiser les études tarifaires des acteurs

4.1 Optimisation de la tarification santé par la théorie de la crédibilité.

Aujourd'hui, Actélior dispose d'un outil de tarification santé créé par le biais de bases de données clients. La digitalisation du monde de l'assurance amène le cabinet de conseil Actélior à réfléchir sur l'utilisation des bases de données nationales disponibles en Open Data, comme la base de données Open Damir. En complément de la compréhension de la base de données et de son utilisation pour le suivi technique, Actélior souhaite intégrer ces informations au sein de sa tarification. L'objectif est donc de joindre la tarification de l'outil d'Actélior avec la tarification de la base Open Damir. C'est dans ce cadre que la théorie de la crédibilité va être utilisée.

4.1.1 Principe de la théorie de la crédibilité

La théorie de crédibilité est un modèle généralement utilisé lorsque les effectifs du portefeuille d'assuré et la profondeur d'historique sont trop peu élevés pour calibrer une loi probabiliste. En assurance, cette théorie est fondée sur la distinction entre la probabilité mathématique d'un événement dans un cadre général (la sinistralité moyenne), et la probabilité de ce même événement dans un cas individuel (la sinistralité individuelle). A chacune de ces probabilités, des poids très différents sont attribués. Théoriquement, la prime individuelle P (Suru A., 2020) s'exprime selon l'équation (4.1) suivante

$$P = \alpha X + (1 - \alpha)\, C, \tag{4.1}$$

où X est l'expérience individuelle (reposant sur l'historique des sinistres propres à l'assuré) et C l'expérience collective (reposant sur l'historique de sinistres de tous les assurés du portefeuille), pondérées par un facteur de crédibilité α, compris entre 0 et 1, défini par la formule (4.2)

$$\alpha = \frac{n}{n + K}. \tag{4.2}$$

Le modèle de crédibilité associe la notion de mutualisation, en attribuant une crédibilité à l'expérience collective, et la notion d'individualisation du tarif, en attribuant une crédibilité à la sinistralité individuelle.

Dans le cadre de la théorie de la crédibilité, le modèle de Bühlmann est le modèle le plus répandu. Cependant, dans cette étude, ce modèle ne sera pas détaillé puisqu'il ne sera pas utilisé lors de nos

prochains traitements. En effet, les bases de données Open Damir n'ont pas cette problématique du manque d'informations sur l'historique des sinistres. De plus, l'objectif de ces travaux est d'associer les deux tarifications qui sont, à ce jour, à disposition du cabinet de conseil Actélior : la tarification santé actuelle et la tarification obtenue par l'outil de tarification construit à l'aide de la base Open Damir dans le cadre de ce mémoire. Pour cela, le facteur de crédibilité devra être intégré dans le calcul tarifaire. Il servira de poids entre ces deux tarifications. Nous pourrons alors définir X comme la tarification obtenue par l'outil actuel, et C comme la tarification nationale obtenue par l'outil de tarification Damir. La mise en place de ce modèle ainsi que la définition du facteur de crédibilité, obtenu par exemple à partir de résultats des intervalles de confiance du Modèle Linéaire Généralisé, seront explicités dans la prochaine partie.

4.1.2 Application de la crédibilité entre la tarification Actélior et la nouvelle tarification Damir

Définition du facteur de crédibilité et de la construction du modèle de crédibilité

Dans le cadre de ce mémoire, un processus de crédibilité a été mis en place. Le principe reste identique à celui d'un modèle de crédibilité : attribuer un certain poids à une tarification individuelle, et le reste à tarification collective. Cependant, le facteur de crédibilité sera défini différemment.
Dans un premier temps, définissons les différents facteurs utilisés dans les équations exposées ci-après :

- π_{Damir} et $\pi_{Actelior}$ désignent respectivement la prime pure obtenue par la tarification Damir et la prime pure obtenue par l'outil de tarification actuel d'Actélior. ;
- π_i avec i allant de 1 à 42 (le nombre de libellé brochure pour lesquels une tarification est effectuée) désigne la prime pure du libellé brochure n^o i, avec la tarification Damir ;
- $\pi_{i,Freq}$ et $\pi_{i,CM}$ désignent respectivement les tarifications obtenues pour la quantité d'actes et le coût moyen pour le libellé brochure n^o i, avec la tarification Damir ;
- $(1-\alpha_i)$ et $(1-\alpha_{i,j})$ désignent respectivement le coefficient de crédibilité associé à la prime pure du libellé brochure n^o i et celui associé à la prime pure du libellé brochure n^o i et pour la modélisation j, obtenues avec la tarification Damir. L'indice j prend la valeur 1 lorsqu'il s'agit de la modélisation du coût moyen et prend la valeur 2 lorsqu'il s'agit de la modélisation de la quantité d'actes ;
- α désigne le coefficient de crédibilité associé à la prime pure obtenue avec la tarification Actélior. $\beta = (1-\alpha)$ désigne le coefficient de crédibilité associé à la prime pure obtenue avec la tarification Damir.

Les termes étant définis, nous pouvons à présent exposer la démarche entreprise pour la mise en place de ce facteur de crédibilité. La prime pure Damir finale s'obtient par la somme des primes pures obtenues pour chaque libellé brochure, associées des facteurs de crédibilité, soit

$$\begin{aligned}\beta \times \pi_{Damir} = (1-\alpha) \times \pi_{Damir} &= \sum_i (1-\alpha_i) \times \pi_i \\ &= (1-\alpha_1)\pi_1 + (1-\alpha_2)\pi_2 + ... + (1-\alpha_{42})\pi_{42}.\end{aligned}$$

Or,

$$(1-\alpha_1)\pi_1 = (1-\alpha_{1,1})\pi_{1,CM} \times (1-\alpha_{1,2})\pi_{1,Freq}.$$

Nous obtenons donc le facteur de crédibilité suivant pour la prime pure Damir d'un libellé brochure

$$(1-\alpha_i) = (1-\alpha_{i,1}) \times (1-\alpha_{i,2}).$$

Il ne reste plus qu'à déterminer la valeur des coefficients de crédibilité pour la tarification de la quantité d'actes et du coût moyen. Nous verrons dans une prochaine partie que le coefficient de crédibilité est défini par

$$\mathbf{1-\alpha_{i,j}} = \frac{Nombre\ de\ variables\ considérées\ comme\ non\ aberrantes\ dans\ le\ modèle\ j}{Nombre\ de\ variables\ utilisées\ dans\ le\ modèle\ j}.$$

Finalement, le coefficient de crédibilité appliqué à la tarification Actélior vaudra

$$\alpha = 1-\beta = 1 - \frac{\sum_i (1-\alpha_i) \times \pi_i}{\pi_{Damir}}.$$

Sinon, ce coefficient de crédibilité peut être déterminé notamment par libellé brochure, ce qui donne

$$\alpha_i = 1-\beta_i = 1 - (1-\alpha_{i,1}) \times (1-\alpha_{i,2}).$$

Application et résultats obtenus

Les résultats sont obtenus pour le libellé brochure « Actes d'anesthésie » et pour les variables d'intérêt DEPENSE_MOY et QUANTITE_ACTES. Les différents résultats appuieront de manière concrète les explications précédentes.

Commençons nos analyses sur le modèle implémenté pour la dépense moyenne. Après avoir validé le modèle, l'objectif est de déterminer le coefficient de crédibilité qui sera appliqué à la prime selon Coût moyen x Fréquence pour les actes d'anesthésie. Dans un premier temps, les intervalles de confiance des coefficients GLM, ainsi que la longueur de ces intervalles sont déterminés pour chacune des variables du modèle implémenté. Le graphique 4.1 est tracé pour visualiser plus clairement la longueur de ces intervalles de confiance.

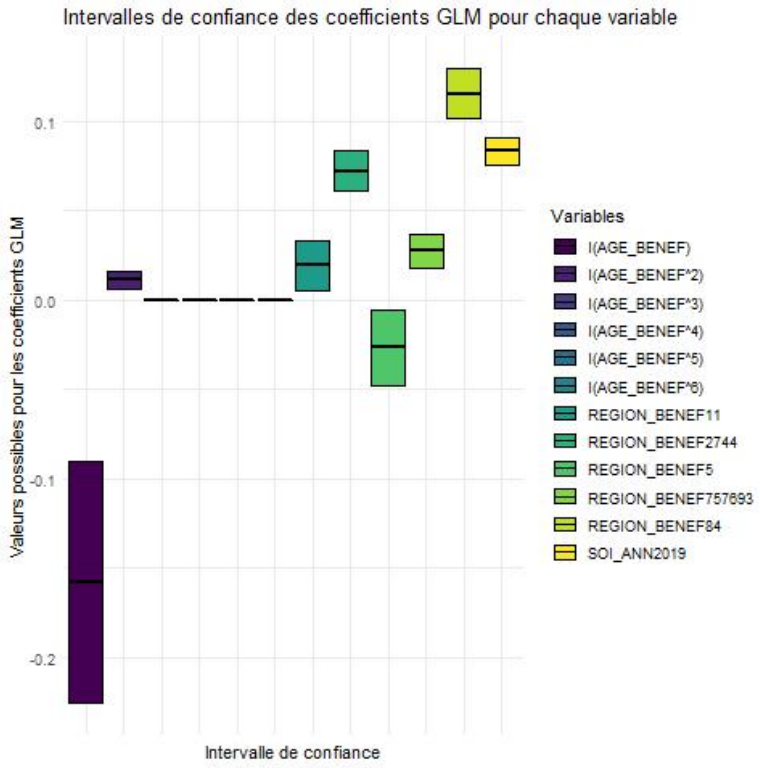


FIGURE 4.1 : Intervalles de confiance obtenus pour chaque variable du GLM de la dépense moyenne des actes d'anesthésie.

Nous remarquons que chaque variable a une longueur différente, ce qui nécessite de réfléchir à la longueur idéale. Comment déterminer les variables pour lesquelles l'intervalle de confiance est trop grand, et donc pour lesquelles la prédiction est de mauvaise qualité ? A partir de quel seuil les intervalles de confiance sont-ils considérés comme trop importants ? Pour cela, le choix d'un seuil a été écarté. La démarche s'appuie sur la détection de valeurs aberrantes, à l'aide d'un tracé de boxplot sur la longueur des intervalles de confiance. Pour rappel, pour une variable quantitative donnée, le boxplot permet de communiquer des informations sur les quantiles, les valeurs extrêmes, la moyenne, la médiane et enfin les points considérés comme aberrants pour la variable en question. Sous *R*, l'identification de valeurs aberrantes à partir d'un boxplot est automatique. Si nous exécutons ce code, deux variables sont considérées comme outliers au niveau de la longueur des intervalles de confiance, et sont visibles sur le boxplot 4.2.

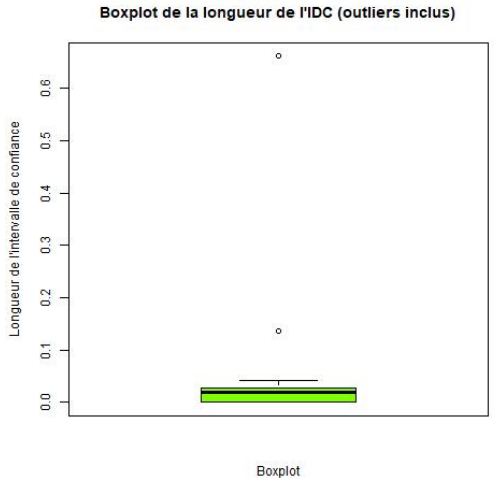


FIGURE 4.2 : Boxplot de la longueur de l'IDC (points aberrants inclus) du GLM de la dépense moyenne des actes d'anesthésie.

Ces deux variables sont INTERCEPT (le coefficient associé aux caractéristiques de l'individu de référence) et AGE_BENEF. En effet, nous pouvons l'observer sur le graphique 4.1 (sauf pour la variable INTERCEPT qui n'est pas affichée car la longueur de son intervalle de confiance est deux fois plus importante par rapport à celle des autres variables). L'intervalle de confiance est important pour la variable AGE_BENEF, par rapport aux autres. Si les valeurs aberrantes sont éliminées, le boxplot 4.3 obtenu est plus cohérent puisque l'ensemble des longueurs des intervalles de confiance sont plus proches les unes des autres.

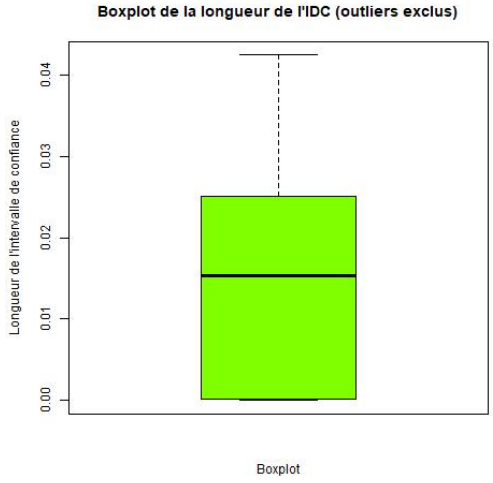


FIGURE 4.3 : Boxplot de la longueur de l'IDC (points aberrants exclus) du GLM de la dépense moyenne des actes d'anesthésie.

Dans ce cas, le coefficient de crédibilité appliqué sur la tarification de la dépense moyenne des actes d'anesthésie est de

$$(\mathbf{1} - \alpha_{\mathbf{ActesAnesthésie,1}}) = \frac{Nombre\ de\ variables\ considérées\ comme\ non\ aberrant}{Nombre\ de\ variables\ utilisées\ dans\ le\ modèle} = \frac{11}{13} = \mathbf{0.8462}.$$

Cette méthode est effectuée une nouvelle fois pour le **modèle GLM sur la quantité d'actes**. Les intervalles de confiance obtenus pour cette variable d'intérêt sont représentés sur le graphique 4.4.

