



Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du
Diplôme d'Actuaire EURIA
et de l'admission à l'Institut des Actuaires

le 16 Décembre 2020

Par : Floriane PIVETEAU

Titre : Pilotage d'un portefeuille de santé individuelle : l'apport de l'Open Data pour analyser l'inflation

Confidentialité : Oui (Durée : 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Marine HABART

Yann QUERE

Renaud CAILLET

Signatures :

Entreprise :

GENERALI

Signature :

Membre présent du jury de l'EURIA :

Pierre AILLIOT

Directeur de mémoire en entreprise :

Alexandre DIAS LOPES

Signature :

Invité :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion

de documents actuariels

(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :

Signature du candidat :

Résumé

La revalorisation des contrats joue un rôle majeur sur le pilotage et la rentabilité d'un portefeuille d'assurance complémentaire santé individuelle. Elle permet d'ajuster les tarifs chaque année afin de faire face à l'évolution des frais de santé. L'inflation de la sinistralité est un élément important à prendre en compte afin de revaloriser les contrats conformément à l'évolution passée du coût des dépenses de santé.

Ce mémoire s'intéresse à l'inflation de la sinistralité et propose une étude de l'inflation dans le but d'améliorer le pilotage du portefeuille de Generali. L'approche proposée consiste à exploiter des données en Open Data afin d'affiner l'étude de l'inflation de la sinistralité du portefeuille de santé individuelle de Generali. La base Open DAMIR, contenant les remboursements de l'assurance maladie tous régimes confondus, est utilisée à cet effet. L'utilisation de cette base permet de disposer d'un volume de prestations important et d'informations supplémentaires par rapport au portefeuille de santé individuelle.

Un portefeuille Generali est reconstitué dans le but de mesurer l'inflation à partir de ces données du marché. Il est construit en tenant compte de la nature des prestations couvertes par une complémentaire santé et du profil de risque du portefeuille de Generali. Des traitements sont réalisés afin de mesurer l'inflation réelle, hors impact de la qualité des données ou des modifications réglementaires. Des données externes sont ajoutées afin d'enrichir l'étude.

Dans un premier temps, l'inflation est calculée au global sur la base d'étude afin de permettre à Generali de positionner son portefeuille par rapport aux données du marché. Dans un second temps, l'inflation est observée de manière plus fine, en utilisant une granularité de calcul composée de la famille de remboursement et des caractéristiques des bénéficiaires des soins (âge, sexe et région de résidence). L'inflation est étudiée dans le but d'identifier des variables qui l'influencent. Dans un esprit de recherche, l'étude se limite à quelques familles de remboursement. L'inflation de la sinistralité est alors étudiée à l'aide d'une méthode de Machine Learning pour chaque famille considérée. Le suivi des variables identifiées comme influentes permettra d'améliorer la compréhension de l'inflation observée sur le portefeuille de Generali.

Mots clés: Complémentaire santé individuelle, Dérive de la sinistralité, Inflation, Open Data, Open DAMIR, Forêts aléatoires

Abstract

Contract revaluation plays a major role in the management and profitability of individual supplemental health insurance portfolios. It enables tariffs to be adjusted each year to keep pace with changes in healthcare costs. Inflation in claims is an important factor to be considered to reevaluate contracts in line with past changes in healthcare expenditures.

This study focuses on inflation of the average cost of claims and proposes an analysis of inflation in order to improve the management of Generali's portfolio. The proposed approach consists in exploiting Open Data to refine the study of inflation of Generali's individual health portfolio. The Open DAMIR database, which contains reimbursements from the entire compulsory health insurance system, is used for this purpose. This database provides a large volume of healthcare acts and additional information compared to the individual health portfolio.

A Generali portfolio is reconstituted to measure inflation based on this market data. It is constructed considering the nature of the healthcare costs covered by a complementary health insurance plan and the risk profile of the Generali portfolio. Adjustments are made to measure actual inflation, excluding the impact of data quality or regulatory changes. External data is added to enrich the study.

Firstly, inflation of the average cost of claims is calculated globally on the study database in order to allow Generali to position its portfolio in relation to market data. In a second step, inflation is observed in more detail, using a granularity of calculation composed of the reimbursement family and characteristics of the care beneficiaries (age, gender and region of residence). Inflation is studied to identify influential variables. In the spirit of research, the study is limited to only a few reimbursement families. Inflation in claims is then studied using a Machine Learning method for each family under consideration. The monitoring of the variables identified as influential could make it possible to improve the understanding of the inflation observed on Generali's portfolio.

Keywords: Individual supplemental health insurance, Change in claims, Inflation, Open Data, Open DAMIR, Random forest

Note de synthèse

Contexte et objectif

Dans le domaine de l'assurance santé individuelle, les tarifs sont ajustés chaque année afin de faire face à l'évolution des frais de santé. La revalorisation des contrats permet aux organismes d'assurance de maintenir leur ratio de sinistres à primes à l'équilibre et joue en ce sens un rôle majeur dans le pilotage et la rentabilité de leur portefeuille. Un taux de majoration est calculé lors de l'exercice annuel de revalorisation. Il se compose généralement de trois parties, liées aux évolutions réglementaires, aux modifications de garanties et à la dérive des dépenses de santé. Cette dernière partie permet de revaloriser les contrats conformément à l'évolution passée de la sinistralité. L'indice de la dérive de sinistralité utilisé par Generali est basé sur des indices de marché et sur la mesure de la dérive du portefeuille de santé individuelle. La dérive peut être liée à la fois à l'évolution du coût moyen des prestations de santé et à la variation de la quantité de biens et de services médicaux consommés par les assurés. L'inflation de la sinistralité étudiée correspond à la variation du coût moyen des dépenses de santé hors effet volume lié à la consommation. Il s'agit d'un phénomène dépendant d'une multitude de facteurs non nécessairement observables. L'étude des éléments influençant l'inflation permet d'améliorer sa compréhension et peut ainsi permettre l'amélioration du pilotage du portefeuille de Generali.

Dans le contexte actuel d'ouverture de données publiques de plus en plus nombreuses et variées, l'utilisation d'Open Data est envisagée afin d'affiner l'étude de l'inflation de la sinistralité du portefeuille de santé individuelle de Generali. La base Open DAMIR (Dépenses d'Assurance Maladie Inter-Régimes) est utilisée à cet effet. Ce jeu de données présente l'avantage de contenir des caractéristiques supplémentaires relatives aux prestations et regroupe un volume d'actes important permettant une étude de l'inflation de manière segmentée. Des données externes sont également ajoutées afin d'enrichir l'étude. L'objectif du mémoire est d'exploiter les informations issues de ces données du marché afin d'en extraire des enseignements permettant d'améliorer la compréhension de l'inflation du portefeuille de santé individuelle.

Un portefeuille Generali est reconstitué à partir de la base Open DAMIR afin de mesurer une inflation comparable à celle du portefeuille sur ces données. L'inflation est dans un premier temps calculée au global sur la base d'étude afin de permettre à Generali de positionner son portefeuille par rapport à l'inflation du marché. L'inflation est ensuite étudiée sur une granularité plus fine dans le but d'identifier les variables qui l'influencent.

Étude de la base Open DAMIR et sélection des données pertinentes

La base Open DAMIR regroupe l'ensemble des prestations prises en charge par l'assurance maladie obligatoire, à l'exception d'une majorité des prestations hospitalières du secteur public. Elle contient les montants associés à chaque remboursement ainsi que des informations concernant l'acte médical, le bénéficiaire des soins et les professionnels de santé exécutant et prescripteur.

Cet ensemble de données a une structure particulière. Il s'agit d'une base agrégée dans laquelle plusieurs lignes peuvent être associées à un même remboursement en raison de l'existence de différents types de remboursements (remboursements de la prestation de référence, remboursements de compléments d'acte et remboursements complémentaires). Différentes variables existent pour tenir compte de cette structure atypique, permettant notamment d'éviter les doubles comptages de montants ou de quantités.

Les bases mensuelles Open DAMIR de 2016 à 2019 sont utilisées pour l'étude. Une sélection des données est réalisée en tenant compte de la nature des prestations couvertes par une complémentaire santé afin de rapprocher le périmètre de la base Open DAMIR de celui du portefeuille de Generali. Cette sélection permet en outre de réduire le volume des bases de données, puisque chaque base mensuelle contient initialement environ 30 millions de lignes.

Une analyse de la qualité des données est réalisée afin d'identifier les données jugées de mauvaise qualité. L'objectif de cette étape est d'augmenter le niveau de confiance accordé aux données. Les contrôles réalisés permettent d'identifier les incohérences à éliminer et de garantir la fiabilité des données.

Préparation des données pour l'étude

La base d'étude doit permettre de mesurer au mieux l'inflation. Les différents remboursements sont alors regroupés en familles de manière à disposer d'une granularité cohérente pour évaluer l'évolution du coût moyen.

Les modifications légales créent une inflation artificielle des frais de santé qui ne doit pas être prise en compte dans le calcul de l'inflation. Dans ce cadre, les natures de prestations apparues ou disparues durant le périmètre temporel de l'étude sont exclues de la base de données. Les montants impactés par une évolution réglementaire sont également retraités afin de mesurer l'inflation réelle. Des montants *as-if* sont calculés de sorte que les montants considérés soient sur la même base réglementaire tout au long de la période étudiée.

L'inflation calculée sur la base d'étude doit pouvoir être comparée à celle observée sur le portefeuille de Generali. Des ajustements sont réalisés afin de reconstituer un portefeuille Generali à partir des données de la base Open DAMIR.

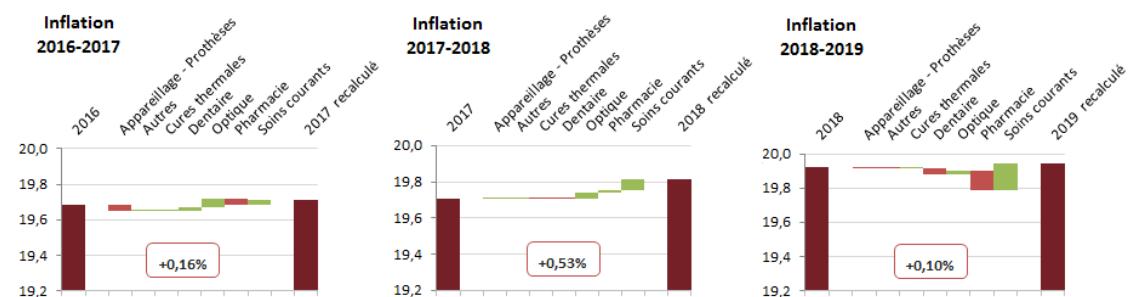
Pour chacune des prestations contenues dans la base d'étude, le montant qui serait remboursé par Generali si elle concernait ses assurés est calculé. Une étape préliminaire consiste à étudier le portefeuille afin de sélectionner les garanties à appliquer. Étant donné la nature et la structure de la base Open DAMIR, l'application des garanties ne permet pas de simuler le montant exact remboursé par Generali. La méthodologie de calcul des remboursements complémentaires est alors testée sur les données du portefeuille avant d'être appliquée à la base d'étude afin d'évaluer les différents biais.

Dans le but de mesurer une inflation globale comparable à celle du portefeuille de Generali, le profil des assurés est approché en appliquant des coefficients de pondération aux quantités d'actes et aux montants remboursés pour chaque tranche d'âges, sexe et région de résidence des bénéficiaires des soins de la base d'étude.

Mesure de l'inflation globale

L'inflation correspond à l'évolution du coût moyen des dépenses de santé des assurés entre deux années consécutives N-1 et N. Étant donné que la base d'étude contient différents types de prestations avec des coûts moyens très différents, l'inflation calculée au global est très dépendante de l'évolution de la composition du portefeuille. Afin de mesurer une inflation globale cohérente, le calcul est réalisé pour chaque famille de remboursements puis les différentes inflations sont agrégées afin d'obtenir l'inflation globale.

L'inflation est ainsi observée au global et par poste. Elle permet à Generali de positionner son portefeuille par rapport aux données nationales.



L'étude des données permet d'observer qu'environ 3% des prestations de l'année 2019 sont manquantes en raison du délai nécessaire à leur remboursement. Les bases Open DAMIR de 2020 ne seront cependant mises en ligne qu'au début de l'année 2021. Il est décidé par prudence de ne pas s'intéresser à l'inflation 2018-2019 dans la suite de l'étude.

Définition des variables de l'étude

Afin d'être analysée, l'inflation est calculée à un niveau plus fin en considérant des profils de risques. Ces derniers doivent être constitués d'un nombre d'actes suffisamment élevé afin de mesurer une inflation cohérente. Les lignes de la base sont alors agrégées selon une granularité composée de la famille de remboursements, de l'âge, du sexe et de la région de résidence des bénéficiaires des soins. Les inflations 2016-2017 et 2017-2018 sont alors observées sur 208 profils différents pour chaque famille de remboursements.

Les variables supplémentaires fournies par la base Open DAMIR et jugées pertinentes sont transformées afin d'être disponibles à la granularité étudiée. En particulier, la variable relative au professionnel de santé prescripteur est utilisée afin de caractériser les lignes agrégées. Pour chaque ligne, la proportion des principales modalités est calculée et est utilisée comme variable explicative. La variation de cette proportion entre deux années est également ajoutée sur chaque ligne de données.

Afin d'enrichir la base d'étude, des données externes sont ajoutées. Elles permettent de caractériser les différents profils d'individus obtenus. En particulier, les variables suivantes sont introduites :

- la densité d'établissements recensés dans le FINESS (Fichier National des Établissements Sanitaires et Sociaux), permettant de caractériser l'offre de soins dans les différentes régions,
- le niveau de vie médian par région, issu de l'INSEE, pouvant refléter la capacité des habitants à assumer des frais de santé plus ou moins élevés,
- le taux de bon niveau général d'état de santé, caractérisant les régions et les sexes, construit à partir de données issues de l'enquête Vie Quotidienne et Santé (VQS) de la DREES.

Méthodologie de l'étude

En étudiant l'inflation sur la base complète, les variables les plus importantes conduisent à une discrimination des différentes familles de remboursements. Une étude est alors réalisée pour chaque famille de prestations dans le but d'expliquer l'inflation.

Dans ce mémoire, l'étude de l'inflation est présentée pour deux familles de remboursements ayant un poids particulièrement important dans le portefeuille de Generali : la famille des médicaments remboursés à 65% par la Sécurité Sociale et la famille des consultations et visites des généralistes. L'objectif est d'établir une relation entre l'inflation et les différentes variables explicatives dans le but d'identifier des variables qui influencent l'inflation.

La méthode des forêts aléatoires est retenue en raison de sa capacité à prendre en compte efficacement les relations non linéaires, compte tenu des liaisons observées sur les données. D'autres éléments ont également contribué à son choix, à savoir son caractère automatique puisqu'elle possède peu de paramètres à régler, sa propriété de modèle non paramétrique, ainsi que sa quantification de l'importance des variables.

Pour chacune des familles, une étude préliminaire des corrélations entre les variables est réalisée afin de comprendre les interactions entre elles. Une forêt aléatoire est implémentée et les paramètres sont optimisés. Une sélection de variables est ensuite réalisée afin d'identifier les variables pertinentes dans le but de faciliter l'interprétation du modèle. Trois méthodes de sélection sont comparées : les méthodes VSURF et de Boruta ainsi que l'algorithme RFE. Les variables les plus importantes et leurs effets sont enfin observés à l'aide de l'indice d'importance par permutation des forêts aléatoires et de la méthode SHAP.

Pour la première famille étudiée, un modèle linéaire est utilisé en comparaison. Ce type de modèle est très utilisé en actuariat pour sa simplicité d'interprétation. Il suppose cependant que certaines hypothèses soient vérifiées et demande un temps de paramétrage plus long afin d'appréhender au mieux les relations non linéaires. Les deux méthodes permettent d'obtenir des conclusions cohérentes entre elles. La méthode des forêts aléatoires s'avère alors pertinente à utiliser dans ce contexte.

L'inflation observée sur la famille "Médicaments PH65"

Les modèles mettent en évidence que l'inflation mesurée sur la famille des médicaments remboursés à 65% est principalement impactée par l'âge du bénéficiaire des soins. Ils permettent également d'observer que l'inflation diffère en fonction des proportions de professionnels de santé prescripteurs, notamment des proportions de prescripteurs gynécologues/sage-femmes, ophtalmologues et dermatologues. Ces proportions de professionnels de santé prescripteurs permettent de caractériser les tranches d'âge. Cette observation permet de mettre en évidence que l'inflation est liée à l'âge en raison du type de médicaments consommés qui diffère en fonction de l'âge. Les proportions de professionnels de santé prescripteurs apportent également des explications supplémentaires. Pour un âge donné, l'inflation peut différer en fonction de la proportion de certains prescripteurs et donc de la proportion consommée de certains types de médicaments.

L'étude permet alors de donner une indication sur les profils d'assurés associés à des inflations plus ou moins élevées. Les variables identifiées comme importantes permettent de caractériser ces profils afin d'expliquer l'inflation.

Les observations pourront être étudiées et affinées sur le portefeuille de Generali. L'information relative au professionnel de santé prescripteur est à récupérer afin d'étudier si les observations issues des données de marché se confirment. L'inflation par âge pourra de plus être mesurée en utilisant des tranches moins étendues.

L'inflation observée sur la famille "Généralistes"

Sur cette seconde famille de remboursements, le modèle met en évidence que l'inflation est liée à l'âge mais dépend principalement de la région de résidence des bénéficiaires des soins. Les variables externes introduites permettent d'enrichir l'étude. Il apparaît que le niveau de vie médian et le nombre d'établissements sanitaires et sociaux (répertoriés dans le FINESS) ramené au niveau de population par région contribuent de manière importante au modèle. Ces variables caractérisent différemment les régions et distinguent d'une autre manière leurs inflations. L'interprétation est cependant limitée par le fait que la base Open DAMIR ne donne pas accès à une variable de localisation plus précise que la région.

Sur le portefeuille de santé individuelle, la localisation est plus détaillée. La construction d'un zonier permettant l'étude de l'inflation selon une granularité géographique plus fine permettra de conclure à une éventuelle influence de la densité d'établissements sanitaires et sociaux et du niveau de vie. Le cas échéant, il sera possible d'identifier des zones géographiques et des profils d'assurés présentant une inflation plus ou moins élevée.

Conclusion et possibilités offertes par l'étude

Les données utilisées dans cette étude apportent des informations supplémentaires mais ne permettent pas de comprendre réellement les phénomènes sous-jacents à l'inflation observée. L'étude permet toutefois d'identifier certaines variables explicatives qui pourraient s'avérer pertinentes à suivre sur le portefeuille Generali.

Pour les deux familles de remboursements présentées, les variables identifiées comme importantes dans les modèles permettent de caractériser des profils d'assurés associés à une inflation plus ou moins élevée. Si les observations issues de la base Open DAMIR se confirment sur le portefeuille de Generali, le suivi de ces variables permettra d'améliorer la compréhension de l'inflation mesurée. En observant les différents indicateurs évoqués dans le temps, certaines actions de prévention pourraient être envisagées afin de réduire l'inflation. Dans le cadre du pilotage du portefeuille de Generali, les actions de prévention sont actuellement identifiées à partir d'études sur le coût moyen des dépenses de santé. Disposer d'observations relatives à l'inflation permettrait alors d'affiner les actions mises en place.

L'une des limites principales de cette étude est la nature agrégée de la base Open DAMIR. Les prestations ne sont pas disponibles ligne à ligne, ce qui engendre un biais dans certains retraitements réalisés. Les études réalisées sont également limitées par le fait que la base Open DAMIR ne donne pas accès à une variable de localisation plus fine que la région et à un âge plus détaillé que les tranches d'âges disponibles. Il convient également de noter que l'étude réalisée ne permet pas d'étudier toutes les garanties couvertes par une complémentaire santé, en particulier pour les soins non remboursés par la Sécurité Sociale comme par exemple les médecines douces (ostéopathie, acupuncture, etc.).

L'étude réalisée apporte une explication de l'inflation mesurée sur les données de la Sécurité Sociale. Les observations sont toutefois valides dans un contexte économique et sanitaire classique. En 2020, la pandémie de Covid-19 et le confinement ont bouleversé le contexte sanitaire. Ils ont modifiés la consommation de biens et de services médicaux et un impact est attendu sur l'inflation 2019-2020. Les observations sont également valides dans un contexte réglementaire stable ou présentant des évolutions facilement identifiables sur les données. A partir du 1er janvier 2020, l'impact de la réforme 100% santé ne sera cependant pas facile à isoler en raison de la mise en place de paniers de soins dentaires, optiques et auditifs intégralement pris en charge. L'inflation hors modifications réglementaires sera alors difficilement mesurable sur ces postes de soins.

Summary

Context and objective

In the area of individual health insurance, tariffs are adjusted each year in order to keep pace with changes in healthcare costs. The revaluation of contracts enables insurance companies to maintain their loss ratio at break-even and thus plays a major role in the management and profitability of their portfolio. An increase rate is calculated during the annual revaluation. It is generally made up of three parts, linked to regulatory changes, changes in guarantees and the drift in health expenses. This last part enables contracts to be revalued in line with past changes in claims. The claim drift index used by Generali is based on market indices and on the measurement of the drift of the individual health portfolio. Drift can be related both to changes in the average cost of health benefits and to changes in the quantity of medical goods and services consumed by policyholders. The inflation in claims studied corresponds to the change in the average cost of healthcare expenditure excluding the volume effect linked to consumption. It is a phenomenon dependent on a multitude of factors not necessarily observable. The study of the factors influencing inflation makes it possible to improve its understanding and can thus improve the management of Generali's portfolio.

In the current context of the opening up of more and more numerous and varied public data, the use of Open Data is being considered in order to refine the study of the inflation of the claims experience of Generali's individual health portfolio. The Open DAMIR (Inter-system Health Insurance Expenditure) database is used for this purpose. This dataset has the advantage of containing additional features relating to benefits and includes a large volume of acts allowing a study of inflation in a segmented manner. External data is also added to enrich the study. The aim of the study is to use the information from this market data to extract insights that will improve the understanding of inflation in the individual health portfolio.

A Generali portfolio is reconstituted from the Open DAMIR database in order to measure inflation comparable to that of the portfolio on this data. Inflation is firstly calculated globally on the study database in order to allow Generali to position its portfolio in relation to market inflation. Inflation is then studied on a finer granularity in order to identify the variables that influence it.

Study of the Open DAMIR database and selection of relevant data

The Open DAMIR database includes all the benefits covered by compulsory health insurance, except for a majority of hospital benefits in the public sector. It contains the amounts associated with each reimbursement as well as information on the medical act, the beneficiary of the treatment and the health professionals performing and prescribing the treatment.

This dataset has a special structure. It is an aggregated base in which several lines can be associated with the same reimbursement due to the existence of different types of reimbursements (reimbursements of the reference benefit, reimbursements of complementary acts and complementary reimbursements). Different variables exist to consider this atypical structure, in particular to avoid double counting of amounts or quantities.

The Open DAMIR monthly bases from 2016 to 2019 are used for the study. A selection of data is made taking into account the nature of the benefits covered by a supplementary health insurance plan in order to bring the scope of the Open DAMIR database closer to that of Generali's portfolio. This selection also makes it possible to reduce the volume of the databases, since each monthly database initially contains approximately 30 million lines.

A data quality analysis is carried out in order to identify data deemed to be of poor quality. The objective of this step is to increase the level of confidence in the data. The checks carried out make it possible to identify the inconsistencies to be eliminated and to guarantee the reliability of the data.

Preparing data for the study

The study database should allow for the best possible measurement of inflation. The various reimbursements are then grouped into families in order to have a coherent granularity to evaluate the evolution of the average cost.

The legal changes create artificial inflation of healthcare costs which must not be taken into account in the calculation of inflation. In this context, the types of acts that have appeared or disappeared during the time frame of the study are excluded from the database. The amounts impacted by a regulatory change are also restated in order to measure real inflation. *As-if* amounts are calculated so that the amounts considered are on the same regulatory basis throughout the study period.

Inflation calculated on the study database must be comparable to that observed in Generali's portfolio. Adjustments are made in order to reconstitute a Generali portfolio based on data from the Open DAMIR database.

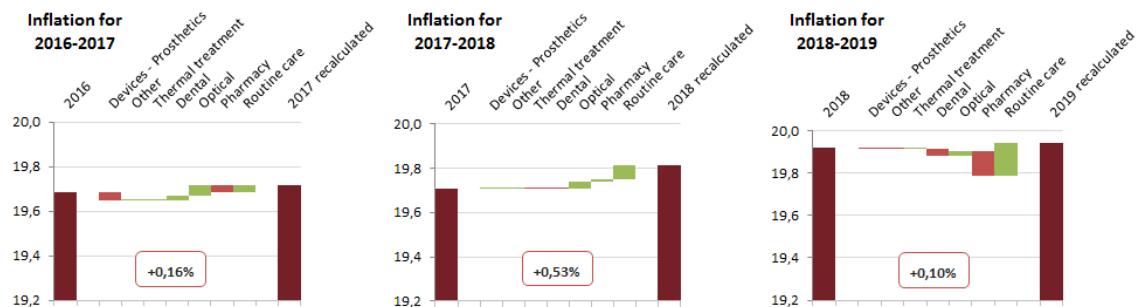
For each of the acts contained in the study database, the amount that would be reimbursed by Generali if it concerned policyholders is calculated. A preliminary step consists of studying the portfolio in order to select the guarantees to be applied. Given the nature and structure of the Open DAMIR database, the application of the guarantees does not make it possible to simulate the exact amount reimbursed by Generali. The methodology for calculating the additional reimbursements is tested on the portfolio data before being applied to the study database in order to evaluate the different biases.