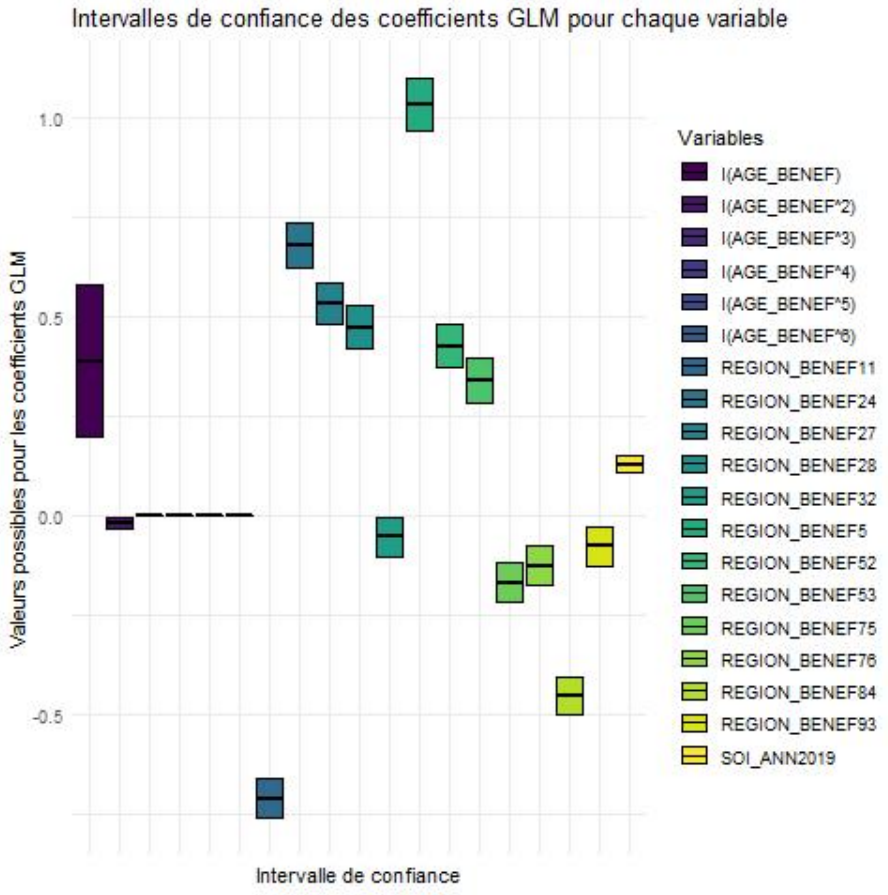


FIGURE 4.4 : Intervalles de confiance obtenus pour chaque variable du GLM de la quantité des actes d'anesthésie.

Nous remarquons que pour chaque variable, la longueur de l'intervalle de confiance des coefficients GLM est plus ou moins importante. Le plus grand intervalle de confiance est associé à la variable AGE_BENEF. Vérifions ce résultat par le biais des boxplots de la longueur de l'intervalle. D'après le graphique 4.5, deux variables sont considérées comme aberrantes au niveau de la longueur des intervalles de confiance, et sont visibles sur le boxplot.

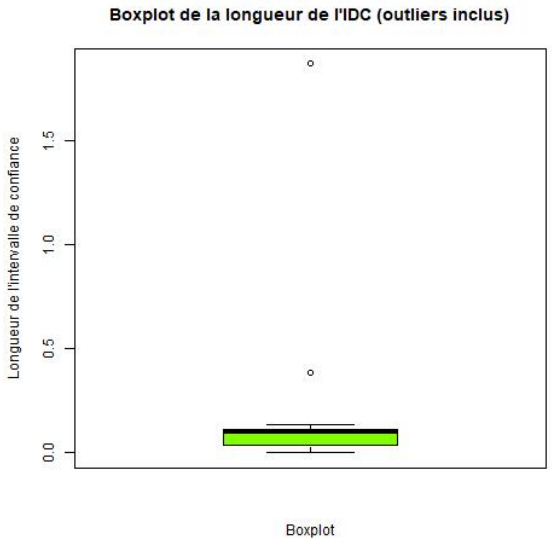


FIGURE 4.5 : Boxplot de la longueur de l'IDC (points aberrants inclus) du GLM de la quantité des actes d'anesthésie.

Ces deux variables sont INTERCEPT (le coefficient associé aux caractéristiques de l'individu de référence) et AGE_BENEF. En effet, nous pouvons l'observer sur le graphique 4.4, sauf pour la variable INTERCEPT car elle n'est pas affichée sur le graphique. Son intervalle de confiance n'est situé au même niveau que pour les autres variables. Les coefficients GLM sont compris entre -11 et -13, ce qui rend le graphique illisible. En revanche, le tableau récapitulant les intervalles de confiances pour chaque variable explicative est disponible dans les annexes A.14 et A.15. Si ces variables sont éliminées, le boxplot 4.6 obtenu est plus cohérent puisque l'ensemble des longueurs des intervalles de confiance sont plus proches les unes des autres.

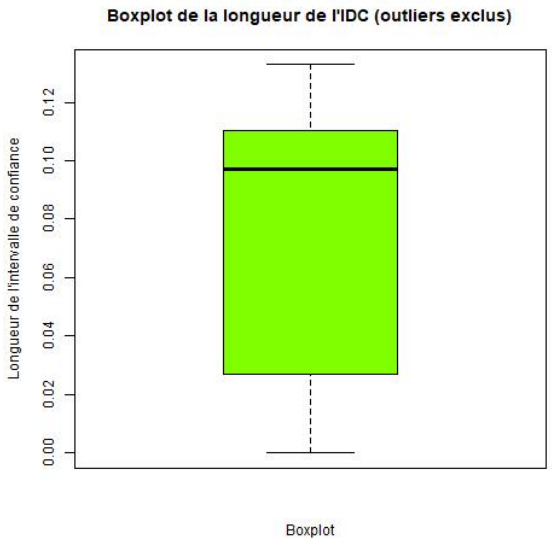


FIGURE 4.6 : Boxplot de la longueur de l'IDC (points aberrants exclus) du GLM de la quantité des actes d'anesthésie.

Dans ce cas, le coefficient de crédibilité appliqué sur la tarification de la quantité d'actes d'anesthésie est de

$$(\mathbf{1} - \alpha_{\mathbf{ActesAnesthésie,2}}) = \frac{18}{20} = \mathbf{0.9}.$$

Finalement, le coefficient de crédibilité qui sera appliqué à la prime pure des actes d'anesthésie obtenue par la tarification Damir vaut

$$(\mathbf{1} - \alpha_{\mathbf{ActesAnesthésie}}) = (\mathbf{1} - \alpha_{\mathbf{ActesAnesthésie,1}}) \times (\mathbf{1} - \alpha_{\mathbf{ActesAnesthésie,2}}) = 0.8456 \times 0.9 = 0.76104.$$

Le coefficient $\alpha_{ActesAnesthésie}$ qui sera appliqué à la prime pure des actes d'anesthésie obtenue par la tarification Actélior vaut

$$\alpha_{\mathbf{ActesAnesthésie}} = 1 - 0.76104 = 0.23896.$$

Par la crédibilité, nous avons donc affiné la tarification santé. Concentrons-nous à présent sur le suivi technique santé Damir, permettant l'analyse du risque santé au niveau national.

4.2 Création d'un suivi technique santé Damir

A ce jour, les suivis techniques réalisés chez Actélior ne s'intéressent qu'aux consommations propres à chaque client. Pour donner plus de corps à ces études, il a été décidé de les enrichir de données nationales, celles des bases Open Damir, de manière à comparer les différences de consommation entre le portefeuille étudié et l'ensemble de la France, d'une région, d'une tranche d'âge ou de tout autre critère exploitable.

4.2.1 Présentation de l'objectif du suivi technique d'Actélior

Le suivi technique, mis en place par Actélior, constitue un outil de reporting actuariel et statistique permettant d'établir une analyse approfondie du portefeuille de chaque client (majoritairement composé de groupes mutualistes). Il est alimenté de façon plus ou moins régulière en fonction des besoins de chaque client (trimestriellement, semestriellement ou encore annuellement). Ce suivi technique est mis en place à partir de bases de données du client, exploitées à l'aide d'outils informatiques de gestion de bases de données et de statistiques, comme *R* et Access. L'exercice en cours de réalisation est projeté jusqu'à la date d'échéance des contrats puis comparé aux exercices précédents.

Grâce à ce processus de traitement de la donnée, le suivi technique permet de répondre à de nombreuses problématiques portant sur l'évolution du portefeuille, et plus particulièrement, l'évolution des effectifs, des dépenses santé et de leur prise en charge, etc. En effet, dans un premier temps, une vision macroscopique est effectuée pour fournir des éléments globaux sur le portefeuille, puis une vision microscopique dans laquelle des analyses plus précises sont fournies. Qu'est-ce que cela signifie précisément ? Ci-dessous, nous donnons chacun des axes exploités et analysés en partant de la vision macroscopique jusqu'à la vision microscopique :

- analyse globale du portefeuille,
- étude des effectifs,
- analyse globale par gamme, par produit et par option,
- analyse globale par type de bénéficiaire,
- analyse similaire mais par bénéficiaire,
- analyse similaire par famille d'actes ,
- analyse similaire par acte.

Enfin, une analyse spécifique des impacts de la réforme 100% santé est réalisée dans le suivi technique. Elle permet de comprendre l'évolution de la consommation santé pour les trois postes concernés par cette réforme : l'optique, le dentaire et l'audiologie (c.f. section 1.1.3). Pour rappel, dans le cadre de ce mémoire, **la consommation santé désigne les frais réels ou la dépense du soin réalisé par l'assuré**. De plus, le suivi technique intègre une analyse poussée du 100% santé puisqu'il s'agit, à ce jour, d'un sujet prédominant sur le marché de l'assurance et qui inquiète beaucoup. Le suivi technique intégrera dans les années à venir d'autres analyses en fonction des évolutions réglementaires, et sera enrichi d'analyses demandées par les clients.

L'ensemble des résultats de ces analyses est présenté dans le suivi technique, créé à partir d'un classeur Excel, sous la forme de tableaux statistiques récapitulatifs, mais aussi de graphiques afin de

distinguer plus facilement les évolutions, et d'en déduire des conclusions correctes. Il s'agit donc d'un outil de reporting précis permettant l'analyse du risque santé en détail du portefeuille client.

Actélior souhaite apporter une dimension nouvelle aux suivis techniques en les enrichissant d'éléments de comparaison. C'est donc dans le cadre de ce mémoire qu'une analyse complémentaire via les bases Open Damir a été réalisée.

4.2.2 Présentation de l'outil de reporting créé à partir de la base Open Damir

Objectif du suivi technique Damir

Le suivi technique réalisé dans le cadre de ce mémoire, a donc pour objectif de compléter l'analyse technique du portefeuille client. Ce suivi technique contiendra des analyses similaires au suivi technique, mais sur un **périmètre national**. Pour cela, la base Open Damir sera utilisée avec un accent porté sur les dépenses de santé et ceci pour tous les postes concernés par le remboursement obligatoire de la Sécurité Sociale. De plus, une analyse des impacts de la réforme 100% santé permettra d'analyser avec détail l'évolution des dépenses nationales pour les trois postes concernés par cette réforme : l'optique, le dentaire et l'audiologie. Ce complément permettra aux clients d'Actélior de comparer les données propres de leur portefeuille aux données nationales. Toutefois, pour des raisons de volumétrie, le suivi technique Damir ne sera mis à jour qu'annuellement à partir des données disponibles à fin juin alors que les suivis techniques des clients d'Actélior sont mis à jour plus régulièrement.

Détails sur le processus effectué et le périmètre étudié

Le suivi technique sera, dans le cadre de ce mémoire, établi sur les années 2018 et 2019. La base d'étude issue des différents retraitements précédents (c.f. section 2.2) contient 10 040 857 lignes, représentant des données agrégées selon des variables qualitatives fixées. Elle représente les remboursements de la Sécurité Sociale, établies alors sur un périmètre national. Le nombre de bénéficiaires de cette base d'étude concerne donc la population assurée au Régime Général de l'Assurance Maladie Obligatoire. D'après le tableau 4.1, pour les années 2018 et 2019, le nombre de bénéficiaires total concerné par les analyses de ce suivi technique s'élève à

TABLE 4.1 : Nombre de bénéficiaires pour les années de survenance 2018 et 2019.

	2018	2019	Evolution
Nombre de bénéficiaires	62 400 k	59 200 k	-5%

Cette diminution de 5% entre les deux années n'est pas expliquée. Aucune information n'est communiquée à ce sujet par l'Assurance Maladie. Les études ci-après prendront en compte cette diminution.

Comme expliqué précédemment en section 2.2, les qualités « Conjoint » et « Autres ayant droit » ont été regroupés avec « Assuré » pour former la qualité de bénéficiaire « Adulte ». Les deux qualités bénéficiaires disponibles pour l'étude sont donc « Adulte » et « Enfant ». Ce retraitement a été effectué *a posteriori* d'une première version du suivi technique. En effet, le montant de la dépense totale pour ces deux catégories était faible. Ces deux qualités de bénéficiaires correspondent à des cas exceptionnels dans le cadre de l'assurance maladie obligatoire. Toute personne résidente en France est assurée à titre individuel à l'Assurance Maladie Obligatoire, contrairement aux contrats complémentaires santé, où

le bénéficiaire peut être affilié au contrat de son conjoint. Le second traitement réalisé sur la base d'étude est l'intégration d'une Provisions pour Prestations A Payer (PPAP) et donc le recalcul des indicateurs de montants à partir de l'estimation de la PPAP. L'explication théorique et le processus calculatoire sont donnés ci-après.

Explication de l'estimation de la PPAP

La Provision pour Prestations à Payer (PPAP), plus communément appelé Provision pour Sinistres à Payer (PSAP) dans d'autres domaines du secteur de l'assurance, comprend l'estimation de la charge des sinistres non survenus. En effet, comme énoncé précédemment, le suivi technique est réalisé chaque trimestre. Pour sa construction, nous choisissons l'année N-1 et l'année N en cours. Or, au mois de mars de l'année N, nous n'avons pas connaissance des sinistres futurs. La comparaison entre l'année N-1 et l'année N est donc impossible. C'est dans ce cadre que la PPAP intervient et donne une estimation des montants de sinistres non survenus pour les mois suivants de l'année N. Le calcul de la PPAP s'effectue donc selon les années de survenance et les années de remboursement. Les années pendant lesquelles chaque règlement est effectué s'appelle les années de développement (charge pour l'organisme assureur), par opposition aux années de survenance, qui correspondent aux années où les sinistres se sont réellement produits.

Dans le cadre du suivi technique Damir, l'estimation de la PPAP permet d'ajouter les prestations santé survenues en 2019 (ou avant) mais qui ne seront payées qu'à partir de 2020.

Pour l'estimation de la PPAP, la méthode de Chain-Ladder a été choisie. Les PPAP seront estimées pour les années de survenance 2018 et 2019 selon deux périmètres :

- le périmètre « Hospitalisation »,
- le périmètre « Hors hospitalisation ».

Cela se justifie par le fait que les cadences de liquidation pour le poste « Hospitalisation » sont très différentes de celles observées sur les autres postes. La méthode de Chain-Ladder sera donc réalisée deux fois.

Dans le cadre de l'estimation des PPAP, un tableau récapitule les charges ultimes ainsi que les montants de remboursements déjà versés, par mois et année de survenance. Les provisions par année de survenance s'obtiennent par différence des deux valeurs (c.f. tableau 4.2).

TABLE 4.2 : Montant des provisions pour le poste « Hospitalisation » et les années de survenance 2018 et 2019.

Poste - Hospitalisation	
Année de survenance	Montant de provision
2018	12 065 378 €
2019	2 157 003 865 €

Soit S_N la charge ultime totale sur l'année de survenance N, S_N^P les montants déjà payés pour l'année N. Le taux de PPAP pour l'année de survenance N, noté τ_{PPAP_N}, calculé pour le poste « Hospitalisation » et pour l'ensemble des autres postes, est alors déterminé par la formule suivante

$$S_N = (1 + \tau_{PPAPN}) \times S_N^P,$$

c'est-à-dire

$$\tau_{PPAPN} = \frac{S_N}{S_N^P} - 1.$$

Après application de la méthode de Chain-Ladder, les taux de PPAP obtenus pour les deux années de survenance 2018 et 2019 et les deux cas sont présentés dans le tableau 4.3.

TABLE 4.3 : Taux de PPAP obtenus pour le poste « Hospitalisation » et « Hors hospitalisation » et les années de survenance 2018 et 2019.

	HORS HOSPITALISATION	HOSPITALISATION
Taux de PPAP 2019	5.34%	9.15%
Taux de PPAP 2018	0.10%	0.05%

De plus, l'hypothèse faite initialement, sur la similarité de l'évolution du comportement de chaque cadence de remboursement, est vérifiée. En effet, d'après les graphiques 4.7 et 4.8, les charges ultimes sont plutôt similaires pour chaque mois et pour les deux années de survenance 2018 et 2019, que ce soit pour les soins hospitaliers ou les soins hors hospitalisation.

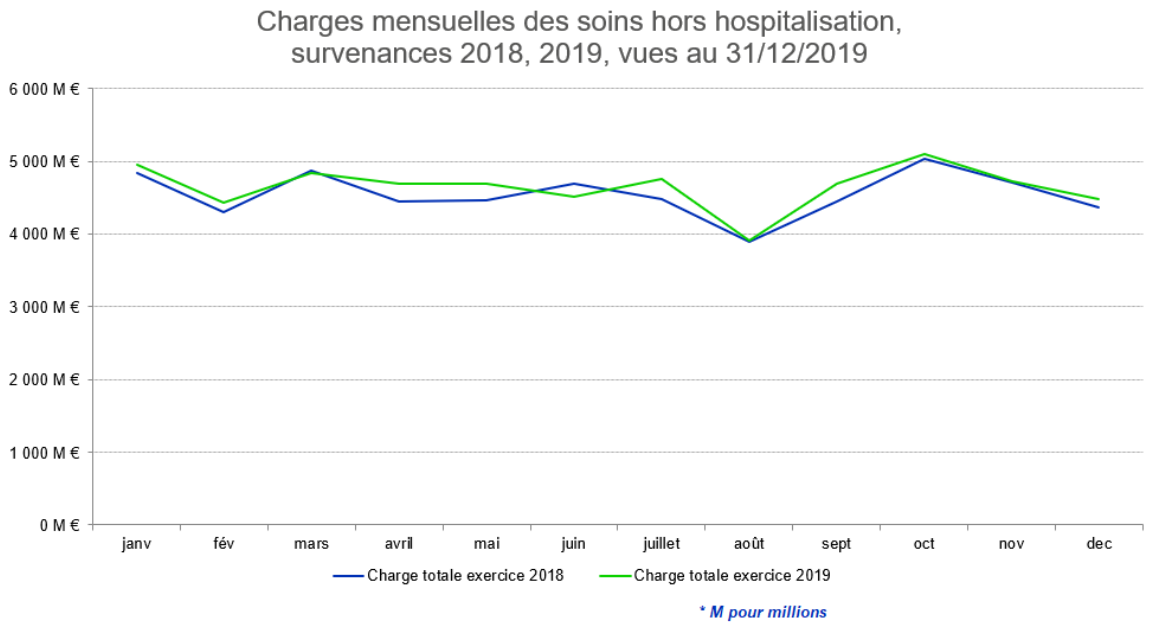


FIGURE 4.7 : Graphique de la charge mensuelle des soins hors hospitalisation pour les années de survenances 2018 et 2019 vu à fin décembre 2019.

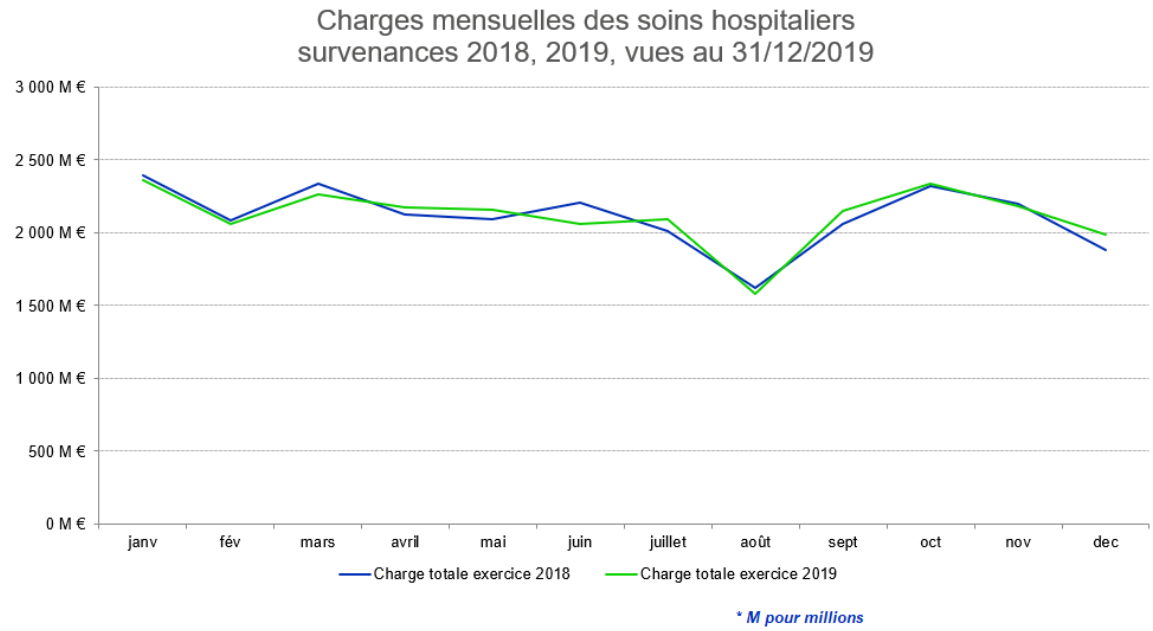


FIGURE 4.8 : Graphique de la charge mensuelle des soins hospitaliers pour les années de survenances 2018 et 2019 vu à fin décembre 2019.

Présentons maintenant le suivi technique Damir créé.

Présentation du suivi technique Damir

La détermination des taux de PPAP est à présent terminée. Ces taux ont été appliqués aux indicateurs de montants de la base d'étude traitée Open Damir. C'est à partir de cette base que le suivi technique Damir va être construit.

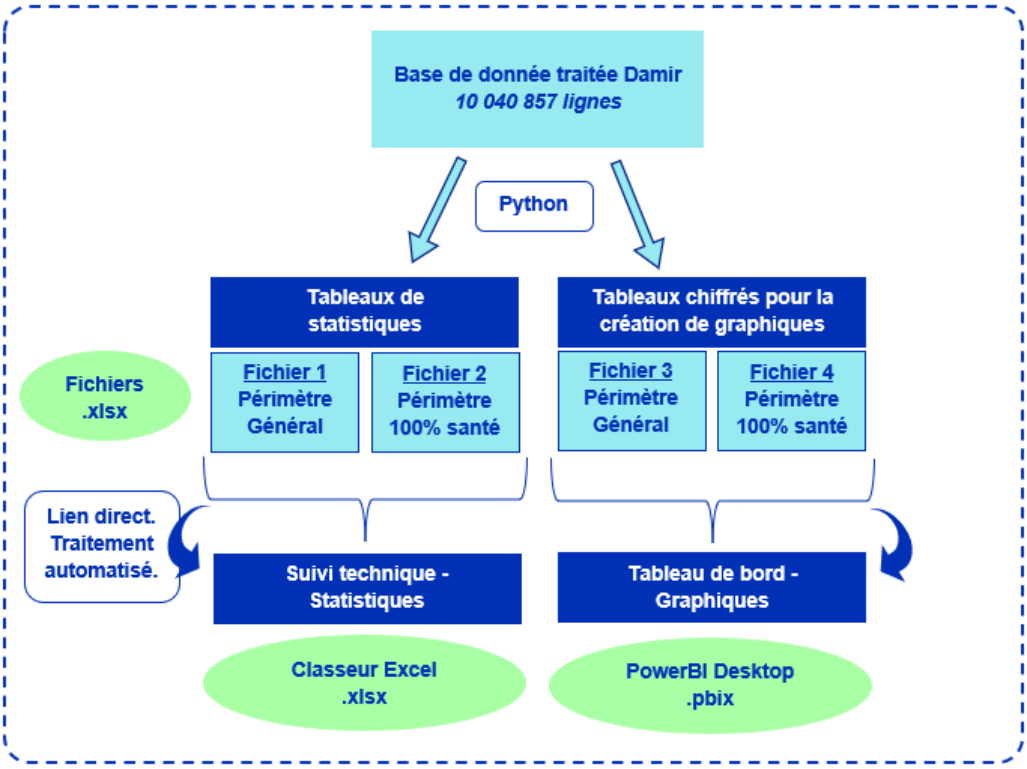


FIGURE 4.9 : Schéma récapitulatif du processus de construction du suivi technique Damir.

Pour la construction de cet outil de reporting, de nombreux outils et logiciels ont été sollicités. En effet, la base d'étude reste une base volumineuse de 286 781 855 lignes après nettoyage des données (sans agrégation). Il n'est donc pas possible de réaliser les calculs manuellement par le biais d'un classeur Excel. La création du suivi technique Damir comportent trois étapes (c.f. schéma 4.9).

Premièrement, les calculs ont été effectués sur $\mathbb{P}ython$ afin d'exporter des tableaux de statistiques bruts au format Classeur Excel. Quatre fichiers sont créés :

- les tableaux de statistiques globales de la base Open Damir à mettre en forme ;
- les tableaux de statistiques globales de la base Open Damir utilisés pour la création de graphiques ;
- les tableaux de statistiques sur les impacts de la réforme 100% santé, à mettre en forme ;
- les tableaux de statistiques sur les impacts de la réforme 100% santé utilisés pour la création de graphiques.

L'utilisation de ℙ*ython* permettra pour les prochaines années un traitement et une exportation des résultats rapides pour la création du suivi technique. La volumétrie des bases Open Data va notamment fortement augmenter d'ici les prochaines années. Le choix de ℙ*ython* permettra donc de contourner la problématique de limite de mémoire ou de volume qu'un fichier Excel pourrait rencontrer s'il contient une base de données aussi volumineuse.

Après l'obtention des statistiques de la base d'étude Open Damir, les résultats chiffrés et les graphiques doivent être mis en forme. Deux fichiers ont donc été créés :

- un suivi technique contenant l'ensemble des statistiques chiffrées sous le format de classeur Excel ;
- un tableau de bord Power BI contenant l'ensemble des graphiques sur le périmètre global de la base Open Damir mais aussi sur chacun des postes concernés par le 100% santé.

Le contenu de ces deux fichiers va être détaillé tout de suite.

Suivi technique via le classeur Excel Ce suivi technique contient 3 onglets : un onglet paramètre, un onglet « Global » qui recense l'intégralité des statistiques globales de la base Open Damir, et un onglet « 100% Santé » qui recense l'intégralité des statistiques sur les postes concernés par la réforme 100% santé. Les informations recensées sont similaires à celles contenues dans le suivi technique Actélior. Les captures 4.10 et 4.11 donnent un aperçu des deux onglets de résultats.

Synthèse globale

	2018	2019	Evolution
Nombre de bénéficiaires	62 400 k	59 200 k	-5%
Tranche d'âge moyenne des bénéficia	40-49 ans	40-49 ans	-

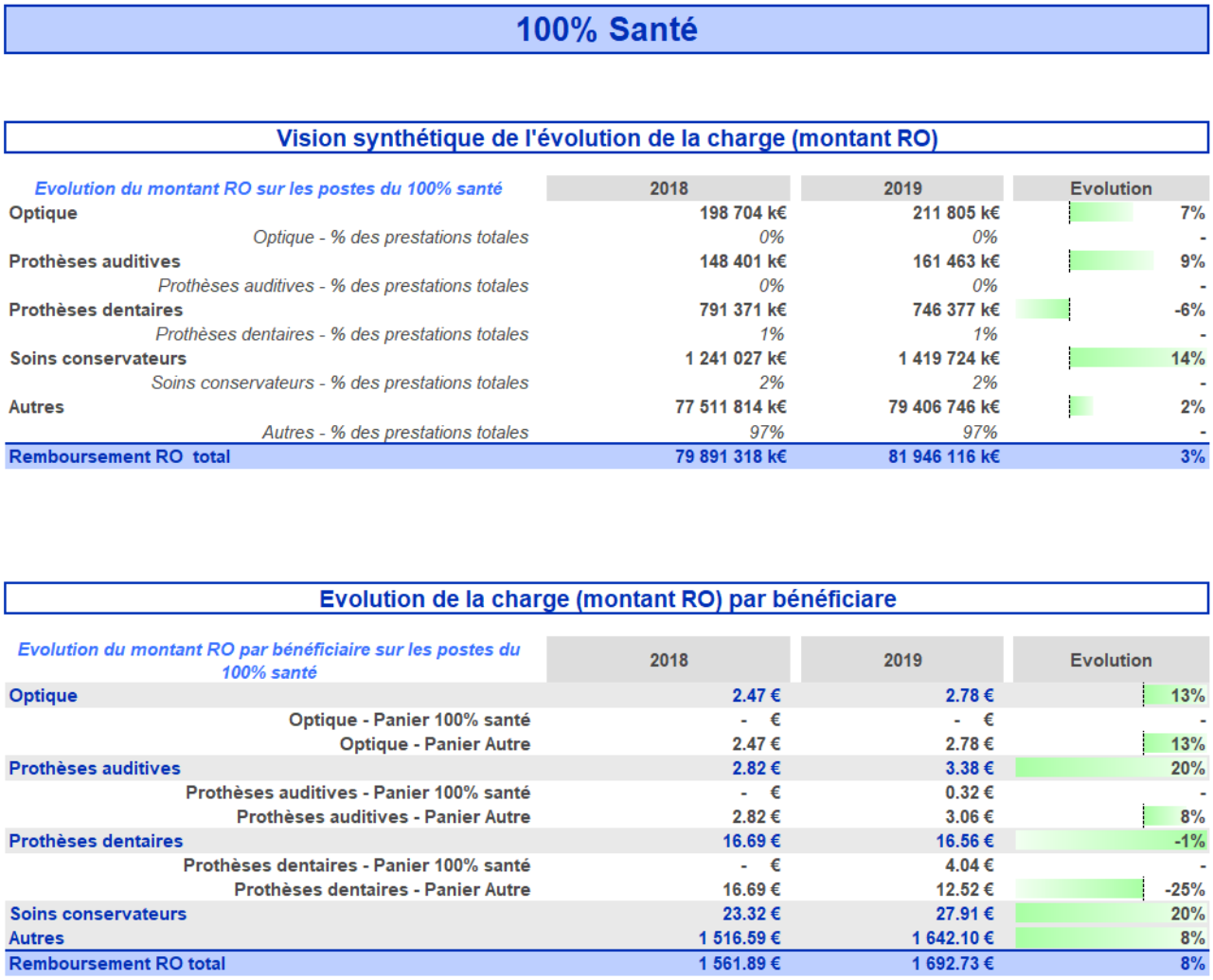
Charge (montant RO) et consommation (Frais réels) totales

	2018	2019	Evolution
Remboursement RO total	79 891 318 k€	81 946 116 k€	3%
Dépense totale	110 491 218 k€	113 961 206 k€	3%

Charge totale (montant RO) par famille d'actes

Montant total du RO par famille d'actes	2018	Part (%) 2018	2019	Part (%) 2019	Evolution
SOINS COURANTS	26 455 995 k€	33%	26 925 316 k€	33%	2%
PHARMACIE	19 222 657 k€	24%	20 164 766 k€	25%	5%
HOSPITALISATION	25 333 219 k€	32%	25 717 949 k€	31%	2%
DENTAIRE	3 264 776 k€	4%	3 414 881 k€	4%	5%
OPTIQUE	198 704 k€	0%	211 805 k€	0%	7%
APPAREILLAGE	5 182 274 k€	6%	5 261 034 k€	6%	2%
AUTRES	233 692 k€	0%	250 365 k€	0%	7%
Remboursement RO total	79 891 318 k€	100%	81 946 116 k€	100%	3%

FIGURE 4.10 : Présentation d'une partie de l'onglet « Global » du suivi technique Damir.