In order to measure overall inflation comparable to that of Generali's portfolio, the profile of policyholders is approximated by applying weighting coefficients to the quantities of acts and amounts reimbursed for each age group, gender and region of residence of the beneficiaries of care in the study database.

Measurement of overall inflation

Inflation corresponds to the change in the average cost of health expenditure between two consecutive years N-1 and N. Since the study base contains different types of benefits with very different average costs, the overall calculated inflation is very dependent on the changing composition of the portfolio. In order to measure a consistent overall inflation, the calculation is performed for each family of reimbursements and then the different inflations are aggregated to obtain overall inflation.

Inflation is thus observed overall and by healthcare category. It enables Generali to position its portfolio in relation to national data.



The study of the data allows us to observe that approximately 3% of the benefits for the year 2019 are missing due to the delay necessary for their reimbursement. However, the 2020 Open DAMIR databases will not be put online until the beginning of 2021. It is decided as a precaution not to take an interest in the inflation 2018-2019 in the rest of the study.

Definition of study variables

In order to be analysed, inflation is calculated at a finer level by considering risk profiles. These must consist of a sufficiently high number of acts to measure consistent inflation. The database lines are then aggregated according to a granularity composed of the family of reimbursements, age, sex and region of residence of the policyholders. The 2016-2017 and 2017-2018 inflations are then observed on 208 different profiles for each family of reimbursements.

The additional variables provided by the Open DAMIR database deemed relevant are transformed in order to be available at the studied granularity. In particular, the variable relating to the prescribing health professional is used to characterize the aggregated lines. For each line, the proportion of the main modalities is calculated and used as an explanatory variable. The variation of this proportion between two years is also added on each data line.

In order to enrich the study base, external data is added. It enables the different profiles of individuals obtained to be characterised. In particular, the following variables are introduced :

- the density of establishments listed in the FINESS (*Fichier National des Établissements Sanitaires et Sociaux*), making it possible to characterise the treatment offer in the different regions,
- the median standard of living per region, from INSEE, which can reflect the capacity of inhabitants to pay for health care to a greater or lesser extent,
- the rate of a good general level of health status, characterising regions and sexes, constructed from data from the DREES' *Vie Quotidienne et Santé* (VQS) survey.

Study methodology

When studying inflation on the complete dataset, the most important variables lead to discrimination between different families of reimbursements. A study is then carried out for each benefit family in order to explain inflation.

In this report, the study of inflation is presented for two families of reimbursement which have a particularly important weight in Generali's portfolio : the family of medicines reimbursed at 65% by Social Security and the family of consultations and visits by general practitioners. The objective is to establish a relationship between inflation and the various explanatory variables in order to identify variables that influence inflation.

The random forest method is chosen because of its ability to effectively consider non-linear relationships, given the observed linkages in the data. Other elements also contributed to its choice, namely its automatic character since it has few parameters to adjust, its non-parametric model property, as well as its quantification of the importance of the variables.

For each of the families, a preliminary study of the correlations between the variables is carried out in order to understand the interactions between them. A random forest is implemented and the parameters are optimised. A selection of variables is then carried out in order to identify the relevant variables in order to facilitate the interpretation of the model. Three selection methods are compared : the VSURF and Boruta methods and the RFE algorithm. The most important variables and their effects are finally observed using the random forest permutation importance index and the SHAP method.

For the first family studied, a linear model is used for comparison. This type of model is widely used in actuarial science for its simplicity of interpretation. However, it supposes that certain hypotheses are verified and requires a longer parameterization time in order to understand non-linear relationships. Both methods allow to obtain results that are consistent with each other. The random forest method is therefore relevant to use in this context.

Inflation observed on the "PH65 medicines" family

The models show that inflation measured on the family of medicines reimbursed at 65% is mainly impacted by the age of the healthcare beneficiary. They also show that inflation differs according to the proportions of prescribing health professionals, in particular the proportions of gynaecologists/midwives, ophthalmologists and dermatologists. These proportions of prescribing health professionals make it possible to characterise the age groups. This observation shows that inflation is linked to age because of the type of medicines consumed, which differs according to age. The proportions of prescribing health professionals also provide additional explanations. For a given age, inflation may differ according to the proportion of certain prescribers and therefore the proportion consumed of certain types of medicines.

The study then makes it possible to give an indication of the profiles of insured persons associated with higher or lower inflation. The variables identified as important make it possible to characterise these profiles in order to explain inflation.

Observations can be studied and refined on Generali's portfolio. The information relating to the prescribing healthcare professional is to be retrieved in order to study whether the observations from the market data are confirmed. In addition, age-specific inflation can be measured using smaller ranges.

Inflation observed in the "General practitioners" family

For this second family of reimbursements, the model shows that inflation is age-related but depends mainly on the region of residence of the care recipients. The external variables introduced enrich the study. It appears that the median standard of living and the number of health and social institutions (listed in FINESS) brought down to the population level by region make an important contribution to the model. These variables characterise the regions differently and distinguish their inflations in a different way. The interpretation is however limited by the fact that the Open DAMIR database does not give access to a more precise location variable than the region.

On the individual health portfolio, the location is more detailed. The construction of a zonier allowing the study of inflation according to a finer geographical granularity will allow to conclude to a possible influence of the density of health and social establishments and the standard of living. If so, it will be possible to identify geographical areas and insurance profiles with higher or lower inflation.

Conclusion and possibilities offered by the study

The data used in this study provide additional information but do not allow a real understanding of the underlying phenomena of the inflation observed. The study does, however, identify certain explanatory variables that could prove relevant to monitor on the Generali portfolio.

For the two families of reimbursements presented, the variables identified as important in the models make it possible to characterise the profiles of insured persons associated with higher or lower inflation. If the observations from the Open DAMIR database are confirmed on Generali's portfolio, the monitoring of these variables will improve the understanding of measured inflation. By observing the various indicators mentioned over time, preventive actions could be envisaged to reduce inflation. Within the framework of the steering of Generali's portfolio, preventive actions are currently identified on the basis of studies of the average cost of health expenditure. Having observations relating to inflation would then make it possible to refine the actions implemented.

One of the main limitations of this study is the aggregated nature of the Open DAMIR database. Benefits are not available on a line-by-line basis, which leads to a bias in some of the adjustments made. The studies carried out are also limited by the fact that the Open DAMIR database does not provide access to a location variable finer than the region and age ranges available. It should also be noted that the study carried out does not make it possible to study all the guarantees covered by supplementary health insurance, in particular for care not reimbursed by Social Security, such as alternative medicine (osteopathy, acupuncture, etc.).

The study provides an explanation of the inflation measured on Social Security data. However, the observations are only valid in a classical economic and health context. In 2020, the Covid-19 pandemic and containment have disrupted the health context. They have modified the consumption of medical goods and services and an impact is expected on inflation in 2019-2020. The observations are also only valid in a stable regulatory context or with easily identifiable changes in the data. From 1 January 2020, the impact of the *100% Santé* reform will not be easy to isolate due to the introduction of baskets of dental, optical and hearing care that are fully covered. Inflation excluding regulatory changes will then be difficult to measure on these care categories.

Remerciements

Je tiens tout d'abord à remercier Rémi BERTHOLON, directeur de la Plateforme Actuariat Province de Generali, de m'avoir accueilli au sein de l'entreprise et de m'avoir permis de réaliser mon stage de fin d'études dans de bonnes conditions.

Je souhaite tout particulièrement remercier mon tuteur Alexandre DIAS LOPES, sans qui ce mémoire n'aurait pas vu le jour. Je le remercie vivement pour son aide et ses conseils avisés, mais aussi pour son soutien, son écoute et pour le temps qu'il m'a accordé tout au long de la réalisation de ce mémoire.

Ma reconnaissance s'adresse également à l'ensemble des actuaires de la Plateforme Actuariat Province et en particulier aux membres de l'équipe Wiz'You pour leurs conseils et leur accueil chaleureux malgré le contexte de télétravail.

Je ne peux écrire ces remerciements sans évoquer mes collègues stagiaires et alternants qui m'ont encouragée tout au long de ce stage. Je les remercie pour les nombreux échanges et pour leur bonne humeur.

Je remercie ensuite Franck VERMET, directeur de l'école, ainsi que toute l'équipe pédagogique de l'EURIA pour la qualité des enseignements reçus au cours de ces trois années de formation. J'adresse également mes remerciements à Karine ARZUR, ma tutrice universitaire, pour ses conseils relatifs à l'élaboration de ce mémoire.

Mes remerciements vont enfin à mes proches qui ont toujours été présents pour me soutenir durant mes études. J'adresse une pensée particulière à mon conjoint, qui me supporte au quotidien.

Table des matières

Introduction	1
1 La prise charge des dépenses de santé en France	4
1.1 Le fonctionnement du système de santé	4
1.1.1 Les régimes du système de santé	4
1.1.2 Les modalités de remboursement	7
1.1.3 Le niveau des remboursements	9
1.2 Les réformes marquantes du système de santé	12
1.3 La dérive des dépenses de santé	16
1.3.1 La maîtrise de la dérive des dépenses de l'assurance maladie	16
1.3.2 La prise en compte la dérive de la sinistralité des organismes d'assurance	17
1.4 La base Open DAMIR parmi l'offre d'Open Data de l'assurance maladie .	18
2 L'utilisation de données en Open Data pour étudier l'inflation	22
2.1 Étude de la base Open DAMIR	22
2.1.1 Description des données	22
2.1.2 Sélection des données pertinentes	26
2.1.3 Analyse de la qualité données	28
2.1.4 Fiabilisation des données	30
2.2 Préparation des données pour une mesure de l'inflation comparable à celle du portefeuille de Generali	32
2.2.1 Catégorisation des prestations	32
2.2.2 Retraitements des évolutions réglementaires	34
2.2.3 Création de la base d'étude	39
2.2.4 Rapprochement de la base d'étude et du portefeuille de Generali .	42
2.3 Statistiques descriptives	46
3 La mesure de l'inflation à partir de données du marché	49
3.1 Mesure et observation de l'inflation globale	49
3.1.1 Définition de l'inflation	49
3.1.2 Observation de l'inflation sur la base d'étude	51
3.1.3 Observation de l'impact des retraitements réalisés	52

3.2	Définition des variables de l'étude	53
3.2.1	Le choix de la variable à expliquer et des données utilisées	53
3.2.2	La définition de l'inflation à étudier	54
3.2.3	Les variables explicatives	57
3.2.4	Les familles de remboursements retenues pour l'étude	62
4	L'apport de la data science pour analyser l'inflation	64
4.1	Sélection des modèles	64
4.1.1	Le modèle linéaire gaussien	64
4.1.2	Les forêts aléatoires	67
4.2	Méthodes d'extraction des informations des modèles	71
4.3	Méthodes de sélection de variables	74
4.4	Étude de l'inflation associée aux médicaments PH65	76
4.4.1	Les associations entre les variables	76
4.4.2	Les forêts aléatoires	80
4.4.3	Le modèle linéaire	89
4.4.4	L'apport des modèles pour la compréhension de l'inflation	94
4.5	Étude de l'inflation associée aux consultations et visites des généralistes .	96
4.5.1	Les associations entre les variables	96
4.5.2	Les forêts aléatoires	98
4.5.3	L'apport du modèle pour la compréhension de l'inflation	103
Conclusion		104
Annexe		109
A Le traitement de la base Open DAMIR		110
A.1	Variables de la base Open DAMIR	110
A.2	Qualité des données	112
A.3	Évolutions réglementaires	113
A.4	Traitement des anomalies après agrégation de la base d'étude	115
B L'analyse de l'inflation		117
B.1	Méthodes de sélection de variables basées sur des forêts aléatoires	117
B.2	Compléments de l'étude de l'inflation sur la famille "Médicaments PH65" .	121
B.3	Compléments de l'étude de l'inflation sur la famille "Généralistes"	125
Bibliographie		128

Introduction

Dans le domaine de l'assurance santé individuelle, il est important d'ajuster les tarifs chaque année afin de faire face à l'évolution des frais de santé. La revalorisation des contrats permet aux organismes d'assurances de maintenir leur ratio de sinistres à primes à l'équilibre et joue en ce sens un rôle majeur dans le pilotage et la rentabilité de leur portefeuille. Un taux de majoration est calculé lors de l'exercice annuel de revalorisation. Il se compose généralement de trois parties, liées aux évolutions réglementaires, aux modifications de garanties et à la dérive des dépenses de santé. Cette dernière partie permet de revaloriser les contrats conformément à l'évolution passée de la sinistralité. L'indice de la dérive de sinistralité utilisé par Generali est basé sur des indices de marché et sur la mesure de la dérive du portefeuille de santé individuelle. La dérive de la sinistralité peut être liée à la fois à l'évolution du coût moyen des prestations de santé et à la variation de la quantité de biens et de services médicaux consommés par les assurés. L'inflation de la sinistralité étudiée correspond à la variation du coût moyen des dépenses de santé hors effet volume lié à la consommation. Il s'agit d'un phénomène dépendant d'une multitude de facteurs non nécessairement observables. L'étude des éléments influençant l'inflation permet d'améliorer sa compréhension et peut ainsi permettre l'amélioration du pilotage du portefeuille de Generali.