100% Santé

Vision synthétique de l'évolution de la charge (montant RO)

Evolution du montant RO sur les postes du 100% santé	2018	2019	Evolution
Optique	198 704 k€	211 805 k€	7%
Optique - % des prestations totales	*0%*	*0%*	-
Prothèses auditives	148 401 k€	161 463 k€	9%
Prothèses auditives - % des prestations totales	*0%*	*0%*	-
Prothèses dentaires	791 371 k€	746 377 k€	-6%
Prothèses dentaires - % des prestations totales	*1%*	*1%*	-
Soins conservateurs	1 241 027 k€	1 419 724 k€	14%
Soins conservateurs - % des prestations totales	*2%*	*2%*	-
Autres	77 511 814 k€	79 406 746 k€	2%
Autres - % des prestations totales	*97%*	*97%*	-
Remboursement RO total	**79 891 318 k€**	**81 946 116 k€**	**3%**

Evolution de la charge (montant RO) par bénéficiaire

Evolution du montant RO par bénéficiaire sur les postes du 100% santé	2018	2019	Evolution
Optique	**2.47 €**	**2.78 €**	**13%**
Optique - Panier 100% santé	- €	- €	-
Optique - Panier Autre	2.47 €	2.78 €	13%
Prothèses auditives	**2.82 €**	**3.38 €**	**20%**
Prothèses auditives - Panier 100% santé	- €	0.32 €	-
Prothèses auditives - Panier Autre	2.82 €	3.06 €	8%
Prothèses dentaires	**16.69 €**	**16.56 €**	**-1%**
Prothèses dentaires - Panier 100% santé	- €	4.04 €	-
Prothèses dentaires - Panier Autre	16.69 €	12.52 €	-25%
Soins conservateurs	**23.32 €**	**27.91 €**	**20%**
Autres	**1 516.59 €**	**1 642.10 €**	**8%**
Remboursement RO total	**1 561.89 €**	**1 692.73 €**	**8%**

FIGURE 4.11 : Présentation d'une partie de l'onglet « 100% santé » du suivi technique Damir.

L'interprétation des résultats obtenus sera détaillée dans la section 4.2.3.

Tableau de bord Power BI Power BI est une application téléchargeable gratuitement et qui appartient au groupe MICROSOFT (2014). L'application permet d'importer une base de données, de la modifier puis de visualiser ces données de façon dynamique. Son interface permet principalement de créer des graphiques via une collection de visuels préexistants mais aussi de créer des tableaux. Cette application concilie l'exploitation de données et la prise de décisions stratégiques par l'affichage de graphiques statistiques complexes et dynamiques, le tout recenser dans un tableau de bord. Power BI Desktop contient trois onglets détaillés ci-après :

- **Onglet « Vue rapport » :** visualiser le tableau de bord en cours de création avec les différents visuels créés ;
- **Onglet « Vue donnée » :** visualiser les bases de données préalablement importées et utilisées pour la création du tableau de bord. Dans le cadre de l'étude, cette vue permettra de changer le format des variables : en monétaire pour les variables de montants et avec séparateur des milliers pour les variables de quantités ;

- **Onglet « Vue modèle » :** visualiser les relations entre les tables utilisées. Les relations servent principalement à rendre le tableau de bord interactif et dynamique, avec l'ajout de filtres par exemple.

De plus, l'éditeur Power Query, présent dans Power BI Desktop, permet de réaliser des modifications de données souhaitées sur une base initiale, avant de créer les graphiques à partir d'une base traitée. Les différentes fonctionnalités sont disponibles. L'unique fonctionnalité utilisée dans le cadre de l'étude est la modification des noms des en-têtes de colonnes (ou nom des variables). En effet, l'application Power BI a ses limites. Pour des bases volumineuses, le traitement de données via Power BI peut alourdir le fichier et ralentir considérablement l'exécution. Nous avons donc réaliser les traitements au préalable sur $\mathbb{P}ython$. Ce travail peut être effectué à l'aide d'outils de gestion ou de traitement de bases de données.

Après cette transformation des données, il est possible de débuter la réalisation de graphiques à partir de la nouvelle base de données modifiée. Ce large choix de visuels permet au créateur du tableau de bord de choisir celui qui représente au mieux ses données ou les statistiques qu'ils souhaitent mettre en évidence. Plusieurs types de visuels existent :

- histogramme,
- filtre,
- graduation temporelle,
- disque / Camembert,
- courbe,
- carte,
- etc.

L'illustration 4.12 montre la première page du tableau de bord Power BI Desktop créée dans le cadre du suivi technique Damir.

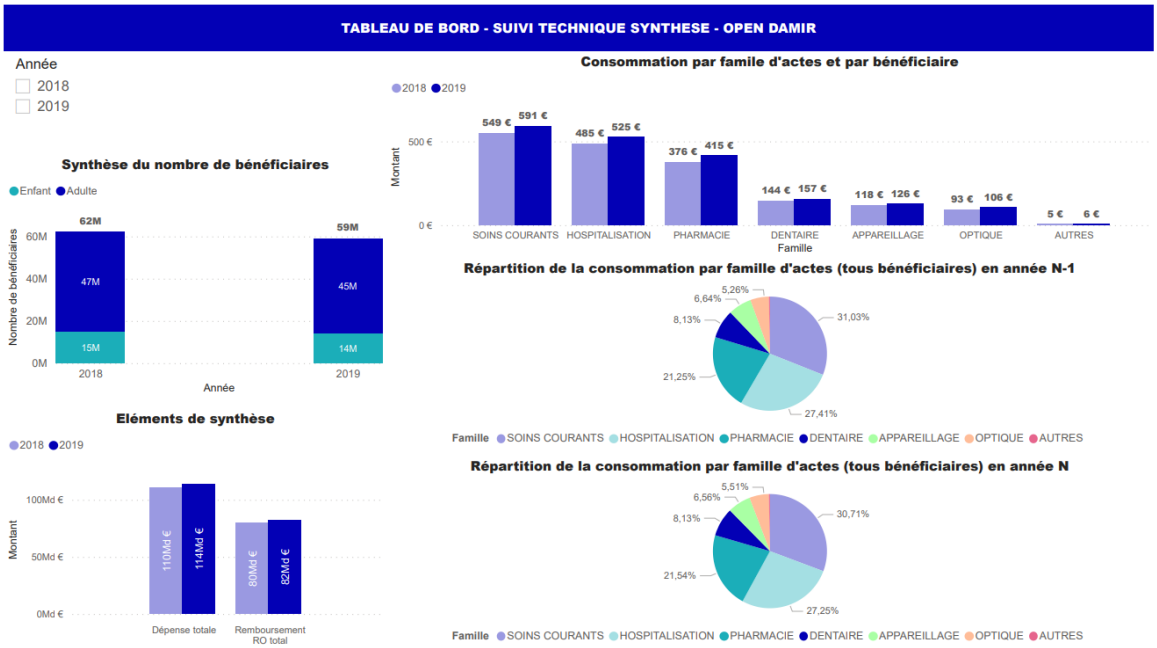


FIGURE 4.12 : Présentation de la première page du tableau de bord Power BI du suivi technique Damir.

Le tableau de bord comporte six pages :

- **Global – Synthèse :** Visualisation de la répartition de la consommation santé selon les familles d'actes et l'année, de la répartition du nombre de bénéficiaires ainsi que des éléments de synthèse sur la charge et la dépense totale par année ;
- **Général - 100% santé :** Visualisation de la charge totale selon les postes concernés par la réforme 100% santé ;
- **Dentaire – 100% santé :** Visualisation de l'évolution de différents facteurs concernant le poste « Dentaire » comme le nombre de prothèses dentaire, au total ou par bénéficiaire, selon le type de paniers ou la tranche d'âge, mais aussi la décomposition de la prise en charge moyenne d'une prothèse dentaire par année ;
- **Audiologie – 100% santé :** Visualisation de l'évolution de différents facteurs concernant le poste « Audiologie » comme le nombre de prothèses auditives, au total ou par bénéficiaire, selon le type de paniers ou la tranche d'âge, mais aussi la décomposition de la prise en charge moyenne d'une prothèse auditive par année ;
- **Optique – 100% santé :** Visualisation de l'évolution de différents facteurs concernant le poste « Optique » comme le nombre total de montures et de verres optique, mais aussi la décomposition de la prise en charge moyenne d'une monture ou d'un verre optique par année ;
- **Optique (2) – 100% santé :** Visualisation de l'évolution de différents facteurs concernant le poste « Optique » comme le nombre moyen d'équipement optique selon la tranche d'âge, mais aussi la décomposition de la prise en charge moyenne d'un équipement optique par année.

Sur ces quatre pages, des filtres peuvent être rajoutés afin de rendre dynamique le tableau de bord. Ces filtres permettent par exemple de regarder les statistiques seulement sur une année, une catégorie

d'âge ou même, une famille d'acte précis. Pour rendre des graphiques interactifs, il ne faut cependant pas oublier de créer le modèle de relation de la base de données. Dans le cadre de l'étude, le suivi technique ne sera pas rendu interactif pour la première version. Cependant, ces outils peuvent être amenés à évoluer. L'objectif du suivi technique reste principalement axé sur la compréhension des flux nationaux en consommation santé, pour faire bénéficier de cette expertise aux nombreux clients d'Actélior. Il est notamment possible de créer des cartes nationales avec le jeu de couleurs associés aux statistiques de la base de données sur Power BI. Une analyse par département pourra par exemple être une fonctionnalité à rajouter dans une nouvelle version du suivi technique.

Tous ces fichiers créés sont automatisés. Ils vont pouvoir être réutilisés les prochaines années. Seul le chemin de la source des données importées doit être modifiée. Attention, il est cependant essentiel que les nouveaux tableaux à importer aient la même structure que les tableaux statistiques initiaux avec lesquels l'outil de reporting a été créés.

4.2.3 Exemple d'interprétation d'éléments de synthèse

Après avoir détaillé la forme du suivi technique Damir, nous allons à présent analyser le fond de ce dernier. Cette analyse permettra de comprendre davantage la tendance des dépenses et des remboursements présents dans la base de données, et fournira une idée sur la consommation santé nationale. Pour rappel, le montant de la charge indiqué au sein du suivi technique correspond au montant du remboursement de la Sécurité Sociale et non pas des organismes de complémentaires santé.

Cette partie n'analysera pas l'ensemble du suivi technique, car nombreux sont les tableaux statistiques et graphiques évolutifs. L'analyse sera axée sur l'objectif de la mise en place de la réforme 100% santé, et non pas son impact. En effet, ce mémoire porte sur les survenances et remboursements des soins des années 2018 et 2019. La réforme n'était donc pas entièrement mise en place pour ces années-là. Seuls les différents paniers (deux pour l'audiologie et l'optique, et trois pour le dentaire) ont été créés. En revanche, le reste à charge n'était pas nul en 2019 pour le panier 1 (panier 100% santé) sauf pour quelques prothèses dentaires. C'est pourquoi l'impact de cette réforme ne peut pas être mesuré sur les années étudiées. Puis, une analyse plus microscopique sera réalisée, notamment par rapport à la dépense et au reste à charge pour l'assuré :

- répartition des frais réels et de la part de reste à charge pour chaque panier et pour le poste dentaire sur l'année 2019 ;
- analyse du poste dentaire et des trois paniers sur les années 2018 et 2019, pour les quantités et montants suivants : nombre de prothèse moyen par bénéficiaire, charge moyenne par prothèse, charge moyenne par bénéficiaire ;
- répartition du nombre de prothèses auditives par tranche d'âge pour les années 2018 et 2019.

Premièrement, sur le graphique de la « Consommation par famille d'actes et par bénéficiaire », présent au sein du graphique 4.12, nous pouvons remarquer que les bénéficiaires du Régime Général de la Sécurité Sociale consomment très peu sur les postes « Optique », « Dentaire », et « Appareillage ». Cela s'explique tout d'abord par la non-nécessité de renouveler chaque année le soin ou l'équipement, pour ces trois postes. D'une part, l'équipement de l'assuré peut toujours fonctionner après un an (pour les prothèses auditives), ou d'autre part, l'assuré n'a subit aucune évolution au niveau de sa vision et n'a donc pas besoin d'effectuer un remplacement de ses verres optique. De plus, très souvent, les organismes de complémentaires santé limitent le renouvellement des équipements optiques ou auditifs de 2 ans jusqu'à 4 ans. De plus, cette consommation faible au niveau de ces trois postes s'explique

notamment par une prise en charge faible de la dépense par la Sécurité Sociale. Le reste à charge pour l'assuré (hors contrat complémentaire santé) est donc beaucoup plus important que sur les autres postes. Cela se remarque notamment sur le graphique 4.13.

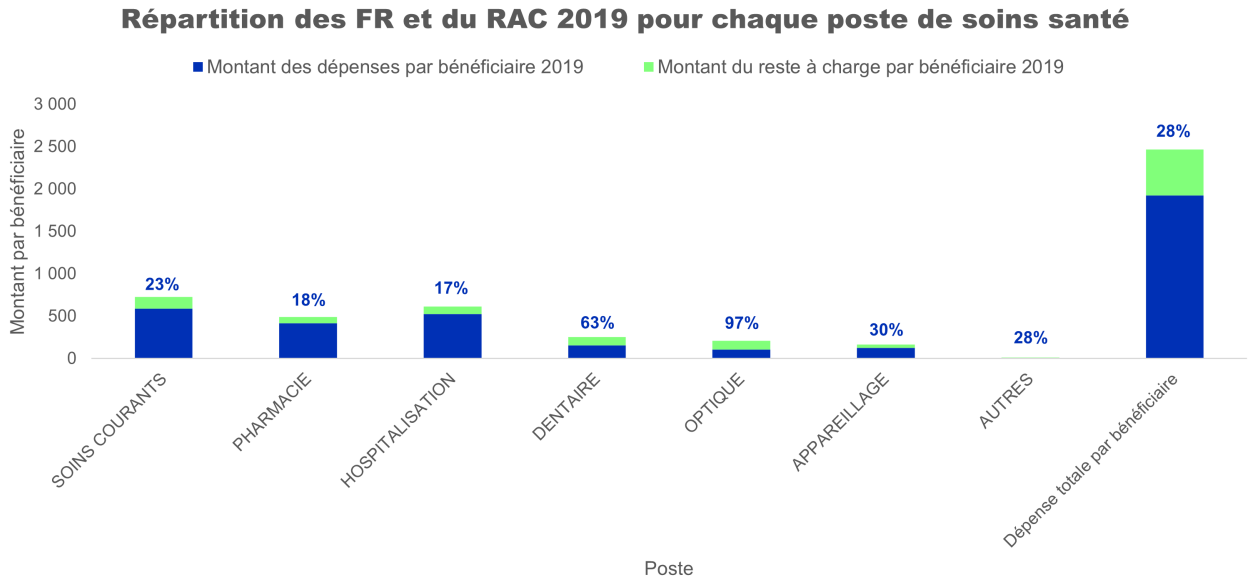


FIGURE 4.13 : Répartition des FR et du RAC pour chaque poste de soins santé et l'année 2019.