Dans le contexte actuel d'ouverture de données publiques de plus en plus nombreuses et variées, l'utilisation d'Open Data est envisagée afin d'affiner l'étude de l'inflation de la sinistralité du portefeuille de santé individuelle de Generali. La base Open DAMIR, contenant les remboursements de l'assurance maladie tous régimes confondus, est utilisée à cet effet. Ces données présentent l'avantage de contenir des caractéristiques supplémentaires relatives aux prestations. La base de données regroupe également un volume d'actes important, pouvant permettre une étude de l'inflation segmentée tout en conservant une certaine pertinence de l'inflation mesurée. L'objectif est d'exploiter les informations supplémentaires issues de cette base afin d'en extraire des enseignements permettant d'améliorer la compréhension de l'inflation du portefeuille. L'inflation est alors mesurée à partir des données de l'assurance maladie obligatoire et des variables l'influencant sont recherchées. Le suivi des variables identifiées comme influentes permettra d'expliquer l'inflation observée sur le portefeuille de Generali.

Le premier chapitre de ce mémoire présente le système de prise en charge des dépenses de santé en France. Il détaille son fonctionnement ainsi que les récentes réformes qui l'ont affecté et ont impacté l'étude réalisée. Ce chapitre présente également la notion de dérive des dépenses de santé et la base Open DAMIR.

Le traitement de cette base de données fait l'objet du deuxième chapitre. L'inflation utilisée dans la revalorisation correspond à l'évolution du coût moyen des dépenses de santé hors effet des évolutions réglementaires passées. Des retraitements sont alors effectués afin de mesurer l'inflation réelle. L'objectif étant ensuite de pouvoir utiliser les observations issues de ces données sur le portefeuille, l'inflation mesurée doit être comparable à celle observée sur les données de Generali. Un portefeuille fictif est alors constitué à partir de la base Open DAMIR en tenant compte de la nature des prestations couvertes par une complémentaire santé et du profil de risque du portefeuille de Generali. Ce chapitre présente l'ensemble des traitements réalisés. La sélection des données et l'étude de la qualité des données sont tout d'abord mentionnées. Les étapes de préparation de la base pour la mesure de l'inflation réelle ainsi que les traitements réalisés dans le but d'approcher le portefeuille de Generali sont ensuite exposés.

Le troisième chapitre est consacré à la présentation de l'inflation mesurée ainsi qu'à la définition de l'inflation étudiée et des variables explicatives. L'inflation calculée au global sur la base d'étude construite est observée, puisqu'elle permet de positionner le portefeuille de Generali par rapport aux données nationales. L'inflation est ensuite calculée en segmentant les données afin d'être analysée. Cette segmentation est définie, puis les variables explicatives issues de la base Open DAMIR et de données externes sont introduites.

Le dernier chapitre aborde la méthode utilisée pour analyser l'inflation et rappelle des éléments théoriques avant de présenter les résultats obtenus. La méthode des forêts aléatoires est employée pour établir une relation entre l'inflation et les différentes variables explicatives et identifier des variables influentes. Cette méthode est retenue pour sa capacité à prendre en compte les relations non linéaires ou non monotones, pour sa quantification de l'importance des variables mais aussi pour son caractère automatique. Les modèles sont interprétés principalement à l'aide de graphiques d'importance des variables et de la méthode SHAP. Cette approche est comparée avec un modèle linéaire, plus traditionnel mais reposant sur plusieurs hypothèses. Les résultats obtenus sont présentés pour deux familles de remboursements. Ils permettent d'identifier des nouvelles pistes d'analyse de l'inflation sur le portefeuille de santé individuelle de Generali.

Chapitre 1

La prise charge des dépenses de santé en France

L'assurance santé individuelle s'inscrit dans le système français de prise en charge des dépenses de santé. La composition de ce système et son fonctionnement sont présentés afin de contextualiser l'étude. La compréhension du système de remboursement des dépenses de santé est également indispensable afin de manipuler correctement les données de l'assurance maladie utilisées dans ce mémoire.

1.1 Le fonctionnement du système de santé

Le système d'assurance maladie français est composé de deux types de régimes : les régimes d'assurance maladie obligatoire de base et les régimes complémentaires. L'État s'occupe de l'assurance santé de base couverte par la Sécurité Sociale tandis que des organismes d'assurance se chargent de la complémentaire santé. Les français bénéficient ainsi d'une double prise en charge des dépenses, leur procurant un taux de remboursement relativement élevé sur les différents postes de soins.

1.1.1 Les régimes du système de santé

Les régimes obligatoires de base

Les régimes d'assurance maladie de base permettent à la population de bénéficier d'une couverture santé pour un ensemble de services, comprenant notamment les consultations médicales, les examens, les soins hospitaliers et les médicaments. L'assurance maladie obligatoire est souvent confondue avec la Sécurité Sociale. En réalité, les régimes d'assurance maladie de base sont gérés par la Sécurité Sociale, mais cette dernière ne se compose pas uniquement d'une branche maladie.

La Sécurité sociale est un service public de l'Etat offrant une couverture de premier niveau face aux risques sociaux. Elle se compose de cinq branches, c'est-à-dire de cinq entités ayant en charge un ou plusieurs risques. Les branches sont les suivantes : Maladie , Accidents du travail et maladies professionnelles, Famille, Retraite et Recouvrement.

La Sécurité Sociale se divise en plusieurs régimes obligatoires. Les assurés sont affiliés à l'un de ces régimes en fonction de leur situation professionnelle. Il existe deux régimes principaux et des régimes spéciaux, se caractérisant par des modalités de gestion et de prise en charge différentes.

- Le régime général prend en charge les travailleurs salariés et indépendants ainsi que les personnes bénéficiant de droits au titre de la résidence, soit la majorité de la population. La gestion des branches Maladie et Accidents du travail et maladies professionnelles de ce régime est assurée par la CNAM (Caisse Nationale de l'Assurance Maladie).
- Le régime agricole prend en charge les exploitants et salariés agricoles. La MSA (Mutualité Sociale Agricole) est chargée de la gestion de l'assurance maladie de ce régime.
- Des régimes spéciaux existent pour certaines entreprises ou branches professionnelles. C'est notamment le cas de la SNCF, de la RATP, des mines, des marins, de l'Assemblée nationale, du Sénat, des clercs et employés de notaires.

Il convient de mentionner que l'Alsace-Moselle bénéficie d'un régime particulier d'assurance maladie. Ce régime est indépendant et intervient en complément du régime général. Les assurés sont mieux remboursés en contrepartie de cotisations plus élevées. Cette assurance maladie a été mise en place par l'Allemagne pendant la guerre, puisque durant cette période la région était annexée à l'Allemagne.

Les régimes complémentaires

Les régimes complémentaires permettent de couvrir tout ou partie des dépenses laissées à la charge des bénéficiaires des soins par les régimes de base. Ils peuvent notamment prendre en charge des prestations qui ne sont pas du tout remboursées par l'assurance maladie obligatoire. Plus de 95% de la population française est couverte par l'assurance complémentaire santé.

Trois types d'organismes se partagent le marché de l'assurance complémentaire santé : les mutuelles, les sociétés d'assurance et les institutions de prévoyance. Les mutuelles sont le premier acteur de ce marché. Elles regroupent 51% du chiffre d'affaires total, tandis que les sociétés d'assurances et les institutions de prévoyance concentrent respectivement 29% et 20% du chiffre d'affaire [9].

Il existe essentiellement deux types de contrats : les contrats individuels et les contrats collectifs. Comme son nom l'indique, un contrat individuel est souscrit à titre individuel et ne comprend que deux signataires : l'assuré et l'assureur. Ce type d'assurance s'adresse principalement aux fonctionnaires, aux étudiants, aux chômeurs, aux retraités, aux indépendants et aux salariés du secteur privé qui souhaitent souscrire une surcomplémentaire individuelle. Les contrats collectifs sont au contraire souscrits par les entreprises au profit de leurs salariés. Ce type de contrat peut être à adhésion obligatoire ou facultative. Depuis 2016, toutes les entreprises du secteur privé ont l'obligation de proposer une assurance complémentaire santé collective à leurs salariés. L'adhésion y est obligatoire, sauf en cas de dispense. Dans le but de compléter les garanties proposées par un contrat à adhésion obligatoire, les entreprises peuvent proposer des contrats collectifs à adhésion facultative. Ces contrats, aussi appelés contrats surcomplémentaires, sont souscrits par l'employeur mais le salarié peut décider d'en bénéficier ou non.

Les dispositifs d'aide à la complémentaire santé

En parallèle, deux dispositifs d'aide à la complémentaire santé ont été mis en place par l'Etat : la CMU-C (Couverture Maladie Universelle Complémentaire) et l'ACS (Aide à la Complémentaire Santé). Ils permettent aux ménages les plus modestes d'accéder à une complémentaire santé. Ces dispositifs sont financés par le Fonds CMU. Les ressources de ce Fonds sont principalement constituées par la taxe de solidarité additionnelle, prélevée sur les contrats des complémentaires santé par les organismes complémentaires.

La CMU-C est une complémentaire santé gratuite attribuée sous condition de ressources et de résidence. Ce dispositif est géré, au choix du bénéficiaire, soit par sa caisse d'assurance maladie, soit par l'un des organismes proposant la CMU-C. Le Fonds CMU rembourse les dépenses de la CMU-C prises en charge par les organismes gestionnaires. L'ACS est une aide financière pour le paiement des cotisations d'un contrat de complémentaire santé s'adressant aux ménages ayant des revenus modestes mais supérieurs au seuil d'éligibilité à la CMU-C. Les bénéficiaires de l'ACS doivent choisir un contrat parmi une liste sélectionnée par le gouvernement.

La Complémentaire santé solidaire a été créée au 1er novembre 2019 en remplacement de la CMU-C et de l'ACS. Il s'agit d'une aide au paiement des dépenses de santé qui, en fonction des ressources, est soit gratuite soit coûte moins de 1€ par jour par personne. La mise en place de la Complémentaire santé solidaire n'entraîne pas de changement pour les bénéficiaires de la CMU-C. En revanche, le fonctionnement de l'ACS est modifié. Les bénéficiaires n'ont plus à choisir entre les différents contrats proposés. Ils ont désormais droit aux prestations de la Complémentaire santé solidaire, qui est une couverture santé plus couvrante que celle de l'ACS.

1.1.2 Les modalités de remboursement

L'assurance maladie de base intervient sur un large panier de soins et laisse une partie des dépenses à la charge des patients. L'assurance complémentaire santé intervient en complément de l'assurance maladie obligatoire.

Les prestations remboursables par l'assurance maladie obligatoire

Le remboursement par l'assurance maladie obligatoire de tout acte ou prestation réalisé par un professionnel de santé est subordonné à son inscription sur la liste des actes et des prestations (LAP) définie par l'article L.162-1-7 du Code de la Sécurité sociale. Cette liste est composée de trois nomenclatures :

- la NGAP (Nomenclature Générale des Actes Professionnels),
- la CCAM (Classification Commune des Actes Médicaux),
- la NABM (Nomenclature des Actes de Biologie Médicale).

Chaque acte possède un code auquel est associé un tarif qui permet de rémunérer les professionnels de santé.

La NGAP est la classification pour les actes techniques en secteur libéral et hospitalier qui existait avant la création de la CCAM. Une partie de la NGAP reste toutefois en vigueur pour un certain nombre d'actes, notamment pour les actes cliniques médicaux et les actes des chirurgiens-dentistes, des sages-femmes et des auxiliaires médicaux. Elle leur attribue un code appelé lettre-clé, correspondant au type d'acte, ainsi qu'un libellé d'acte. La NGAP comprend également un coefficient indiquant la valeur relative de chaque acte professionnel, qui permet de différencier différentes prises en charge pour un même type d'acte. Par exemple, pour les actes relevant de la lettre-clé C, dont le libellé est "consultation par le médecin généraliste, le chirurgien-dentiste omnipraticien ou la sage-femme", le coefficient 1 est associé à une consultation simple tandis que le coefficient 2 désigne un avis ponctuel de spécialiste.

La CCAM est la nomenclature qui a remplacé la NGAP pour les actes techniques réalisés par les médecins. En plus d'identifier les actes, la codification CCAM permet de fixer leurs honoraires. Elle recense actuellement plus de 7000 codes alphanumériques. Chaque code permet de désigner l'appareil, l'organe, l'action effectuée et la technique utilisée.

La NABM est la classification qui contient toutes les informations sur les actes de biologie remboursés par la Sécurité Sociale.

La LPPR (Liste des Prestations et Produits Remboursables), prévue à l'article L-165-1 du Code de la Sécurité Sociale, est une nomenclature qui répertorie les dispositifs médicaux destinés au diagnostic et au traitement des maladies ou des blessures, les matériels d'aide à la vie quotidienne, les orthèses et prothèses externes, les dispositifs implantables, les véhicules pour les handicapés physiques, etc. A chaque produit de la LPPR est attribué une codification, une description précise du produit, son utilité, la base de remboursement appliquée ainsi que le prix de vente autorisé.

Le mécanisme de remboursement

Pour chaque acte remboursé, il existe une assiette de remboursement, appelée base de remboursement de la Sécurité Sociale. Le remboursement de l'assurance maladie obligatoire est calculé à partir de cette base de remboursement et d'un taux de remboursement.

La différence entre la base de remboursement et le remboursement de la Sécurité Sociale est appelée ticket modérateur. L'assurance complémentaire santé intervient sur cette partie de la dépense.

Les dépenses de santé peuvent comporter un reste à charge lié à la liberté tarifaire. Des dépassements d'honoraires sont notamment appliqués par certains professionnels de santé en supplément de la base de remboursement de la Sécurité Sociale. Certains produits ou actes possèdent en outre des tarifs libres, comme par exemple en optique. Ces montants peuvent être partiellement ou totalement couverts par la complémentaire santé, selon le niveau de couverture de l'assuré. Ainsi, un reste à charge peut éventuellement subsister.

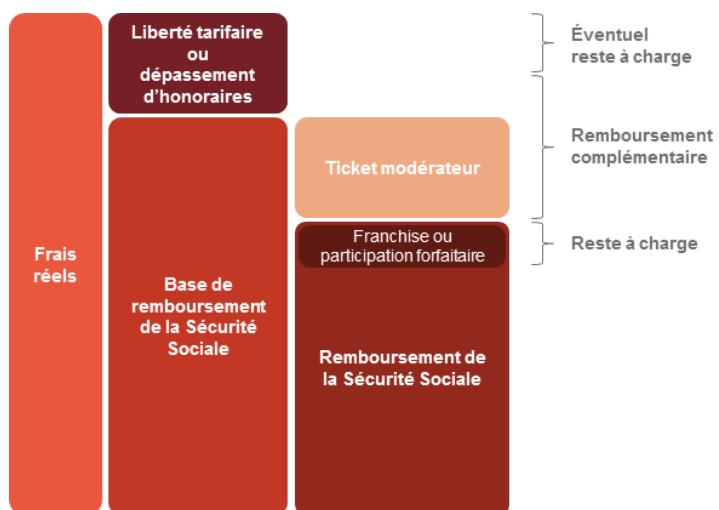


FIGURE 1.1 – Décomposition des remboursements

Pour certains actes, une participation financière est demandée aux patients. Elle est déduite du montant remboursé par la Sécurité Sociale. Pour les consultations ou actes pratiqués par un médecin, les examens radiologiques et les analyses de biologie médicale, une participation forfaitaire égale à 1€ est laissée à la charge des patients. De la même manière, une franchise médicale s'applique sur les boîtes de médicaments, les actes paramédicaux et les transports.

Il est à noter que pour les actes dont le tarif est supérieur ou égal à 120€ dans la CCAM et ceux ayant un coefficient supérieur ou égal à 60 dans la NGAP, le ticket modérateur est remplacé par une participation forfaitaire de 24€.

Par ailleurs, l'assurance complémentaire santé peut prendre en charge des soins non remboursés par la Sécurité Sociale. C'est notamment le cas des médecines douces comme l'ostéopathie ou l'acupuncture ainsi que de certains frais d'optique dont la plupart des lentilles, la chirurgie de la myopie, etc.

1.1.3 Le niveau des remboursements

L'assurance maladie de base intervient avec des niveaux de remboursements différents en fonction du type de prestation. Le niveau de prise en charge peut également être modifié lorsque les patients ou les professionnels de santé se trouvent dans certaines situations particulières. Le remboursement de l'assurance complémentaire santé dépend ensuite du reste à charge laissé aux assurés par l'assurance maladie obligatoire et de leur niveau de couverture.

Les remboursements de l'assurance maladie obligatoire

Le niveau de remboursement de l'assurance maladie obligatoire diffère selon la nature de l'acte. Le tableau ci-dessous présente les principaux taux de remboursement s'appliquant au régime général.

Prestation	Taux de remboursement
Médecins, praticiens	70%
Auxiliaires médicaux	60%
Médicaments	15%, 30%, 65%, 100%
Optique	60%
Dentaire	70%
Auditif, orthopédie	60%
Hospitalisation	80%
Frais de transport	65%
Cures thermales	65%, 70%

TABLE 1.1 – Principaux taux de remboursements du régime général

Il convient de remarquer que les médicaments ne sont pas tous remboursés au même niveau. Ils font l'objet d'un classement en plusieurs catégories, déterminé en fonction de leur SMR (Service Médical Rendu). Un médicament ayant un SMR faible possède un taux de remboursement égal à 15%, un médicament avec un SMR modéré est remboursé à hauteur de 30%, un médicament au SMR important à un taux de remboursement égal à 65% et un médicament reconnu comme irremplaçable et coûteux est remboursé à hauteur de 100%.

Les prises en charge particulières

Dans certains cas particuliers, la Sécurité Sociale peut rembourser les dépenses de santé à hauteur de 100% de la base de remboursement. Certains patients peuvent aussi être exonérés de participation financière. Voici deux exemples de situations particulières de prises en charge.

- Les patients souffrant d'ALD (Affection de Longue Durée)

Une ALD est une maladie chronique à caractère grave, qui peut être "exonérante", c'est-à-dire prise en charge à 100% par l'assurance maladie obligatoire. Il s'agit d'une affection nécessitant un traitement prolongé et coûteux, ce qui justifie la suppression du ticket modérateur.

- Les bénéficiaires de la CMU-C

Outre le fait que la complémentaire santé soit gratuite pour les bénéficiaires de la CMU-C, ces derniers bénéficient de prises en charge particulières. Ils sont notamment exonérés de la participation forfaitaire de 1€ et de la franchise appliquée sur les médicaments, les actes paramédicaux et les transports sanitaires. Le ticket modérateur est intégralement remboursé sur tous les soins courants et les dépassements d'honoraires sont interdits.

Les facteurs impactant le remboursement

Certains facteurs influent sur le remboursement de l'assurance maladie. Il s'agit par exemple du respect ou non du parcours de soins coordonnés et du secteur auquel appartient le médecin (secteur conventionné ou non).

- Le parcours de soins coordonnés

Le parcours de soins est un processus qui impose aux assurés de la Sécurité Sociale de déclarer un médecin traitant et de le consulter en priorité. Ce médecin traitant doit aussi être consulté avant toute consultation chez un spécialiste. Son rôle est d'assurer la coordination des soins et le suivi médical des patients. Le non respect du parcours de soins entraîne l'application d'une pénalité financière, correspondant à un abaissement de 40% du taux de remboursement de la Sécurité sociale. Pour de nombreux actes, le remboursement passe alors de 70% à 30%. L'assurance complémentaire ne couvre pas la pénalité financière appliquée par le régime obligatoire.

Il convient de noter qu'il existe des exceptions au parcours de soins. Certaines spécialités sont notamment accessibles directement sans passer par le médecin traitant. C'est par exemple le cas des gynécologues et des psychiatres.

- Le conventionnement des médecins

Les honoraires des médecins et le montant qui sert de base de remboursement à la Sécurité Sociale varient en fonction de la discipline du médecin mais aussi en fonction de son conventionnement ou non et de son secteur d'activité (secteur 1 ou 2).

Les médecins conventionnés de secteur 1 appliquent le tarif fixé par la convention nationale, sans dépassement d'honoraires. Ce tarif correspond à la base de remboursement de la Sécurité Sociale, qui est notamment égale à 25€ pour les médecins généralistes de secteur 1.

Dans le secteur 2, le prix des consultations varie car les honoraires des médecins sont libres. Les médecins de secteur 2 sont néanmoins signataires de la convention médicale, qui leur impose de fixer leurs tarifs avec "tact et mesure". Leurs consultations sont prises en charge sur la base du tarif de référence de la Sécurité sociale, notamment égal à 23€ pour les médecins généralistes de secteur 2. Les médecins peuvent en outre adhérer à l'OPTAM (qui sera présentée dans la partie 1.2.4) pour assurer à leurs patients que leurs dépassements d'honoraires sont modérés et stables. Ils sont ainsi remboursés sur la même base de remboursement que les médecins conventionnés de secteur 1, ce qui diminue le dépassement d'honoraires.

Certains médecins sont dits non conventionnés ou de secteur 3 car ils n'ont pas signé la convention médicale. Ils pratiquent des tarifs totalement libres, avec dépassement d'honoraires. Le remboursement Sécurité Sociale de leurs consultations est minime.

Les remboursements de l'assurance complémentaire santé

Il existe différents niveaux de couverture pour les contrats d'assurance complémentaire. En fonction de ce niveau, les contrats peuvent être classés en trois catégories : les contrats d'entrée de gamme, de milieu de gamme et de haut de gamme.

Les remboursements associés sont généralement présentés sous forme d'une grille de garantie, indiquant le taux de remboursement s'appliquant à chaque acte médical. Ces taux peuvent s'exprimer de différentes manières, en particulier :

- en pourcentage de la base de remboursement de la Sécurité Sociale,
- en pourcentage du montant de frais réels,
- en montant forfaitaire.

Le remboursement exprimé sous forme de forfait est souvent utilisé pour le remboursement des soins dont la base de remboursement de la Sécurité Sociale est très faible par rapport à la dépense moyenne réelle, comme par exemple en optique.

Il est à noter que le remboursement total versé par tous les organismes ne peut jamais excéder les frais réels engagés par l'assuré. L'assuré ne peut donc pas s'enrichir grâce à sa consommation de soins.

1.2 Les réformes marquantes du système de santé

Plusieurs réformes ont fait évoluer le système de santé français depuis quelques années. Ces réformes ont un triple objectif : responsabiliser financièrement les patients, assurer un socle de prise en charge minimal et contenir les prix pratiqués par les professionnels de santé.

1.2.1 L'accord National Interprofessionnel de 2013

L'ANI (Accord National Interprofessionnel) de 2013 est un accord portant sur les conditions de travail et les garanties sociales dont peuvent bénéficier les salariés au sein des entreprises. Il a été adopté par le parlement le 14 juin 2013 et est entré en vigueur au 1er janvier 2016. Cet accord a instauré la généralisation de la complémentaire santé dans les entreprises du secteur privé. Depuis le 1er janvier 2016, les employeurs du secteur privé ont l'obligation de proposer à leurs salariés une complémentaire santé collective et de participer à son financement. Les garanties doivent de plus respecter un panier de soins dit ANI, comprenant certaines prises en charge minimales.

Cette réforme a eu pour impact d'accroître la part de salariés couverts par une complémentaire santé collective. Cette évolution s'explique en partie par les salariés nouvellement couverts, mais principalement par un transfert des salariés couverts par une couverture complémentaire individuelle vers une complémentaire collective.

1.2.2 Le contrat responsable

La notion de contrat responsable concerne l'assurance complémentaire santé et s'applique aux contrats collectifs et individuels. Elle permet à un contrat de complémentaire santé de bénéficier d'aides fiscales et sociales à condition de respecter un cahier des charges précis. Un contrat responsable doit inciter l'assuré à avoir une attitude responsable au regard de ses dépenses de santé.

Le contrat responsable est entré en vigueur au 1er janvier 2006 mais a depuis fait l'objet d'un renforcement de son cahier des charges. Les dispositions du nouveau contrat responsable ont été définies par le décret du 18 novembre 2014 et sont entrées en vigueur à compter du 1er avril 2015.

Depuis 2006, un contrat responsable doit respecter des minima de remboursements sur certaines prestations et à l'interdiction de prendre en charge certaines franchises ou majorations. Par exemple, il ne doit pas rembourser la participation forfaitaire de 1€ et la franchise sur les médicaments. Avec le décret de 2015, un nouveau panier minimal des garanties est demandé pour qu'un contrat soit considéré comme responsable.

Des plafonds de prise en charge sont aussi mis en place pour certains postes de soins. Par exemple, un encadrement de la prise en charge des dépenses d'optique est mis en place grâce à des plafonds et des planchers de remboursement mais aussi grâce à la limitation au remboursement d'un équipement tous les deux ans (sauf pour les mineurs pour lesquels un équipement peut être remboursé tous les ans en cas d'évolution du besoin de correction). Parmi ces dispositions se trouve également l'obligation de couvrir l'intégralité de la participation de l'assuré pour l'ensemble des dépenses de santé (sauf pour les frais de cure thermale, les médicaments dont le service médical rendu a été classé faible ou modéré et l'homéopathie). Le dispositif complet est décrit dans les articles L. 871-1, R. 871-1 et R. 871-2 du code de la Sécurité Sociale et dans le décret n° 2014-1374 du 18 novembre 2014.