Remarque : le poste « Appareillage » comprend les remboursements des prothèses auditives mais aussi des autres appareillages. Nous ne retiendrons donc pas cet élément pour l'analyse suivante car il ne représente pas uniquement le reste à charge des prothèses auditives.

Dans le graphique 4.13, le pourcentage indiqué pour chaque poste représente le montant de reste à charge en pourcentage du montant des frais réels. Nous observons que les taux de restes à charge moyens sur le dentaire, l'optique et l'appareillage sont respectivement de 63%, 97% et 30% (les 3 taux les plus élevés). Ces constats mettent en évidence la nécessité de la réforme 100% santé. Effectivement, nous remarquons que pour la plupart des postes (postes pour lesquels le pourcentage est faible), les soins sont très bien remboursés par la Sécurité Sociale, sauf pour les postes « Dentaire » et « Optique » où le reste à charge pour l'assuré reste très important. De plus, si nous effectuons un zoom sur le poste dentaire, avec une distinction des trois paniers, nous remarquons sur le graphique 4.14 que le Remboursement Obligatoire de la Sécurité Sociale est très faible.

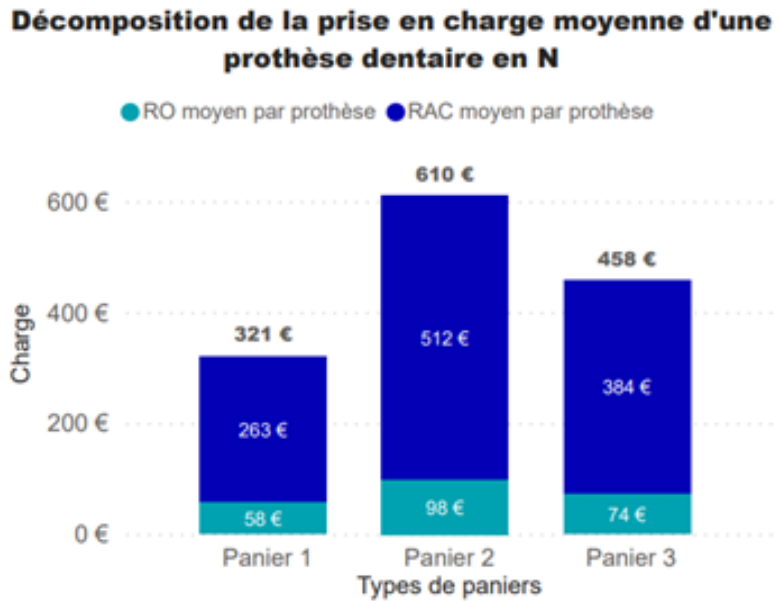


FIGURE 4.14 : Décomposition de la prise en charge moyenne d'une prothèse dentaire en 2019.

Le suivi technique Damir permet donc de mettre en avant et de comprendre l'objectif du projet de la réforme 100% santé. Il permettra pour les prochaines années de suivre l'impact de cette réforme, principalement au niveau de la consommation (les dépenses), mais aussi au niveau du recours aux soins. En effet, le suivi technique Damir permet de chiffrer le nombre de prothèses ou le nombre de soins effectués par bénéficiaire sur l'année, et donne ainsi une tendance du recours aux soins à l'échelle nationale. Le volet « 100% santé » du suivi technique Damir permet donc d'analyser l'évolution du recours aux soins sur deux années consécutives, et pour les trois postes concernés par la réforme 100% santé. Cette statistique pourra être comparée à celle du portefeuille client, afin de déterminer une éventuelle surconsommation sur ces postes par rapport à la tendance nationale (c.f. graphique 4.15).

Charge (montant RO) par bénéficiaire pour le poste "Prothèses dentaires"

	RO par bénéficiaire | Prothèses dentaires	2018	2019	Evolution
Nombre de prothèse moyen par bénéficiaire	Panier 1	0.00	0.07	0%
	Panier 2	0.00	0.04	0%
	Panier 3	0.20	0.12	-41%
	Total	0.20	0.23	13%
Charge moyenne par prothèse	Panier 1	- €	58.25 €	0%
	Panier 2	- €	97.83 €	0%
	Panier 3	83.34 €	74.06 €	-11%
	Total	83.34 €	73.31 €	-12%
Charge moyenne par bénéficiare	Panier 1	- €	4.04 €	0%
	Panier 2	- €	3.82 €	0%
	Panier 3	16.69 €	8.70 €	-48%
	Total	16.69 €	16.56 €	-1%
Nombre de bénéficiaires (hors enfant)		*47 205 k*	*44 862 k*	*-5%*

FIGURE 4.15 : Montant remboursé par le Régime Obligatoire pour le poste « Prothèses dentaires ».

Enfin, d'autres axes au niveau microscopique sont exploités comme l'obtention de résultats par catégorie d'âge. La consommation globale donne une tendance générale qui peut être totalement différente selon la catégorie d'âge. Cette analyse microscopique pourra mettre en évidence l'échantillon de population le plus consommant sur un certain poste, et pourra être comparée avec la tendance au niveau du portefeuille clients. Cela entrainera donc la détection des risques auxquels le client se soumet face aux répercussions de cette réforme. Par exemple, pour les prothèses auditives, le graphique 4.16 donne le nombre de prothèses « consommées » par bénéficiaire selon la tranche d'âge.

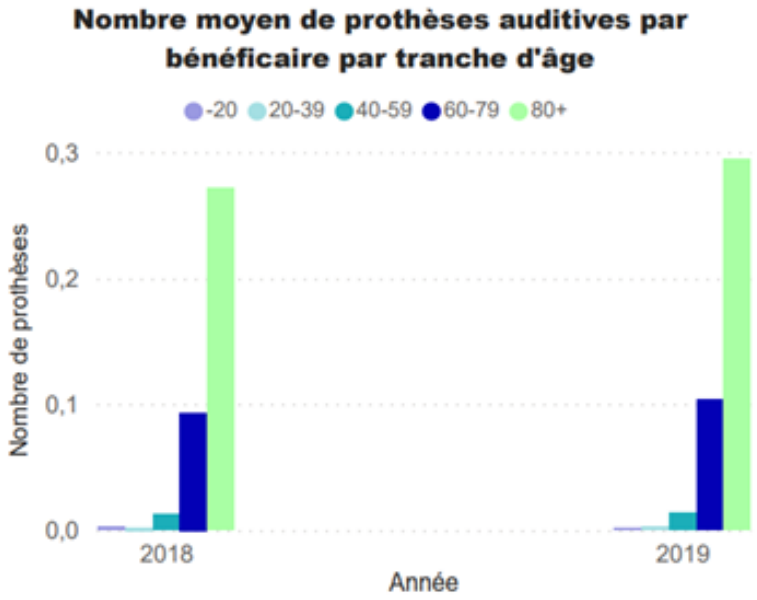


FIGURE 4.16 : Nombre moyen de prothèses auditives par bénéficiaire et par tranche d'âge.

Ce graphique est plutôt cohérent puisqu'une personne âgée a souvent recours à ce type d'équipement, car souvent dans le besoin. La population jeune a recours aux prothèses auditives non pas par confort mais par besoin. Elle concerne une très faible proportion de personnes.

Conclusion

Ce mémoire avait pour objectif d'intégrer la base Open Damir, disponible en Open Data, dans divers travaux actuariels liés à l'assurance santé, et plus particulièrement, ceux réalisés par le cabinet de conseil Actélior. Les deux outils enrichis de cette base de données sont :

- le suivi technique santé, sur le périmètre global puis sur les impacts de la réforme 100% santé ;
- l'outil de tarification santé.

Dans le cadre de la transformation digitale, cette base de données est une ressource riche d'informations permettant d'améliorer les outils existants. L'exploitation de ces ressources permettra de comparer le comportement en santé des adhérents d'un portefeuille client par rapport au comportement national en santé.

Pour cela, un des principaux objectifs de ce mémoire était le traitement de ces données volumineuses, par l'utilisation d'outils informatiques accessibles pour tous. Tous les traitements ont par conséquent été réalisés sur $\mathbb{P}ython$, et les modélisations GLM sur R. Les interfaces de développement, respectivement *Anaconda (Spyder)* et *Rstudio*, des langages $\mathbb{P}ython$ et R sont en téléchargement gratuit. Aujourd'hui, $\mathbb{P}ython$ et R sont prépondérant sur le marché du digital et fortement utilisés dans l'actuariat. De nombreuses fonctions pré implémentées sont disponibles pour l'exécution de méthodes de Machine Learning mais aussi pour la réalisation de différentes étapes des modélisations GLM. Un package « *Actuar* » est notamment disponible sur R, destiné aux actuaires. Ces travaux peuvent donc être aisément repris par d'autres utilisateurs n'ayant pas accès à des solutions performantes. N'oublions pas qu'une des plus grandes complexités de ces bases est la volumétrie allongeant les temps de traitement. Le forum Open Damir, disponible sur le site qui lui est dédié (Datagouv, 2019), permet de répondre aux différentes questions des utilisateurs de cette base de données. L'index, complémentaire aux bases Open Damir, fournit notamment des informations sur le contenu de ces dernières. Cependant, pour comprendre ces données, des analyses approfondies sont nécessaires. Les éléments de ce mémoire pourront aider les futurs utilisateurs de la base Open Damir à comprendre et à exploiter correctement cette dernière.

Le deuxième objectif de ce mémoire était la création de deux outils, similaires à ceux utilisés par Actélior, avec les bases de données Open Damir. D'un point de vue métier, le suivi technique apporte une réelle vision nationale de la consommation et des remboursements en santé. Il apportera d'autant plus d'informations sur les impacts de la réforme 100% santé pour les années à venir. Ce mémoire ne concernant uniquement que les années de remboursement et de survenance 2018 et 2019, la réforme était au tout début de son entrée en vigueur. A ce jour, la réforme 100% santé constitue une réelle inquiétude pour les complémentaires santé pour lesquelles les prestations payées s'envolent en dentaire et en audiologie. Il sera donc vraiment intéressant d'analyser les bases Open Damir 2020-2021 afin d'identifier les évolutions de comportement, de coûts moyens ou encore de nombres d'actes avec 2018

et 2019.
Concernant la tarification santé, le modèle de crédibilité, intégré dans le processus de tarification, apporte l'ajustement nécessaire. Les intervalles de confiance des coefficients GLM pourront notamment être utilisés de façon différente. En effet, cela permettra d'aider les clients dans la définition de leur appétence aux risques et de les accompagner dans la décision finale du tarif.

Certaines difficultés ont cependant été rencontrées pour la construction de la table de correspondance, utilisée lors du retraitement des données. En effet, peu d'informations sont disponibles sur les codes actes répertoriés dans l'index de la base Open Damir. Ces informations sont à collecter par le biais de données disponibles sur différents sites, par l'analyse quantitative des données présentes au sein de cette base, et enfin, par recherche sur le site Ameli. Cette table de correspondance a donc été finalisée dans le cadre de ce mémoire. Elle sera mise à jour chaque année en étudiant les nouveaux codes actes présents dans l'index. Les analyses présentées au sein de ce mémoire seront donc actualisées régulièrement.

Néanmoins, les applications réalisées dans le cadre de ce mémoire font face à certaines limites. En effet, le choix des lois s'ajustant au mieux aux données ne sont pas adaptées pour la modélisation de certains libellés brochures. Pour améliorer les résultats, d'autres lois, non étudiées dans le cadre de ce mémoire, pourraient être testées. Pour cela, nos travaux pourront être liés à ceux du mémoire précédemment réalisé au sein du cabinet d'Actélior consistant à automatiser la construction d'un GLM. Dans notre cas, cette automatisation sera utile pour le choix de la loi qui ajuste le mieux les données parmi un nombre conséquent de lois qui existent. Cela n'a pas été mis en place au sein de ce mémoire car le temps d'exécution de cette implémentation est très long pour un tel volume de données. Enfin, les travaux réalisés dans le cadre de ce mémoire ont été réalisés sur les bases de données des années de remboursements 2018 et 2019. Ceux-ci devront être appliqués avec précaution sur les bases de données Open Damir 2020 et 2021, compte tenu de la pandémie du Covid-19 qui a eu un effet non négligeable sur le système de santé et sur le recours aux soins des français.

Les différents travaux réalisés dans le cadre de ce mémoire seront mis à jour chaque année avec l'actualisation des données mises à disposition, les besoins des clients et les éventuelles évolutions réglementaires.

Enfin, maintenant que les bases ont été bien analysées sur la santé, un travail similaire reste à faire sur la prévoyance. Cela permettra très certainement au cabinet d'apporter des éléments complémentaires aux études et à la tarification du risque prévoyance.