1.2.3 La convention médicale 2016-2021

La convention médicale 2016-2021 est un accord signé entre les syndicats de médecins libéraux et l'Union Nationale des Caisses d'Assurance Maladie le 25 août 2016. Elle définit les règles et obligations respectives de l'assurance maladie et des médecins qui exercent en cabinet, en maison de santé ou en clinique privée, pour une durée de cinq ans. Plusieurs dispositifs sont mis en place par cette nouvelle convention médicale afin de répondre à deux objectifs clés : renforcer l'accès aux soins et améliorer la prise en charge médicale. Les principales mesures adoptées sont les suivantes :

- Au 1er mai 2017, le tarif de la consultation des médecins de secteur 1 et des médecins de secteur 2 ayant signé un contrat destiné à limiter leurs dépassements d'honoraires évolue de 2€, passant à 25€ contre 23€ auparavant.
- La consultation coordonnée est revalorisée de 28€ à 30€ au 1er juillet 2017. Au 1er juin 2018, les avis ponctuels de consultant sont aussi revalorisés de 2€.
- De nouveaux tarifs sont mis en place pour mieux prendre en charge les consultations complexes et très complexes, demandant plus de temps et d'attention. Par exemple, le suivi d'une sclérose en plaque entre dans le champs des consultations complexes, tandis que la mise en place d'un dossier de greffe fait l'objet d'une consultation très complexe.
- Le dispositif de maîtrise des dépassements d'honoraires, autrefois appelé CAS (Contrat d'Accès aux Soins), est rénové. Tous les médecins peuvent signer le nouveau dispositif appelé OPTAM (Option Pratique Tarifaire Maîtrisée), y compris désormais les chirurgiens et les obstétriciens. Cette nouvelle option repose sur le même principe que le CAS mais y apporte des adaptations permettant d'améliorer son attractivité auprès des médecins. Avec l'OPTAM, ces derniers bénéficient notamment d'une prime qui valorise leur activité réalisée à tarif opposable.
- A partir de 2018, un forfait patientèle unique est créé en remplacement des différents forfaits octroyés par l'assurance maladie obligatoire pour compléter la rémunération des médecins généralistes.

1.2.4 L'Option Pratique Tarifaire Maîtrisée

L'OPTAM (Option Pratique Tarifaire Maîtrisée) et l'OPTAM-CO (Option Pratique Tarifaire Maîtrisée Chirurgie et Obstétrique) sont deux options à adhésion facultative applicables aux médecins de secteur 2. Plus particulièrement, l'OPTAM-CO concerne les médecins exerçant une spécialité de chirurgie ou une spécialité de gynécologie-obstétrique. Ces options ont été mises en place au 1er janvier 2017, en remplacement du contrat d'accès aux soins (CAS) entré en vigueur en décembre 2013.

L'objectif principal de ces options est d'améliorer l'accès aux soins en limitant les dépassements d'honoraires et en favorisant l'activité à tarifs opposables. En souscrivant à l'une de ces options conventionnelles, les praticiens s'engagent à respecter un taux moyen de dépassement et un taux moyen d'activité facturée sans dépassement. En contrepartie, une rémunération spécifique est mise en place pour les médecins ayant respecté les engagements de l'option souscrite. L'assurance maladie s'engage également à faire bénéficier les médecins qui adhèrent à l'une des options des tarifs de remboursement en vigueur pour le secteur 1. La signature de l'OPTAM permet ainsi aux patients du praticien d'obtenir une meilleure prise en charge par l'assurance maladie obligatoire, puisque la base de remboursement de certains actes en secteur 1 est plus élevée qu'en secteur 2, mais également par sa complémentaire santé responsable.

Le taux de dépassement et le pourcentage d'activité réalisée à tarif opposable sur lesquels s'engage un médecin sont fixés au regard de sa pratique tarifaire sur les trois années précédant l'entrée en vigueur de la convention médicale, soit 2013, 2014 et 2015¹. Il convient de noter que ces taux sont recalculés, ce qui signifie qu'ils ont pour base le tarif de remboursement opposable au secteur 1 et non le tarif opposable au secteur 2. Le taux de dépassement recalculé ne doit pas être supérieur à 100%.

Lorsque des évolutions des tarifs de remboursement ont lieu, le taux de dépassement et la part d'activité réalisée à tarif opposable sont recalculés en fonction des nouveaux tarifs. Par exemple, pour un médecin généraliste de secteur 2 adhérent à l'OPTAM qui facture sa consultation 30€, le dépassement est de 7€ avant le 1er mai 2017 puisque la base de remboursement de la Sécurité Sociale est de 23€. Le taux de dépassement moyen est alors de 30,4% ($\frac{30-23}{23}$). Au 1er mai 2017, le tarif de remboursement est augmenté à 25€. Le taux de dépassement moyen autorisé n'est alors plus que de 20% ($\frac{30-25}{25}$). Le médecin adhérent à l'option peut toujours facturer 30€, mais il ne peut plus appliquer un dépassement de 30,4% sur les 25€ de la base de remboursement qui mettrait sa consultation à 32,60€.

L'option repose donc sur la stabilité du montant de l'acte et non sur la stabilité du montant du dépassement, ce qui engendre une diminution du taux de dépassement au fil des évolutions des bases de remboursement.

1. Si le médecin n'a pas d'activité pour l'année précédant l'entrée en vigueur de la convention, les indicateurs sont calculés sur la moyenne des médecins éligibles de la même spécialité dans leur région (ou au niveau national pour certaines spécialités).

1.2.5 La réforme 100% santé

Mise en place à partir de 2019, la réforme 100% santé a pour principal objectif de réduire le renoncement aux soins en optique, audiologie et dentaire [11].

Ces trois secteurs de la santé étant généralement peu remboursés, notamment en raison des tarifs fixés librement et ne correspondant pas toujours aux bases de remboursement de la Sécurité Sociale, certains ménages sont dans l'impossibilité de se soigner. L'objectif de la réforme est alors de permettre à tous les assurés disposant d'une complémentaire santé d'accéder à une offre de soins de qualité sans reste à charge après l'intervention de l'assurance maladie obligatoire et de la complémentaire santé.

Le principe est le même pour les trois secteurs. Des paniers de soins sont définis :

- Panier reste à charge zéro "RAC 0" :
Ce panier de soins est valable pour les trois secteurs concernés par la réforme, à savoir optique, audiologie et dentaire. Les soins de ce panier sont totalement remboursés, ne laissant pas de reste à charge aux assurés.
- Panier "RAC modéré" :
Ce panier s'applique uniquement pour les soins dentaires. Les dépenses associées sont partiellement prises en charge par l'assurance maladie de base et la complémentaire santé.
- Panier "honoraires libres" :
Ce panier concerne à nouveau les trois secteurs concernés par la réforme. Il correspond à tous les soins qui ne sont pas présents dans les paniers "RAC 0" et "RAC modéré". Le reste à charge associé aux soins de ce panier peut être élevé.

La réduction du reste à charge, que ce soit pour le panier "RAC 0" ou le panier "RAC modéré", est obtenue grâce à la mise en place de plafonds limites de vente. Certains soins sont de plus revalorisés au moyen d'une évolution de leur base de remboursement.

Avec la mise en place de ces paniers, les nomenclatures CCAM et LPP sont revues, moyennant la création de nouveaux codes prestations et une redéfinition des codifications affinées. Par exemple, la constitution de paniers d'actes prothétiques nécessite de découper certains codes CCAM en plusieurs codes. Cette révision des nomenclatures permet aussi de fixer des honoraires limites de facturation distincts.

La réforme 100% santé se déployera progressivement jusqu'en 2021. Des mesures préparatoires ont été mises en place dès 2019, comme notamment la modification des nomenclatures et le plafonnement des tarifs des prothèses dentaires. Depuis le 1er janvier 2020, le 100% santé est garanti sur les soins d'optique et sur une partie du dentaire. À partir du 1er janvier 2021, le 100% santé sera garanti en audiologie et sur le reste du panier dentaire.

1.3 La dérive des dépenses de santé

Chaque année, les dépenses de santé évoluent, que ce soit en raison de l'évolution des coûts des soins ou de l'évolution de la consommation des patients. La dérive des dépenses de santé correspond à l'évolution naturelle des prestations payées au titre des dépenses de santé entre deux années consécutives. Cette évolution impacte d'une part la Sécurité Sociale, qui tente de la maîtriser en mettant en place des stratégies de régulation du système de santé. D'autre part, les prestations versées par les organismes d'assurance complémentaire évoluent. Afin de répondre à leurs obligations de solvabilité et à leurs objectifs de rentabilité, ces organismes répercutent l'évolution sur les primes d'assurances payées par leurs assurés.

1.3.1 La maîtrise de la dérive des dépenses de l'assurance maladie

La dérive des dépenses de santé contribue au déficit de la Sécurité Sociale. Pour contenir cette dérive naturelle et ainsi réduire le déficit de l'assurance maladie, des politiques visant à mieux maîtriser l'évolution des dépenses de santé sont mises en place.

Dans le cadre de la loi de financement de la Sécurité Sociale, l'ONDAM (Objectif National des Dépenses d'Assurance Maladie) est fixé chaque année par le Parlement pour l'année à venir. Il s'agit du montant des dépenses à ne pas dépasser par l'assurance maladie, compte tenu du niveau des recettes prévues. Il a pour but d'aider à la maîtrise du système de santé mais n'est pas un plafond, ce qui signifie que les remboursements des patients ne sont pas cessés une fois que cet indicateur est atteint. Les prestations contenues dans le champ de l'ONDAM représentent l'essentiel des prestations réglées par les branches maladie et AT/MP (accidents du travail et maladies professionnelles) du régime général de la Sécurité Sociale. Il se décompose en sous-objectifs, notamment pour les soins de ville, les établissements de santé publics et privés et les établissements médico-sociaux. L'ONDAM était fixé à +2,5% en 2019 par rapport à 2018 et à +2,3% en 2020 par rapport à 2019.

Afin d'aboutir à l'objectif de dépenses fixé, des économies sont nécessaires. Chaque année, de nouvelles mesures de régulation sont mises en place. Elles s'organisent selon différents axes, comme par exemple :

- la maîtrise tarifaire des actes médicaux,
- les actions sur la pertinence et le bon usage des soins,
- la structuration de l'offre de soins,
- la diminution de certaines prises en charge par l'assurance maladie,
- la mise en place d'actions de prévention,
- le renforcement du contrôle et la lutte contre la fraude.

Certaines actions permettent de réduire les dépenses de santé en diminuant les coûts des soins. D'autres mesures visent à limiter l'augmentation de la consommation de soins afin de respecter l'objectif de dépenses.

1.3.2 La prise en compte la dérive de la sinistralité des organismes d'assurance

La dérive des dépenses de santé a également un impact sur les organismes d'assurance proposant des contrats de complémentaire santé, puisque ces derniers voient les frais de santé de leurs assurés évoluer. Afin de répondre à leurs obligations de solvabilité et à leurs objectifs de rentabilité, les compagnies d'assurance ajustent leurs tarifs chaque année en tenant compte de cette dérive de la sinistralité. L'augmentation des primes d'assurance mise en place une fois par an est généralement appelée revalorisation des contrats ou indexation tarifaire.

Lors de l'exercice annuel de revalorisation des contrats de santé, un taux de majoration est calculé. Il se compose généralement de trois parties, liées aux évolutions réglementaires, aux modifications de garanties et à la dérive naturelle des dépenses de santé.

La première composante a pour but de prendre en compte l'impact des évolutions réglementaires, telles que la modification d'une base de remboursement de la Sécurité Sociale par exemple, sur la sinistralité future. Des études sont réalisées lors de chaque modification légale afin d'estimer l'évolution des montants de prestations remboursées en résultant. La deuxième composante permet de considérer l'impact des modifications des garanties des contrats sur la sinistralité à venir. Ces modifications peuvent être de différents types, comme l'évolution d'un pourcentage de remboursement ou l'ajout d'une garantie au contrat. Cette partie du taux de majoration est à nouveau basée sur des études estimant l'impact de chaque modification sur les montants versés par l'assurance.

La dernière composante du taux de majoration est relative à la dérive naturelle de la sinistralité du portefeuille. Elle permet de revaloriser les contrats de santé conformément à l'évolution passée des frais de santé des assurés. L'indice de la dérive utilisé peut provenir d'indices de marché ou peut également provenir de la mesure de la dérive de sinistralité sur le portefeuille de santé individuelle.

Concernant les indices de marché utilisés pour la revalorisation, l'ONDAM est souvent pris en considération, notamment en raison de son impact sur la consommation de soins des patients. Des études réalisées par d'autres organismes peuvent également être prises en compte, notamment les études du BIPE (Bureau d'Information et de Prévisions Économiques). Pour que l'indice de la dérive de la sinistralité utilisé soit représentatif du portefeuille de l'assureur, il peut provenir de la dérive mesurée directement sur le portefeuille. Cette dernière correspond à la variation du montant total versé par l'assurance complémentaire santé d'une année à l'autre, à effectif constant. La dérive mesurée ne doit pas contenir les effets des évolutions passées liées à des modifications réglementaires ou à des modifications de garanties pour être utilisée dans la revalorisation. Lors de la mesure de la dérive de sinistralité sur le portefeuille, les évolutions passées sont donc retraitées.