Liste des symboles

ACP Analyse en Composante Principale

ACPR Autorité de Contrôle Prudentiel et de Résolution

ACS Aide la Complémentaire Santé

AIC Akaike Information Criterion (Critère d'Information d'Akaike)

ALD Affections Longues Durées

AM Alsace-Moselle

AMC Assurance Maladie Complémentaire

AMO Assurance Maladie Obligatoire

ANI Accord National Interprofessionnel

AT Accident de travail

BIC Bayesian Information Criterion (Critère d'Information Bayésien)

BR Base de Remboursement

BRSS Base de Remboursement de la Sécurité Sociale

CAF Caisse des Allocations Familiales

CARSAT Caisse d'Assurance Retraite et de la Santé Au Travail

CAS Contrat d'Accès aux Soins

CCAM Classification Commune des Actes Médicaux

CGSS Caisse Générale de la Sécurité Sociale

CMU Couverture Maladie Universelle

CMU-C Couverture Maladie Universelle Complémentaire

CNAM Caisse Nationale de l'Assurance Maladie

CNAMTS Caisse Nationale d'Assurance Maladie des Travailleurs Salariés

CNMSS Caisse Nationale Militaire de Sécurité Sociale

CPAM Caisse Primaire d'Assurance Maladie

CSBM Consommation de Soins et de Biens Médicaux

CSP Catégorie SocioProfessionnelle

CSS Complémentaire Santé Solidaire

DPTAM Dispositif de Pratique Tarifaire Maîtrisée

DREES Direction de la Recherche, des Études, de l'Évaluation et des Statistiques

DRSM Direction Régionale du Service Médical

DSS Direction de la Sécurité Sociale

EPCI Etablissement Public de Coopération Intercommunale

FIR Fond d'Intervention Régional

FR Frais réels

GLM Generalized Linear Model

HCAAM Haut Comissariat pour l'Avenir de l'Assurance Maladie

IA Intelligence Artificielle

IDC Intervalle De Confiance

iForest Isolation Forest

INSEE Institut National de la Statistique et des Etudes Economiques

IP Institutions de Prévoyance

k Milliers

LGBM Light Gradient Boosting Machine

M Millions

Md Milliards

ML Machine Learning

MLG Modèle Linéaire Généralisé

MP Maladies Professionnelles

MSA Mutualité Sociale Agricole

MSE Mean Squard Error (Erreur quadratique moyenne)

NAF Nomenclature d'Activités Française

OPTAM Option Pratique Tarifaire Maîtrisée

OPTAM-CO Option Pratique Tarifaire Maîtrisée Chirurgie et Obstétrique

PDF Portable Document Format

PIB Produit Intérieur Brut

PowerBI Power Business Intelligence

PPAP Provision pour Prestation A Payer

PS Professionnel de Santé

PSAP Provision pour Sinistre A Payer

PUMA Protection Universelle Maladie

QPV Quartiers Prioritaires de la politique de la Ville

RAC Reste A Charge

RATP Régie Autonome des Transports Parisiens

RGPD Réglement Général sur la Protection des Données

RMSE Root Mean Squard Error (Racine de l'erreur quadratique moyenne)

RO Régime Obligatoire

RSI Régime Social des Indépendants

RSS Remboursement de la Sécurité Sociale

S2 Solvabilité 2

SA Société Anonyme

SAM Société d'Assurance Mutuelle

SAS Statistical Analysis System

SCR Somme des Carrés des Résidus

SNCF Société Nationale des Chemins de fer Français

SNIIRAM Système National d'Information Inter-Régimes de l'Assurance Maladie

SS Sécurité Sociale

TAA Tarification A l'Activité

TM Ticket Modérateur

UGECAM Union de Gestion des Etablissements de Caisse d'Assurance Maladie

URSSAF Unions de Recouvrement des cotisations de la Sécurité Sociale et d'Allocations Familiales

VBA Visual Basic Application

Bibliographie

A. Saini (2021). Data Science Blogathon - Gradient Boosting Algorithm : a complete guide for beginners. url : https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/ (visité le 24/10/2021).

Ameli (2021). Site de l'Assurance Maladie Obligatoire. https://assurance-maladie.ameli.fr/. (Visité le 24/10/2021).

Annuaire du site de l'Assurance Maladie (2021). Annuaire des médecins signataires à l'OPTAM/OPTAM-CO. http://annuairesante.ameli.fr/. (Visité le 24/10/2021).

Assurance Maladie Obligatoire (juillet 2021). Liste des codes actes des remboursements à destination des professionnels de santé. https://www.ameli.fr/sites/default\/files/2021_composition-postes-depenses_cartographie_0.pdf. (Visité le 24/10/2021).

Bel, L., Daudin, J., Etienne, M., Lebarbier, E., Mary-Huard, T., Robin, S. et Vuillet, C. (2016). Le Modèle Linéaire et ses Extensions. http://moulon.inra.fr/modelstat/supports/ModeleLineaireEt_Extensions-compressed.pdf.

Chery, C. (2015). Construction d'un outil de tarification de contrats complémentaire santé. Mémoire d'actuariat. Lyon : ISFA.

Chevalier P.A. (23 janvier 2015). Github de la base Open Damir. https://github.com/SGMAP-AGD/DAMIR/wiki/OpenDamir. (Visité le 24/10/2021).

Datagouv (2019). Site de téléchargement des bases Open Damir. https://www.data.gouv.fr/fr/datasets/open-damir-base-complete-sur-les-depenses-dassurance-maladie-inter-regimes/. (Visité le 24/10/2021).

Datagouv (2021). Site des données publiques de l'Etat français. https://www.data.gouv.fr/fr/. (Visité le 24/10/2021).

Delignette-Muller M-L. et Dutang C. (2014). fitdistrplus: An R Package for Fitting Distributions. https://cran.r-project.org/web/packages/fitdistrplus/vignettes/-paper2JSS.pdf. (Visité le 24/10/2021).

Direction de la Sécurité Sociale (2018). Rapport - Les chiffres clés de la Sécurité Sociale 2018. https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2019/CHIFFRES%20CLES%202019.pdf. (Visité le 24/10/2021).

Direction de la Sécurité Sociale (2019). Rapport - Les chiffres clés de la Sécurité Sociale 2019. https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2020/CHIFFRES%20CLES%202020%20ED2019.pdf. (Visité le 24/10/2021).

DREES (2019a). Rapport - La complémentaire santé - Acteurs, bénéficiaires, garanties. https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-10/cs2019.pdf. (Visité le 24/10/2021).

DREES (2019b). Rapport - Les dépenses de santé en 2019 - Résultats des comptes de la santé). https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-04/Les%20d%C3%A9penses%20de%20sant%C3%A9%20en%202019%20-%20R%C3%A9sultats%20des%20comptes%20de%20la%20sant%C3%A9%20-%20%C3%89dition%202020.pdf. (Visité le 24/10/2021).

Fun Mooc - Grenoble Alpes (2021). Explications du principe du test V de Cramer. https://lms.fun-mooc.fr/asset-v1:grenoblealpes+92001+session01+type@asset+block/mod6-cap2.pdf. (Visité le 24/10/2021).

HCAAM (2021). Rapport du HCAAM sur la régulation du système de santé. https://www.securite-sociale.fr/home/hcaam/zone-main-content/rapports-et-avis-1/rapport-du-hcaam-sur-la-regulati.html. (Visité le 24/10/2021).

INSEE (2021). Estimation de la population au 1^{er} janvier 2021, INSEE. https://www.insee.fr/fr/statistiques/1893198. (Visité le 24/10/2021).

J. Brownlee (2021). Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost. https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/. (Visité le 24/10/2021).

Lamon, C. (2019). Modélisation et analyse de comportements clients en assurance dommage - application au changement de véhicules et à la résiliation de contrats. Mémoire d'actuariat. Paris : Université Paris-Dauphine.

Lazic, S. (2020). Cours santé/prévoyance - Partie Santé. Suppport de cours. Paris : Université Paris-Dauphine.

Lemakistatheux (2013). Présentation des coefficients de Pearson et de Spearman. url : https://lemakistatheux.wordpress.com/2013/05/22/le-coefficient-correlation-et-le-test-de-spearman/ (visité le 24/10/2021).

M. Clarke (2021). Practical Data Science - Isolation Forest. https://practicaldatascience.co.uk/machine-learning/how-to-use-the-isolation-forest-model-for-outlier-detection. (Visité le 24/10/2021).

M. Friendly (2000). « *goodfit* » R function - vcd: Visualizing Categorical Data. R package version 1.4-9. https://www.rdocumentation.org/packages/vcd/versions/1.4-8/topics/goodfit. SAS Institute, Cary, NC. (Visité le 24/10/2021).

Meyer D., Zeileis A. et Hornik K. (2021). Discrete Data Analysis with R. vcd: Visualizing Categorical Data. R package version 1.4-9.

Microsoft (2014). Documentation Power BI Desktop. https://docs.microsoft.com/fr-fr/power-bi/fundamentals/power-bi-overview. (Visité le 24/10/2021).

Ndibi Otakana, A. (2017). Calibration des chocs primes et réserves en prévoyance - Partie étudiée : Résidus d'Anscombe. Mémoire. Lyon : ISFA, Université Claude Bernard Lyon 1.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. url : https://www.R-project.org/.

Suru A. (2020). Assurance IARD - Les dessous d'un secteur qui vous protège.

Vernin-Biancale, J. (2020). Mémoire d'actuariat - Création d'un outil permettant d'estimer les impacts de la réforme 100% santé à partir de l'historique d'un portefeuille. Mémoire d'actuariat. Lyon : ISFA.

Annexe A

Résultats des traitements et tableaux descriptifs des variables

A.1 Exemples de traitements réalisés pour observer la répartition des codes actes selon le type d'assurance

TABLE A.1 : Résultats de la répartition de 20 codes actes selon le type d'assurance.

	ASU_NAT										
PRS_NAT	10	11	12	22	30	40	50	70	80	90	99
0											
1096											
1097											
1098	98.04%				0.53%	1.43%					
1099	98.99%				0.54%	0.47%					
1100	98.83%				1.17%						
1101	98.65%			0.01%	0.74%	0.60%					
1102	98.46%			0.02%	0.48%	1.04%					
1103	96.93%			0.01%	1.16%	1.89%					
1104	35.45%				64.55%						
1105	65.67%				34.33%						
1106	90.54%				9.23%	0.23%					
1107	96.11%				3.87%	0.02%					
1108	99.38%				0.57%	0.06%					
1109	93.30%			0.02%	1.58%	5.10%				0.00%	
1110	93.01%			0.07%	1.82%	5.04%				0.06%	
1111	92.07%			0.03%	6.81%	0.74%				0.35%	
1112	94.29%			0.02%	3.68%	2.00%				0.01%	
1113	97.89%			0.09%	0.51%	1.50%					

A.2 Tableau récapitulatif des taux d'anomalies au sein de chaque base de données mensuelle Open Damir

TABLE A.2 : Taux d'anomalies des 24 bases de données Open Damir à l'étude.

Nom de la base	Taux d'anomalies	Nombre de lignes après nettoyage de la base
A2018_01	21.49%	5 855 551
A2018_02	20.73%	9 166 973
A2018_03	20.38%	10 704 348
A2018_04	20.35%	10 923 306
A2018_05	20.37%	11 208 325
A2018_06	20.19%	11 595 641
A2018_07	20.25%	11 941 777
A2018_08	19.91%	10 951 375
A2018_09	19.91%	11 131 558
A2018_10	20.10%	12 722 842
A2018_11	20.08%	11 964 314
A2018_12	19.75%	11 854 684
A2019_01	22.25%	13 051 996
A2019_02	23.10%	12 387 987
A2019_03	23.07%	12 868 185
A2019_04	23.02%	13 123 030
A2019_05	22.64%	12 979 926
A2019_06	22.66%	12 735 062
A2019_07	22.56%	13 826 925
A2019_08	22.56%	12 238 297
A2019_09	22.16%	12 851 501
A2019_10	22.18%	14 137 082
A2019_11	22.36%	12 851 857
A2019_12	22.17%	13 709 313
Moyenne pondérée	**21.50%**	**11 949 244**

A.3 Retraitement de la base de données de la démographie française, extraite de l'INSEE

TABLE A.3 : Visualisation de la base de l'INSEE sur la démographie française retraitée - Hommes.

Année 2018

Régions_DAMIR	Hommes							
	0 à 19 ans	20 à 29 ans	30 à 39 ans	40 à 49 ans	50 à 59 ans	60 à 69 ans	70 à 79 ans	80 ans et +
84	884 235	401 392	436 990	459 052	452 724	386 799	266 171	154 742
27	290 038	127 747	138 814	152 967	163 339	158 310	107 085	63 779
53	354 471	153 156	167 719	189 924	193 816	184 520	118 552	71 119
24	273 452	113 357	126 708	142 718	148 555	140 795	93 325	59 523
44	580 275	281 837	296 664	310 302	331 531	297 192	180 746	103 727
32	699 956	311 322	328 594	336 547	333 191	292 246	168 495	91 133
11	1 418 265	716 356	743 594	729 345	653 425	482 173	301 596	167 153
28	359 333	155 429	169 009	180 418	190 143	180 874	112 811	67 142
75	589 161	264 657	293 880	332 591	345 538	340 131	231 788	145 611
76	602 138	283 735	299 018	324 233	337 153	316 888	222 419	136 008
52	426 045	179 622	199 774	216 809	210 760	194 786	124 856	78 973
93	547 416	243 728	273 624	293 327	306 804	279 688	210 919	125 476
5	309 512	100 362	98 177	110 468	119 798	86 750	46 829	23 064

TABLE A.4 : Visualisation de la base de l'INSEE sur la démographie française retraitée - Femmes.

Année 2018

Régions_DAMIR	Femmes							
	0 à 19 ans	20 à 29 ans	30 à 39 ans	40 à 49 ans	50 à 59 ans	60 à 69 ans	70 à 79 ans	80 ans et +
84	842 540	396 210	453 060	461 905	465 263	427 324	313 945	278 055
27	273 548	120 417	141 966	154 757	168 708	170 821	123 663	114 148
53	336 106	144 144	169 621	189 227	197 677	201 227	144 619	135 108
24	260 827	109 995	133 586	144 538	153 191	153 058	109 422	102 448
44	550 556	270 798	303 563	313 322	340 317	323 461	210 914	192 337
32	668 023	306 274	338 541	341 312	343 988	329 586	211 562	185 736
11	1 367 091	757 203	804 490	744 287	697 246	551 228	365 556	304 216
28	340 616	153 079	173 991	184 304	197 452	199 577	136 848	126 905
75	561 300	261 788	307 972	342 534	365 889	377 006	273 867	255 743
76	573 290	279 225	313 654	334 691	355 902	348 980	260 675	230 672
52	404 364	174 885	203 722	216 554	216 387	213 578	149 193	140 406
93	518 604	243 806	294 933	310 278	333 712	317 011	254 752	214 038
5	306 947	113 799	126 787	131 503	133 364	97 180	56 813	39 453

TABLE A.5 : Visualisation de la base de l'INSEE finale jointe avec la base Open Damir.

SOI_ANN	SEXE_BENEF	AGE_BENEF	REGION_BENEF	EXPO_BENEF
2018	1	0	84	933901.0794
2018	1	0	27	309283.1446
2018	1	0	53	375999.3464
2018	1	0	24	290748.1755
2018	1	0	44	618333.4262
2018	1	0	32	745737.4706
2018	1	0	11	1496862.461
2018	1	0	28	382704.8715
2018	1	0	75	623927.7286
2018	1	0	76	635305.4184
2018	1	0	52	451186.1754
2018	1	0	93	578964.6636
2018	1	0	5	329359.6673
2018	1	20	84	425920.532

A.4 Tableau récapitulatif du nombre de lignes pour chaque base mensuelle après les retraitements détaillés en section 2.2

TABLE A.6 : Nombre de lignes pour chaque base après les retraitements détaillés en section 2.2.