La dérive naturelle de la sinistralité peut être décomposée en deux effets dus au comportement de consommation des patients et à l'inflation du coût unitaire des dépenses de santé.

D'une part, l'évolution du nombre d'actes consommés est responsable d'une partie de la dérive, ce qui explique notamment que certaines mesures mises en place par la Sécurité Sociale visent à limiter l'évolution de la consommation des patients. D'autre part, l'évolution du coût moyen des soins, hors évolutions réglementaires ou modifications de garanties, contribue à la dérive naturelle de la sinistralité. Cette inflation de la sinistralité peut par exemple être la conséquence d'une évolution des dépassements d'honoraires ou des prix des dispositifs médicaux. L'inflation dépend en particulier de la composition du portefeuille et un certain nombre de facteurs peuvent l'influencer. Elle est par conséquent difficile à comprendre et à anticiper. Ce mémoire s'inscrit dans ce contexte en proposant une étude de l'inflation de la sinistralité à partir d'Open Data dans le but d'améliorer la compréhension de l'inflation.

1.4 La base Open DAMIR parmi l'offre d'Open Data de l'assurance maladie

L'assurance maladie publie sur son site Internet *ameli.fr* des études de santé publique, des articles et des données statistiques. Ces publications concernent les dépenses d'assurance maladie, les professionnels de santé et la consommation de soins. Depuis quelques années, elle met aussi à disposition des bases brutes extraites du SNIIRAM dans un but d'ouverture et de partage des données, tout en garantissant la protection des données individuelles. L'un des objectifs de la mise à disposition de ces données est d'améliorer la compréhension du fonctionnement du système de soins français, grâce à l'exploitation et l'appropriation des données par le grand public.

1.4.1 Le SNIIRAM et le SNDS, des données protégées

Le SNIIRAM, Système National d'Information Inter-Régimes de l'Assurance Maladie, est un ensemble de bases de données nationales pseudonymisées relatives aux prestations remboursées par l'assurance maladie. Il s'agit d'un outil géré par la CNAMTS (Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés) pour le compte des régimes obligatoires. Il regroupe des données provenant des organismes gérant un régime de base d'assurance maladie (régime général, régime agricole, etc.) et de l'ATIH (Agence Technique de l'Information sur l'Hospitalisation). Ce regroupement de données a pour finalité l'amélioration de la qualité des soins, une meilleure gestion de l'assurance maladie et des politiques de santé ainsi que la transmission aux prestataires de soins d'informations pertinentes concernant leur activité. Les données du SNIIRAM sont regroupées avec d'autres données dans le SNDS (Système National des Données de Santé). Crée par la loi de modernisation du système de santé de 2016, ce système a pour but de centraliser et de mettre à disposition de manière sécurisée les données d'hospitalisation, les données de consommation de soins, les données liées aux décès, etc.

Les informations regroupées dans le SNDS sont considérées comme des données de santé, c'est-à-dire des données à caractère personnel considérées comme sensibles d'après la loi "Informatique et Libertés". Son accès est donc restreint. La liste des organismes disposant d'un accès permanent au SNDS est fixée par le décret n° 2016-1871 du 26 décembre 2016. Il s'agit d'organisations chargées d'une mission de service public, comme les caisses nationales des régimes d'assurance maladie obligatoire, les agences régionales de santé, l'agence nationale de santé publique, l'établissement français du sang, etc. La loi prévoit également la possibilité pour les organismes publics comme privés d'accéder aux données du SNDS pour un projet de recherche, d'étude ou d'évaluation présentant un intérêt public. Cet accès nécessite une autorisation de la CNIL (Commission Nationale de l'Informatique et des Libertés). La loi interdit l'utilisation des données contenues dans ce système à des fins de promotion de produits de santé, d'exclusion de garanties de contrats d'assurance ou de modification de primes d'assurance.

1.4.2 L'Open Data des données de santé

L'Open Data désigne "l'ouverture et le partage de données par leur mise en ligne dans des formats ouverts, en autorisant la réutilisation libre et gratuite par toute personne"². Les Open Data peuvent donc être utilisées par tous les organismes, notamment dans une démarche en lien avec des contrats d'assurance.

Des extractions du SNIIRAM sont accessibles en Open Data. Ces bases de données ont été agrégées préalablement à leur ouverture au public afin que l'identification d'un bénéficiaire ou d'un professionnel de santé soit impossible. Bien que la contrainte de l'anonymisation soit une limite à l'utilisation des données, il s'agit d'une composante incontournable de l'Open Data afin de protéger la vie privée des personnes concernées. L'offre de données en Open Data de l'assurance maladie contient notamment les données Open Bio présentant des données relatives à la biologie médicale, les données Open LPP sur le thème des dispositifs médicaux inscrits sur la liste des produits et prestations, les données Open Medic concernant les médicaments délivrés en officine de ville et la base Open DAMIR relative aux dépenses d'assurance maladie.

La base Open DAMIR est la version Open Data de la base DAMIR, qui est accessible via le SNIIRAM pour les utilisateurs habilités. Elle a été publiée en 2015 par l'assurance maladie, conformément aux préconisations de la commission "Open Data en santé", qui avait pour mission de débattre des enjeux en matière d'accès aux données de santé. Avant sa publication, un hackathon avait été organisé [6] afin d'expérimenter son ouverture en Open Data. Il s'agit d'une compétition durant laquelle des développeurs réalisent un projet de programmation informatique dans un temps limité et sur un thème précis, la base Open DAMIR ici. L'expérience s'était montrée concluante et avait permis de mettre en lumière la diversité des utilisateurs et des usages potentiels.

2. Définition de l'Open Data issue du rapport de la commission "Open Data en santé".

La base Open DAMIR a été le premier jeu de données sources mis en ligne par l'assurance maladie. Plusieurs bases ont par la suite été publiées et cette dynamique devrait se poursuivre par la mise à disposition auprès du grand public de données plus nombreuses et plus variées, contribuant ainsi à répondre à l'objectif d'amélioration de la compréhension du fonctionnement du système de soins français.

1.4.3 La base Open DAMIR

La base Open DAMIR, Dépenses d'Assurance Maladie Inter-Régimes, regroupe l'ensemble des remboursements de l'assurance maladie, à l'exception d'une partie des prestations hospitalières du secteur public puisque la base contient seulement les prestations hospitalières facturées directement à l'assurance maladie. Elle est téléchargeable en ligne sur le site Internet de l'assurance maladie[1] sous la forme de bases mensuelles, chaque base représentant environ 5 Go.

Les remboursements contenus dans ce jeu de données concernent tous les régimes pour la période de 2009 à 2019. Ils sont décrits par 55 variables, dont 13 indicateurs : montant total de la dépense, dépassement, base de remboursement, montant remboursé, quantité, dénombrement et coefficient global. Les informations présentes pour chaque remboursement concernent l'acte médical associé ainsi que les 3 acteurs intervenant dans le remboursement de l'acte.

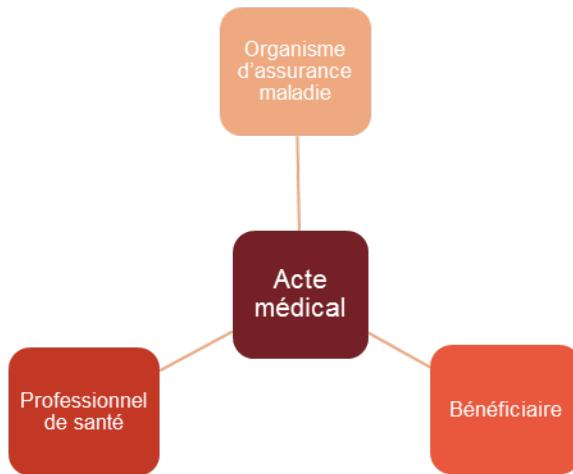


FIGURE 1.2 – Les 3 acteurs du remboursement d'un acte médical

Les prestations médicales sont décrites notamment grâce à la nature de la prestation, la nature d'assurance et le complément d'acte. Concernant les bénéficiaires, leur sexe, leur âge, leur qualité de bénéficiaire et leur lieu de résidence sont présents dans la base. L'organisme d'assurance maladie est quand à lui caractérisé par sa localisation.

En ce qui concerne le professionnel de santé, il convient de souligner qu'un acte médical est en général lié à deux professionnels de santé, puisqu'il y a toujours un exécutant et souvent un prescripteur. Par exemple, un acte infirmier de soins, tel qu'un prélèvement sanguin, est prescrit par un médecin et exécuté par un infirmier. Des informations relatives aux professionnels de santé exécutants et prescripteurs sont disponibles dans la base Open DAMIR, comme par exemple leur spécialité médicale, leur statut juridique, leur localisation. Lorsque le professionnel est un établissement de santé, il est aussi décrit par le type d'établissement et la catégorie. Enfin, la période est un axe d'analyse des remboursements présent dans les données grâce à l'année et le mois de remboursement ainsi que l'année et le mois de soins.

Avant sa mise en Open Data, la base DAMIR a été agrégée mais aussi retraitée afin d'exclure certaines données considérées comme trop discriminantes pour conserver l'anonymat des professionnels de santé et des bénéficiaires des soins. Les localisations sont notamment limitées à 13 zones géographiques correspondant, à quelques regroupements près, aux régions françaises. De même, les âges sont regroupés en tranches d'âges. Il est ainsi impossible de ré-identifier les individus.

Chapitre 2

L'utilisation de données en Open Data pour étudier l'inflation

La base Open DAMIR contient les montants remboursés par l'assurance maladie, les bases de remboursement associées aux actes remboursés ainsi que les montants des frais réels. Il est alors possible de calculer les montants que rembourseraient une assurance complémentaire santé si ces prestations concernaient ses assurés.

Dans le but d'étudier l'inflation sur des données marché, un portefeuille Generali est reconstitué à partir de la base Open DAMIR. Il est construit de manière à se rapprocher de la typologie du portefeuille Generali (âge, sexe et région des assurés) et des garanties proposées. L'objectif est de pouvoir utiliser les enseignements obtenus en étudiant ce portefeuille fictif sur les données de Generali.

2.1 Étude de la base Open DAMIR

La compréhension des données utilisées est indispensable afin de les utiliser correctement et d'être en mesure de leur faire confiance. Pour obtenir des résultats fiables, les données doivent aussi être cohérentes. Les contrôles de qualité des données permettent d'identifier les données jugées de mauvaise qualité afin de les fiabiliser ou de les exclure pour qu'elles n'altèrent pas les études réalisées par la suite.

2.1.1 Description des données

Pour rappel, la base Open DAMIR est composée de 55 variables, dont 13 indicateurs de montant et de volume. Il s'agit d'une base agrégée sur l'ensemble des indicateurs, ce qui signifie qu'une ligne correspond à une combinaison des 42 variables descriptives avec sommation sur les 13 indicateurs.

La liste complète des variables est présente en Annexe A.1. Elle provient du lexique Open DAMIR disponible en ligne, contenant les libellés des variables ainsi que les libellés de leurs modalités, puisque les variables sont codées en nombres entiers. Les variables principales sont présentées dans cette partie.

Les variables descriptives, combinées pour l'agrégation

Variable	Description
SOI_ANN	Année de soins
SOI_MOI	Mois de soins
PRS_NAT	Nature de prestation Code de l'acte, associé à un libellé dans le lexique Open DAMIR. Par exemple, le code 3313 correspond au libellé "Pharmacie 65%".
PRS_Rem_TYP	Type de remboursement Permet de distinguer les prestations de référence, les compléments d'acte et les prises en charge complémentaires (ticket modérateur hors CMU, supplément Alsace Moselle, supplément hors Alsace Moselle, etc.).
CPL_COD	Complément d'acte Complète le type de remboursement pour les compléments d'acte et permet de distinguer les majorations de nuit, d'urgence, de dimanche et jour férié.
PRS_Rem_TAU	Taux de remboursement
BEN_SEX_COD	Sexe du bénéficiaire
AGE_BEN_SNDS	tranche d'âges du bénéficiaire au moment des soins Segmentation des âges en tranches de 10 ans, excepté pour les âges inférieurs à 20 ans (une seule tranche) et supérieurs à 80 ans (idem).
BEN_RES_REG	Région de résidence du bénéficiaire Champs correspondant aux régions françaises, à l'exception de deux modalités : les régions Corse et Provence-Alpes-Côte d'Azur sont regroupées, de même que les régions d'outre-mer.
PSE_ACT_CAT	Catégorie de l'exécutant Prend les modalités suivantes : établissement, médecin, fournisseur, laboratoire, dentiste, sage-femme, infirmier, masseur, pédicure-podologue, orthoptiste - orthophoniste.
PSE_ACT_SNDS	Nature d'activité du professionnel de santé exécutant Précise la catégorie de l'exécutant.
PSE_SPE_SNDS	Spécialité médicale du professionnel de santé exécutant Précise la nature d'activité lorsque le prescripteur appartient à la catégorie des médecins.