Nom de la base Damir	Base Damir	Base nettoyée	Base agrégée en mensuel	Base agrégée en annuelle
A2018_01	34 003 028	5 855 551	1 247 283	1 183 415
A2018_02	31 064 879	9 166 973	2 172 089	1 467 284
A2018_03	33 246 390	10 704 348	2 664 467	1 590 820
A2018_04	32 219 249	10 923 306	2 810 301	1 581 970
A2018_05	31 683 059	11 208 325	2 910 118	1 575 629
A2018_06	32 786 189	11 595 641	3 132 225	1 639 169
A2018_07	32 937 644	11 941 777	3 181 591	1 656 625
A2018_08	30 160 185	10 951 375	2 959 818	1 531 934
A2018_09	30 378 977	11 131 558	3 050 337	1 566 971
A2018_10	34 304 782	12 722 842	3 437 955	1 739 311
A2018_11	31 975 235	11 964 314	3 261 544	1 662 662
A2018_12	31 358 125	11 854 684	3 219 864	1 606 654
A2019_01	35 883 180	13 051 996	3 387 600	2 297 183
A2019_02	33 861 814	12 387 987	3 187 165	1 934 686
A2019_03	35 348 452	12 868 185	3 314 576	1 870 944
A2019_04	35 556 250	13 123 030	3 356 761	1 841 006
A2019_05	34 990 081	12 979 926	3 389 177	1 799 555
A2019_06	34 246 739	12 735 062	3 319 475	1 736 233
A2019_07	36 909 000	13 826 925	3 602 243	1 848 025
A2019_08	32 870 015	12 238 297	3 160 059	1 621 080
A2019_09	34 163 250	12 851 501	3 407 746	1 710 754
A2019_10	37 429 359	14 137 082	3 733 801	1 858 978
A2019_11	33 794 832	12 851 857	3 404 452	1 717 058
A2019_12	35 563 651	13 709 313	3 603 994	1 745 239
TOTAL	806 734 365	286 781 855	74 914 641	40 783 185
Nombre de lignes de la base finale obtenue	10 040 857			

A.5 Description des variables disponibles dans la base Open Damir

TABLE A.7 : Description des variables disponibles dans la base Open Damir.

Variable	Libellé	Variable	Libellé
PERIODE DE TRAITEMENT		**EXECUTANT**	
FLX_ANN_MOI	Année et mois du remboursement	EXE_INS_REG	Région du PS exécutant (valables à partir de 2015 pour les tables commençant par A)
PRESTATION		PSE_STJ_SNDS	Statut juridique du PS exécutant
PRS_NAT	Nature de la prestation	MFT_COD	Mode de fixation des tarifs de l'établissement de l'exécutant
ASU_NAT	Nature de l'assurance	ETE_REG_COD	Région d'implantation de l'établissement de l'exécutant (valables à partir de 2015 pour les tables commençant par A)
ATT_NAT	Nature de l'Accident du Travail	ETE_TYP_SNDS	Type d'éttablissement de l'exécutant
CPT_ENV_TYP	Type d'enveloppe	ETE_CAT_SNDS	Catégorie de l'établissement de l'exécutant
CPL_COD	Complément d'acte	DDP_SPE_COD	Discipline de prestation de l'établissement de l'exécutant
EXO_MTF	Motif d'exonération du Ticket Modérateur	MDT_TYP_COD	Mode de traitement dans l'établissement de l'exécutant
PRS_REM_TAU	Taux de remboursement	**PRESCRIPTEUR**	
PRS_PPU_SEC	Code secteur privé/public	PSP_ACT_CAT	Catégorie du prescripteur
PRS_FJH_TYP	Type de prise en charge du Forfait Journalier	PSP_SPE_SNDS	Spécialité Médicale du PS prescripteur
ETE_IND_TAA	Indicateur TAA privé/public	PSP_ACT_SNDS	Nature d'Activité PS Prescripteur
PRS_PDS_QCP	Code qualificatif du parcours de soins (sortie)	PRE_INS_REG	Région du PS Prescripteur (valables à partir de 2015 pour les tables commençant par A)
DRG_AFF_NAT	Nature du destinataire de règlement affiné	PSP_STJ_SNDS	Statut juridique du PS prescripteur
PRS_REM_TYP	Type de remboursement	ETP_REG_COD	Région d'implantation de l'établissement du prescripteur (valables à partir de 2015 pour les tables commençant par A)
ORGANISME		ETP_CAT_SNDS	Catégorie de l'établissement du prescripteur
ORG_CLE_REG	Région de l'organisme de liquidation (valables à partir de 2015 pour les tables commençant par A)	**TOP PS5**	
PERIODE		TOP_PS5_TRG	Prestations du périmètre hors CMU C et CMU C
SOI_ANN	Année du soin	**INDICATEURS**	
SOI_MOI	Mois du soin	FLT_ACT_COG	Coefficient global de la prestation préfiltrée
BENEFICIAIRE		FLT_ACT_NBR	Dénombrement de la prestation préfiltrée
BEN_SEX_COD	Sexe du bénéficiaire	FLT_ACT_QTE	Quantité de la prestation préfiltrée
AGE_BEN_SNDS	Tranche d'âge du bénéficiaire au moment des soins	FLT_DEP_MNT	Montant du dépassement de la prestation préfiltrée
BEN_QLT_COD	Qualité du bénéficiaire	FLT_PAI_MNT	Montant de la dépense de la prestation préfiltrée
BEN_RES_REG	Région de résidence du bénéficiaire (valables à partir de 2015 pour les tables commençant par A)	FLT_REM_MNT	Montant versé/remboursé préfiltré
MTM_NAT	Modulation du Ticket Modérateur	PRS_ACT_COG	Coefficient global
BEN_CMU_TOP	Bénéficiaire de la CMU-C	PRS_ACT_NBR	Dénombrement
EXECUTANT		PRS_ACT_QTE	Quantité
	**PS : professionnel de santé*	PRS_DEP_MNT	Montant du dépassement
PSE_ACT_CAT	Catégorie de l'exécutant	PRS_PAI_MNT	Montant de la dépense
PSE_SPE_SNDS	Spécialité médicale du PS exécutant	PRS_REM_MNT	Montant versé/remboursé
PSE_ACT_SNDS	Nature de l'activité du PS exécutant	PRS_REM_BSE	Base de remboursement

A.6 Tableaux récapitulatifs de l'indépendance entre la fréquence et le coût moyen pour chaque libellé brochure.

TABLE A.8 : Résultats de l'indépendance entre la fréquence et le coût moyen pour chaque libellé brochure (1).

FAMILLE	LIBELLE	Coefficient de corrélation - Pearson	Indépendance - Pearson	Coefficient de corrélation - Spearman	Indépendance - Spearman
SOINS COURANTS	Consultations de généralistes	-0.1902	Indépendance	-0.1176	Indépendance
SOINS COURANTS	Consultations de spécialistes	0.0202	Indépendance	0.0708	Indépendance
SOINS COURANTS	Consultations de psychiatres	0.0733	Indépendance	0.1108	Indépendance
SOINS COURANTS	Actes techniques médicaux	0.0388	Indépendance	0.1000	Indépendance
SOINS COURANTS	Auxiliaires médicaux	0.0686	Indépendance	0.4946	Corrélation positive
SOINS COURANTS	Analyses médicales	0.0254	Indépendance	0.1039	Indépendance
HOSPITALISATION	Actes d'anesthésie	0.1125	Indépendance	0.2500	Corrélation positive
HOSPITALISATION	Actes de chirurgie	0.0084	Indépendance	0.0094	Indépendance
HOSPITALISATION	Actes d'échographie	0.0822	Indépendance	0.1277	Indépendance
HOSPITALISATION	Actes d'imagerie	0.0378	Indépendance	0.0747	Indépendance
HOSPITALISATION	Actes d'obstétrique	0.1686	Indépendance	0.2336	Indépendance
HOSPITALISATION	Chambre particulière maternité	Non disponible	Non disponible	Non disponible	Non disponible
HOSPITALISATION	Chambre particulière médicale et chirurgicale	-0.1338	Indépendance	-0.2162	Indépendance
HOSPITALISATION	Frais de séjour médical et chirurgical	-0.0104	Indépendance	0.0900	Indépendance
HOSPITALISATION	Forfait hospitalier	0.0663	Indépendance	0.6839	Corrélation positive
HOSPITALISATION	Frais de transport	0.0270	Indépendance	0.0470	Indépendance
HOSPITALISATION	Frais d'accompagnant	Non disponible	Non disponible	Non disponible	Non disponible
HOSPITALISATION	Autres	0.0976	Indépendance	0.0408	Indépendance

TABLE A.9 : Résultats de l'indépendance entre la fréquence et le coût moyen pour chaque libellé brochure (2).

FAMILLE	LIBELLE	Coefficient de corrélation - Pearson	Indépendance - Pearson	Coefficient de corrélation - Spearman	Indépendance - Spearman
DENTAIRE	Prothèse dentaire - Panier 1	0.0544	Indépendance	0.1645	Indépendance
DENTAIRE	Prothèse dentaire - Panier 2	0.0399	Indépendance	0.1020	Indépendance
DENTAIRE	Prothèse dentaire - Panier 3	0.0750	Indépendance	0.1622	Indépendance
DENTAIRE	Parodontologie	0.0752	Indépendance	0.2685	Corrélation positive
DENTAIRE	Implant dentaire	0.0340	Indépendance	0.0247	Indépendance
DENTAIRE	Soins dentaires	0.0210	Indépendance	0.4606	Corrélation positive
DENTAIRE	Orthodontie acceptée / refusée	0.1528	Indépendance	0.1937	Indépendance
OPTIQUE	Monture	0.1278	Indépendance	0.2021	Indépendance
OPTIQUE	Verres	0.0975	Indépendance	0.1908	Indépendance
OPTIQUE	Lentilles acceptées	0.0873	Indépendance	0.1540	Indépendance
OPTIQUE	Chirurgie réfractive	Non disponible	Non disponible	Non disponible	Non disponible
OPTIQUE	Autres	Non disponible	Non disponible	Non disponible	Non disponible
PHARMACIE	Pharmacie 100%	0.0209	Indépendance	0.2228	Indépendance
PHARMACIE	Pharmacie 15%	0.1844	Indépendance	0.4315	Corrélation positive
PHARMACIE	Pharmacie 30%	0.1124	Indépendance	0.2898	Corrélation positive
PHARMACIE	Pharmacie 65%	0.0209	Indépendance	0.2099	Indépendance
PHARMACIE	Vaccins anti-grippe	0.1168	Indépendance	0.1450	Indépendance
PHARMACIE	Autres	0.0040	Indépendance	-0.0002	Indépendance
APPAREILLAGE	Accessoires	0.0244	Indépendance	0.2098	Indépendance
APPAREILLAGE	Petit appareillage	0.0193	Indépendance	0.1745	Indépendance
APPAREILLAGE	Grand appareillage	0.1164	Indépendance	0.2774	Corrélation positive
APPAREILLAGE	Prothèse auditive I	0.0492	Indépendance	-0.0092	Indépendance
APPAREILLAGE	Prothèse auditive II	0.0178	Indépendance	0.0553	Indépendance

TABLE A.10 : Résultats de l'indépendance entre la fréquence et le coût moyen pour chaque libellé brochure (3).

FAMILLE	LIBELLE	Coefficient de corrélation - Pearson	Indépendance - Pearson	Coefficient de corrélation - Spearman	Indépendance - Spearman
AUTRES	Cure thermale - Hébergement	Non disponible	Non disponible	Non disponible	Non disponible
AUTRES	Cure thermale - Transport	0.0723	Indépendance	0.2178	Indépendance
AUTRES	Cure thermale - Soins	-0.0298	Indépendance	-0.2581	Indépendance
AUTRES	Médecine douce	0.0453	Indépendance	-0.0531	Indépendance
AUTRES	Dépistage	Non disponible	Non disponible	Non disponible	Non disponible
AUTRES	Contraception	Non disponible	Non disponible	Non disponible	Non disponible
AUTRES	Arrêt du tabac	0.0179	Indépendance	0.0560	Indépendance
AUTRES	Soins à l'étranger	-0.1660	Indépendance	-0.1016	Indépendance
AUTRES	Prime de naissance ou d'adoption	Non disponible	Non disponible	Non disponible	Non disponible
AUTRES	Maternité	Non disponible	Non disponible	Non disponible	Non disponible
AUTRES	Autres vaccins	Non disponible	Non disponible	Non disponible	Non disponible

A.7 Tableaux récapitulatifs des lois de la dépense moyenne et de la quantité d'actes pour chaque libellé brochure.

TABLE A.11 : Lois de la dépense moyenne et de la quantité d'actes pour chaque libellé brochure (1).

Famille	Libellé	Loi de la dépense moyenne	Loi de la quantité d'actes
SOINS COURANTS	Consultations de généralistes	Gamma	Binomiale négative
SOINS COURANTS	Consultations de spécialistes	Weibull	Binomiale négative
SOINS COURANTS	Consultations de psychiatres	LogNormale	Binomiale négative
SOINS COURANTS	Actes techniques médicaux	LogNormale	Binomiale négative
SOINS COURANTS	Auxiliaires médicaux	Weibull	Binomiale négative
SOINS COURANTS	Analyses médicales	Weibull	Binomiale négative
HOSPITALISATION	Actes d'anesthésie	LogNormale	Binomiale négative
HOSPITALISATION	Actes de chirurgie	LogNormale	Binomiale négative
HOSPITALISATION	Actes d'échographie	Gamma	Binomiale négative
HOSPITALISATION	Actes d'imagerie	Gamma	Binomiale négative
HOSPITALISATION	Actes d'obstétrique	LogNormale	Binomiale négative
HOSPITALISATION	Chambre particulière maternité	Non disponible	Non disponible
HOSPITALISATION	Chambre particulière médicale et chirurgicale	Gamma	Binomiale négative
HOSPITALISATION	Frais de séjour médical et chirurgical	LogNormale	Binomiale négative
HOSPITALISATION	Forfait hospitalier	LogNormale	Binomiale négative
HOSPITALISATION	Frais de transport	Weibull	Binomiale négative
HOSPITALISATION	Frais d'accompagnant	Non disponible	Non disponible
HOSPITALISATION	Autres	Weibull	Binomiale négative

TABLE A.12 : Lois de la dépense moyenne et de la quantité d'actes pour chaque libellé brochure (2).

Famille	Libellé	Loi de la dépense moyenne	Loi de la quantité d'actes
DENTAIRE	Prothèse dentaire - Panier 1	Weibull	Binomiale négative
DENTAIRE	Prothèse dentaire - Panier 2	Weibull	Binomiale négative
DENTAIRE	Prothèse dentaire - Panier 3	Weibull	Binomiale négative
DENTAIRE	Parodontologie	LogNormale	Binomiale négative
DENTAIRE	Implant dentaire	Weibull	Binomiale négative
DENTAIRE	Soins dentaires	LogNormale	Binomiale négative
DENTAIRE	Orthodontie acceptée / refusée	LogNormale	Binomiale négative
OPTIQUE	Monture	Weibull	Binomiale négative
OPTIQUE	Verres	Weibull	Binomiale négative
OPTIQUE	Lentilles acceptées	Weibull	Binomiale négative
OPTIQUE	Chirurgie réfractive	Non disponible	Non disponible
OPTIQUE	Autres	Non disponible	Non disponible
PHARMACIE	Pharmacie 100%	LogNormale	Binomiale négative
PHARMACIE	Pharmacie 15%	Gamma	Binomiale négative
PHARMACIE	Pharmacie 30%	LogNormale	Binomiale négative
PHARMACIE	Pharmacie 65%	LogNormale	Binomiale négative
PHARMACIE	Vaccins anti-grippe	Weibull	Binomiale négative
PHARMACIE	Autres	LogNormale	Binomiale négative
APPAREILLAGE	Accessoires	LogNormale	Binomiale négative
APPAREILLAGE	Petit appareillage	LogNormale	Binomiale négative
APPAREILLAGE	Grand appareillage	Gamma	Binomiale négative
APPAREILLAGE	Prothèse auditive I	Weibull	Binomiale négative
APPAREILLAGE	Prothèse auditive II	Weibull	Binomiale négative

TABLE A.13 : Lois de la dépense moyenne et de la quantité d'actes pour chaque libellé brochure (3).

Famille	Libellé	Loi de la dépense moyenne	Loi de la quantité d'actes
AUTRES	Cure thermale - Hébergement	Non disponible	Non disponible
AUTRES	Cure thermale - Transport	Weibull	Binomiale négative
AUTRES	Cure thermale - Soins	Gamma	Binomiale négative
AUTRES	Médecine douce	Gamma	Binomiale négative
AUTRES	Dépistage	Non disponible	Non disponible
AUTRES	Contraception	Non disponible	Non disponible
AUTRES	Arrêt du tabac	Weibull	Binomiale négative
AUTRES	Soins à l'étranger	Weibull	Poisson
AUTRES	Prime de naissance ou d'adoption	Non disponible	Non disponible
AUTRES	Maternité	Non disponible	Non disponible
AUTRES	Autres vaccins	Non disponible	Non disponible

A.8 Intervalles de confiance des coefficients GLM pour la modélisation de la dépense moyenne et la fréquence des actes d'anesthésie

TABLE A.14 : Intervalles de confiance des coefficients GLM pour la modélisation de la dépense moyenne des actes d'anesthésie.

Variables	Borne inférieure	Borne supérieure	Longueur de l'intervalle
(Intercept)	4.8548	5.5162	0.6614
I(AGE_BENEF)	-0.2255	-0.0904	0.1352
I(AGE_BENEF^2)	0.0063	0.0158	0.0095
I(AGE_BENEF^3)	-0.0005	-0.0002	0.0003
I(AGE_BENEF^4)	0.0000	0.0000	0.0000
I(AGE_BENEF^5)	0.0000	0.0000	0.0000
I(AGE_BENEF^6)	0.0000	0.0000	0.0000
REGION_BENEF11	0.0052	0.0330	0.0279
REGION_BENEF2744	0.0610	0.0833	0.0223
REGION_BENEF5	-0.0480	-0.0055	0.0425
REGION_BENEF757693	0.0176	0.0370	0.0195
REGION_BENEF84	0.1016	0.1294	0.0278
SOI_ANN2019	0.0757	0.0910	0.0153

TABLE A.15 : Intervalles de confiance des coefficients GLM pour la modélisation de la quantité des actes d'anesthésie.

Variables	Borne inférieure	Borne supérieure	Longueur de l'intervalle
(Intercept)	-13.8017	-11.9302	1.8715
I(AGE_BENEF)	0.1964	0.5785	0.3821
I(AGE_BENEF^2)	-0.0332	-0.0063	0.0268
I(AGE_BENEF^3)	0.0001	0.0010	0.0009
I(AGE_BENEF^4)	0.0000	0.0000	0.0000
I(AGE_BENEF^5)	0.0000	0.0000	0.0000
I(AGE_BENEF^6)	0.0000	0.0000	0.0000
REGION_BENEF5	0.9640	1.0968	0.1329
REGION_BENEF11	-0.7595	-0.6628	0.0966
REGION_BENEF24	0.6213	0.7321	0.1109
REGION_BENEF27	0.4792	0.5815	0.1022
REGION_BENEF28	0.4164	0.5268	0.1104
REGION_BENEF32	-0.1053	-0.0039	0.1014
REGION_BENEF52	0.3696	0.4798	0.1102
REGION_BENEF53	0.2821	0.3959	0.1138
REGION_BENEF75	-0.2179	-0.1190	0.0989
REGION_BENEF76	-0.1751	-0.0787	0.0964
REGION_BENEF84	-0.5026	-0.4054	0.0972
REGION_BENEF93	-0.1271	-0.0298	0.0973
SOI_ANN2019	0.1053	0.1488	0.0436

Annexe B

Définition et démonstration des formules utilisées dans le mémoire

B.1 Calculs de la prime pure dans le cadre de la modélisation « Coût moyen × Fréquence »

Soit S la variable aléatoire correspondante à la charge totale des prestations de l'adhérent sur une période donnée et π la prime (ou cotisation) pure, déterministe. La charge totale de prestations de l'adhérent est définie par l'équation (3.1)

$$S = \sum_{i=1}^{N} X_i,$$

où N est la variable aléatoire à valeurs entières et strictement positives représentant le nombre d'actes santé sur la période considérée et $(X_i)_{i\in\mathbb{N}}$ la suite des prestations individuelles. Les X_i sont positifs et à valeurs réelles. Cette définition de la charge totale des prestations est valide avec l'hypothèse que les X_i sont i.i.d (indépendants et identiquement distribués).

L'estimation de la charge totale des prestations et donc de la prime pure se caractérise donc par la formule $\pi = \mathbb{E}(S)$. Déterminons l'expression de la prime pure dans le cadre de la modélisation « Coût moyen × Fréquence »

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S|N)).$$

Pour $N = n$,

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

Par linéarité de l'espérance conditionnelle, l'égalité devient

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

Selon l'hypothèse d'indépendance entre les X_i et le nombre de prestations N aléatoire, l'égalité devient

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

Or, par hypothèse, les X_i sont i.i.d, ce qui donne

$$\mathbb{E}(S|N=n) = \sum_{i=1}^{n} \mathbb{E}(X) = n\mathbb{E}(X), \tag{B.1}$$

soit

$$\mathbb{E}(S|N) = N\mathbb{E}(X). \tag{B.2}$$

Finalement, d'après les équations (B.1) et (B.2), la prime pure se réécrit selon l'équation (B.3)

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S|N)) = \mathbb{E}(N\mathbb{E}(X))$$
$$\mathbb{E}(S) = \mathbb{E}(N) \times \mathbb{E}(X), \tag{B.3}$$

avec :

- $\mathbb{E}(N)$ qui représente la fréquence des prestations ;
- $\mathbb{E}(X)$ qui représente l'estimation du coût moyen de la prestations santé.

Concernant la variance de la charge totale des prestations S de l'adhérent, elle s'obtient par la formule de décomposition de la variance, définie par

$$\mathbb{V}(X) = \mathbb{E}(\mathbb{V}(X|G) + \mathbb{V}(\mathbb{E}(X|G),$$

avec X et G deux variables aléatoires sur un même espace de probabilité, et $\mathbb{V}(X) < \infty$. La variance de la variable aléatoire S est donc définie par

$$\mathbb{V}(S) = \mathbb{E}(S^2) - \mathbb{E}(S)^2.$$

Or,

$$\begin{aligned}\mathbb{E}(S^2) &= \mathbb{E}(\mathbb{E}(S^2|N)) \\ &= \mathbb{E}[\mathbb{E}(\sum_{i=1}^{N} {X_i}^2|N) + \mathbb{E}(\sum_{i=1}^{N}\sum_{j\neq i} X_i X_j|N)]\end{aligned}$$

D'après les traitements réalisés à l'équation (B.2) :

$$\begin{aligned}\mathbb{E}(S^2) &= \mathbb{E}[N\mathbb{E}(X^2) + N(N-1)\mathbb{E}(X)^2] \\ &= \mathbb{E}[N(\mathbb{E}(X^2) - \mathbb{E}(X)^2) + N^2\mathbb{E}(X)^2],\end{aligned}$$

ce qui donne

$$\mathbb{E}(S^2) = \mathbb{E}(N)\mathbb{V}(X) + \mathbb{E}(N^2)\mathbb{E}(X)^2. \tag{B.4}$$

D'après les égalités (B.3) et (B.4) obtenues précédemment, la variance de la variable aléatoire S s'écrit

$$\begin{aligned}\mathbb{V}(S) &= \mathbb{E}(S^2) - \mathbb{E}(S)^2 \\ &= \mathbb{E}(N)\mathbb{V}(X) + \mathbb{E}(N^2)\mathbb{E}(X)^2 - (\mathbb{E}(N)\mathbb{E}(X))^2 \\ &= \mathbb{E}(N)\mathbb{V}(X) + \mathbb{E}(X)^2[\mathbb{E}(N^2) - \mathbb{E}(N)^2].\end{aligned}$$

Nous obtenons donc

$$\mathbb{V}(S) = \mathbb{E}(N)\mathbb{V}(X) + \mathbb{E}(X)^2\mathbb{V}(N). \tag{B.5}$$

B.2 Définition des indicateurs SCR, MSE et RMSE

Trois indicateurs sont utilisés dans le cadre de ce mémoire pour évaluer les écarts observables sur des variables quantitatives. Notons y_i la valeur observée de la variable d'intérêt et $\hat{\mu}_i$ la valeur prédite de la variable d'intérêt.

Dans un premier temps, présentons l'indicateur SCR. La Somme des Carrés des Résidus, permet de mesurer l'écart entre la valeur observée et la valeur prédite. L'écart étant élevé au carré, les erreurs importantes sont donc majorées de manière considérable

$$SCR = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2.$$

Parallèlement à l'utilisation de cet indicateur, nous pouvons notamment compléter l'analyse avec les deux autres indicateurs : le MSE et le RMSE.

Le carré moyen des erreurs (ou Mean Square Error) est défini comme la moyenne arithmétique des carrés des écarts entre les valeurs observées et les valeurs prédites. Dans le cadre de régression de bonne qualité, le MSE est très faible. Cet indicateur mesure donc la précision de l'estimateur $\hat{\mu}_i$ obtenu. Le MSE est défini par

$$MSE = \mathbb{E}[(\hat{\mu}_i - y_i)^2].$$

Enfin, l'erreur quadratique moyenne (RMSE) est défini par

$$RMSE = \sqrt{MSE}.$$

B.3 Démonstration de la formule de passage du montant de la dépense moyenne au montant de remboursement complémentaire moyen

Cette section détaille les calculs intermédiaires pour le passage de la dépense moyenne à l'obtention du remboursement complémentaire moyen. Soit les termes définis suivants :

- $\mathbb{E}(Y_{RC})$ désigne la prise en charge moyenne de l'organisme complémentaire santé pour un acte donné ;
- R_{Mut} désigne le remboursement complémentaire (hors remboursement de la Sécurité Sociale) ;
- R_{SS} désigne le remboursement de la Sécurité Sociale ;
- $F_{\mu,\sigma}$ désigne la fonction de répartition de la loi Log normale de paramètres μ et σ ;
- ϕ désigne la fonction de répartition de la loi normale centrée réduite.

Nous obtenons alors les calculs suivants

$$\begin{aligned}
\mathbb{E}(Y_{RC}) &= \mathbb{E}(\min(\max(Y - \tau_{RSS} \times BRSS, 0), (\tau_{RC} - \tau_{RSS}) \times BRSS)) \\
&= \mathbb{E}(\min(\max(Y - R_{SS}, 0), R_{Mut} - R_{SS})) \\
&= \mathbb{E}(0 \times 1_{Y \in [0, R_{SS}]} + (Y - R_{SS}) \times 1_{Y \in [R_{SS}, R_{Mut}]} + \\
&\quad (R_{Mut} - R_{SS}) \times 1_{Y \in [R_{Mut}, +\infty]}) \\
&= \mathbb{E}((Y - R_{SS}) \times 1_{Y \in [R_{SS}, R_{Mut}]}) + (R_{Mut} - R_{SS}) \times \mathbb{P}(Y \in [R_{Mut}, +\infty]).
\end{aligned}$$

Or,

$$\mathbb{E}((Y - R_{SS}) \times 1_{Y\in[R_{SS},R_{Mut}]}) = \mathbb{E}(Y \times 1_{Y\in[R_{SS},R_{Mut}]}) - R_{SS} \times \mathbb{P}(Y \in [R_{SS}, R_{Mut}]),$$

et

$$\mathbb{E}(Y \times 1_{Y\in[R_{SS},R_{Mut}]}) = \frac{1}{\sigma 2\pi} \int_{R_{SS}}^{R_{Mut}} \exp(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}) \, \mathrm{d}x.$$

Nous effectuons le changement de variable $s = \ln(x)$

$$\begin{aligned}\mathbb{E}(Y \times 1_{Y\in[R_{SS},R_{Mut}]}) &= \frac{1}{\sigma 2\pi} \int_{\ln(R_{SS})}^{\ln(R_{Mut})} \exp(-\frac{(s - \mu)^2}{2\sigma^2}) \exp(s) \, \mathrm{d}s \\ &= \frac{\exp(\mu + \frac{\sigma^2}{2})}{\sigma\sqrt{2\pi}} \int_{\ln(R_{SS})}^{\ln(R_{Mut})} \exp(-\frac{(s - (\mu + \sigma^2)^2)}{2\sigma^2}) \, \mathrm{d}s.\end{aligned}$$

Nous effectuons un second changement de variable $t = \frac{s-(\mu+\sigma^2)^2}{\sigma}$

$$\begin{aligned}\mathbb{E}(Y \times 1_{Y\in[R_{SS},R_{Mut}]}) &= \frac{\exp(\mu + \frac{\sigma^2}{2})}{\sqrt{2\pi}} \int_{\frac{\ln(R_{SS})-(\mu+\sigma^2)}{\sigma}}^{\frac{\ln(R_{Mut})-(\mu+\sigma^2)}{\sigma}} \exp(-\frac{t}{2} \, \mathrm{d}t \\ &= \exp(\mu + \frac{\sigma^2}{2})(\phi(\frac{\ln(R_{Mut}) - (\mu + \sigma^2)}{\sigma}) - \phi(\frac{\ln(R_{SS}) - (\mu + \sigma^2)}{\sigma})).\end{aligned}$$

Finalement, l'équation permettant d'obtenir le remboursement complémentaire moyen est donnée par

$$\begin{aligned}\mathbb{E}(Y_{RC}) &= \exp(\mu + \frac{\sigma^2}{2})(\phi(\frac{\ln(R_{Mut}) - (\mu + \sigma^2)}{\sigma}) - \phi(\frac{\ln(R_{SS}) - (\mu + \sigma^2)}{\sigma})) \\ &\quad - R_{SS} \times (F_{\mu,\sigma}(R_{Mut}) - F_{\mu,\sigma}(R_{SS})) + (R_{Mut} - R_{SS}) \times (1 - F_{\mu,\sigma}(R_{Mut})) \\ &= \exp(\mu + \frac{\sigma^2}{2})(\phi(\frac{\ln(R_{Mut}) - (\mu + \sigma^2)}{\sigma}) - \phi(\frac{\ln(R_{SS}) - (\mu + \sigma^2)}{\sigma})) \\ &\quad - R_{SS} \times (1 - F_{\mu,\sigma}(R_{SS})) + R_{Mut} \times (1 - F_{\mu,\sigma}(R_{Mut})).\end{aligned}$$