PSP_ACT_CAT	Catégorie du prescripteur
PSP_ACT_SNDS	Nature d'activité du professionnel de santé prescripteur
PSP_SPE_SNDS	Spécialité médicale du professionnel de santé prescripteur
ETE_TYP_SNDS	Type d'établissement exécutant Renseignée pour les prestations effectuées dans un établissement : lucratif, non lucratif, etc.
PRS_PPU_SEC	Code secteur Privé/Public

TABLE 2.1 – Principales variables descriptives

Les indicateurs de montant et de volume, sommés pour l'agrégation

Deux types d'indicateurs existent dans la base de données : les indicateurs bruts (prefixés par PRS_) et les indicateurs préfiltrés (prefixés par FLT_). Les principaux indicateurs sont les suivants :

Variable	Description
PRS_PAI_MNT (FLT_PAI_MNT)	Montant de la dépense (resp. Montant de la dépense préfiltré)
PRS_DEP_MNT (FLT_DEP_MNT)	Montant du dépassement (resp. Montant du dépassement préfiltré) Contient les dépassements d'honoraires.
PRS_REM_MNT (FLT_REM_MNT)	Montant versé/remboursé (resp. Montant versé/remboursé préfiltré)
PRS_REM_BSE	Base de Remboursement
PRS_ACT_QTE (FLT_ACT_QTE)	Quantité (resp. Quantité de la prestation préfiltrée)

TABLE 2.2 – Principaux indicateurs de montant et de volume

Les indicateurs préfiltrés permettent d'éviter les doubles comptages de montants ou de quantités pouvant être engendrés par la structure atypique de la base. Ils sont, comme leur nom l'indique, préfiltrés par rapport aux indicateurs bruts.

Le montant de la dépense préfiltré, le montant du dépassement préfiltré et la quantité de la prestation préfiltrée correspondent aux indicateurs bruts pour les prestations de référence uniquement. Ce sont ces indicateurs qui doivent être utilisés afin d'éviter les doubles comptages. La base de remboursement existe uniquement en indicateur brut. Elle ne doit être prise en compte que pour les prestations de référence et les compléments d'acte. Le montant versé/remboursé préfiltré concerne uniquement la part des prestations prise en charge par le régime obligatoire. Il correspond au montant versé/remboursé pour les remboursements de type prestation de référence ou complément d'acte (majoration de nuit, d'urgence, de dimanche et jour férié).

Dans le cas où une part complémentaire, non obligatoire, de l'acte est prise en charge, elle n'est pas comprise dans les indicateurs préfiltrés. C'est notamment le cas des compléments liés à la prévention et l'Alsace Moselle.

Une structure des données particulière

- Plusieurs lignes sont associées à un même acte.

Le remboursement d'un acte peut être composé de plusieurs types de remboursements. Il existe toujours un remboursement de la prestation de référence, auquel peuvent être associés des remboursements de compléments d'acte et/ou des remboursements complémentaires.

Lorsqu'un acte est concerné par plusieurs types de remboursement, il existe une ligne pour chaque type dans la base Open DAMIR. Voici un exemple simple de remboursement de consultations de médecins généralistes de janvier 2016 dans la région Grand Est.

PRS_ACT_QTE	PRS_PAIS_MNT	PRS_REM_BSE	PRS_REM_MNT	FLT_ACT_QTE	FLT_PAIS_MNT	FLT_REM_MNT	PRS_REM_TAU	PRS_REM_TYP
7	161 €	161 €	112,70 €	7	161 €	112,70 €	70 %	Prestation de référence
7	161 €	0 €	32,20 €	0	0 €	0 €	20 %	Supplément Alsace-Moselle

FIGURE 2.1 – Exemple illustratif de la structure de la base de données

Ces deux lignes concernent les mêmes actes : il s'agit de 7 consultations dont le tarif de base est de 23€ par prestation ($7 \times 23 = 161$). Il n'y a pas de dépassement d'honoraires. Le régime de base a remboursé 112,70€ ($7 \times 23 \times 70\%$, soit 16,10€ par acte) et le régime local d'Alsace Moselle a pris en charge 32,20€ ($7 \times 23 \times 20\%$, soit 4,60€ par consultation). Ainsi, ces actes sont remboursés à 90%. Il est nécessaire de prendre en compte ces deux lignes afin d'identifier la part non remboursée par l'Assurance Maladie.

Cet exemple illustre aussi le fait que l'utilisation des indicateurs bruts engendrerait des doubles comptages. La somme des montants de dépenses non préfiltrés (PRS_PAIS_MNT) correspond au double des montants réels ; tandis que la somme des montants de dépenses préfiltrés (FLT_PAIS_MNT) correspond bien aux montants réels.

- Des montants négatifs sont présents dans la base.

La base Open DAMIR contient des montants négatifs. Ils s'expliquent de deux manières différentes.

Certains montants sont négatifs à cause de la nature de la prestation. Les participations forfaitaires de 1€, les franchises et les participations assuré de 18€/24€ sont dans cette situation.

En effet, les montants des participations et des franchises sont compris dans les remboursements des prestations leur donnant lieu, puis sont déduits sur une ligne particulière ne mentionnant que la franchise/participation en montant négatif.

Dans l'exemple illustré par la figure 2.1, le montant remboursé par le régime obligatoire est égal à 70% de la base de remboursement, soit 16,10€ par prestation. Or, s'agissant d'une consultation de médecin généraliste, une participation forfaitaire de 1€ est laissée à la charge de l'assuré et le régime obligatoire ne rembourse en réalité que 15,10€. Cette information est contenue dans la base Open DAMIR dans une ligne particulière, ayant comme nature de prestation "participation forfaitaire" et comme montant remboursé -1€ multiplié par le nombre d'actes, soit - 7€ pour cet exemple.

PRS_ACT_QTE	PRS_PA1_MNT	PRS_Rem_BSE	PRS_Rem_MNT	FLT_ACT_QTE	FLT_PA1_MNT	FLT_Rem_MNT	PRS_Rem_TAU	PRS_NAT
7	161 €	161 €	112,70 €	7	161 €	112,70 €	70 %	Consultation du généraliste
7	0 €	0 €	-7 €	7	0 €	-7 €	100 %	Participation forfaitaire

FIGURE 2.2 – Suite de l'exemple illustratif de la structure de la base de données

Les montants négatifs qui ne sont pas associés aux actes cités précédemment correspondent à des régularisations. Cependant, ces montants ne se sont pas annulés avec les montants positifs symétriques lors de l'agrégation de la base DAMIR. Ils sont donc considérés comme des anomalies de données et sont exclus de la base d'étude.

2.1.2 Sélection des données pertinentes

Bien que la base Open DAMIR soit agrégée, chaque base mensuelle représente environ 30 millions de lignes et environ 5 Go. Une sélection des variables pertinentes a donc été nécessaire afin de réduire le volume des données et de pouvoir obtenir une unique base d'étude.

Parmi les 55 variables de la base initiale, seules 35 sont jugées utiles. Ces variables ne sont pas toutes conservées dans la base d'étude. Certaines sont exploitées seulement pour la sélection des données, d'autres servent uniquement pour la construction de la base d'étude. Des variables seront aussi regroupées par la suite afin d'optimiser le volume des données.

Une sélection des données est réalisée afin de rapprocher le périmètre de la base Open DAMIR de celui du portefeuille d'assurance santé individuelle de Generali. Cette sélection permet en outre de réduire le volume des bases de données et permet ainsi de les manipuler plus facilement.

- Le périmètre est restreint à la maladie et la prévention maladie grâce à la variable "Nature d'assurance".

- La base Open DAMIR contient des actes qui ne font pas partie du périmètre de l'assurance complémentaire. C'est notamment le cas des indemnités journalières mais également des rémunérations forfaitaires et des aides financières, comme le forfait médecin traitant, la rémunération sur objectifs de santé publique, le forfait structure, le forfait d'aide à l'installation du médecin, le contrat incitatif infirmier, etc. Ces actes sont retirés de la base d'étude en utilisant le code d'acte ("PRS_NAT").
- Les actes remboursés entièrement par l'Assurance Maladie sont exclus puisque l'assurance complémentaire n'intervient pas dans leur remboursement.
- Les participations forfaitaires de 1€ et les franchises sont supprimées puisqu'elles sont incluses dans le remboursement de la sécurité sociale renseigné dans la base et ne peuvent pas être remboursées par l'assurance complémentaire dans le cadre des contrats responsables. En revanche, les participations assuré de 18€ / 24€ peuvent être prises en charge par l'assurance complémentaire santé et sont donc conservées dans la base d'étude.
- Étant donné que la CMU-C prend en charge la part complémentaire des dépenses de santé de ses bénéficiaires, ces derniers ne sont pas concernés par l'assurance complémentaire santé individuelle. Ils sont donc exclus du périmètre d'étude grâce à la variable "Top bénéficiaire CMU-C", permettant d'identifier les bénéficiaires de la CMU-C.
- Les patients atteints d'une ALD ont une consommation de soins qui est régie par leur maladie et n'est donc pas représentative de la consommation du portefeuille de Generali. Les actes en rapport avec une ALD sont donc exclus de l'étude grâce à la variable "Motif d'exonération du ticket modérateur", qui permet d'identifier les actes exonérés de ticket modérateur et détaille le motif d'exonération. Cette variable permet notamment de distinguer les patients atteints d'une ALD et indique si l'acte remboursé est en rapport ou non avec l'ALD. A noter, les soins associés à des patients atteints d'une ALD mais sans rapport avec l'affection sont conservés dans la base d'étude puisque cette consommation n'est pas biaisée par l'ALD.
D'autres remboursements assortis d'un motif d'exonération du ticket modérateur sont supprimés de la base, comme les prestations remboursées à 100% au titre d'une pension militaire d'invalidité par exemple.
- Le périmètre temporel retenu s'étend de 2016 à 2019. Les données antérieures à 2016 n'ont pas été utilisées en raison de la modification du cahier des charges des contrats responsables survenue au cours de l'année 2015. Ces règles ayant pour objectif de sensibiliser l'assuré à ne pas abuser de ses garanties santé, l'inflation observée avant cette réforme n'est pas représentative de l'inflation lui étant postérieure.

2.1.3 Analyse de la qualité données

Contrôler la qualité des données est indispensable afin d'avoir confiance dans les résultats obtenus en les exploitant. L'analyse de la qualité des données permet d'identifier les incohérences à éliminer et de garantir la fiabilité des données.

Selon la variable étudiée, jusqu'à 4 contrôles sont réalisés :

- Contrôle d'exhaustivité : la variable est-elle bien renseignée ?
- Contrôle de vraisemblance : la valeur de la variable est-elle logique ?
- Contrôle de cohérence : la variable est-elle cohérente avec les autres variables ?
- Contrôle de distribution : n'y a-t-il pas une occurrence avec un poids trop important ?

Un tableau récapitulatif des contrôles effectués est disponible en annexe A.2.

Pour chacune des variables, un pourcentage d'observations satisfaisant les conditions d'exhaustivité, de vraisemblance ou de cohérence est calculé. Les variables possédant beaucoup d'anomalies et nécessitant une fiabilisation sont ainsi mises en évidence. Les observations incohérentes sont aussi identifiées.

Étant donné la volumétrie importante de la base Open DAMIR, ces contrôles sont effectués sur chaque base mensuelle. Les résultats sont présentés à titre illustratif sur la base mensuelle de juin 2018.

Les contrôles d'exhaustivité et de vraisemblance ont mis en évidence l'absence de valeurs manquantes dans la base Open DAMIR et l'existence de modalités spécifiques pour les valeurs inconnues.

Nom	Libellé	Contrôle d'exhaustivité	Contrôle de vraisemblance	Contrôle de cohérence	Contrôle distribution
FLX_ANN_MOI	Année et Mois de Traitement	100%			
SOI_ANN	Année de Soins	100%	100%	100%	✓
SOI_MOI	Mois de Soins	100%	100%		✓
BEN_SEX_COD	Sexe du Bénéficiaire	100%	99,9%		✓
AGE_BEN_SNDS	Tranche d'Age du Bénéficiaire au moment des soins	100%	99,8%		✓
BEN_RES_REG	Région de Résidence du Bénéficiaire	100%	95,6%		✓
PRS_NAT	Nature de Prestation	100%	99,9%	100%	
PRS_ACT_QTE	Quantité	100%	99,8%		✓
PRS_PA1_MNT	Montant de la Dépense	100%	99,9%	99,8%	
PRS_DEP_MNT	Montant du Dépassement	100%	99,9%	99,9%	
PRS_REM_TAU	Taux de Remboursement	100%	99,9%	99,5%	
PRS_REM_BSE	Base de Remboursement	100%	99,6%	99,8%	
PRS_REM_MNT	Montant Versé/Remboursé	100%	99,6%	98,9%	✓

FLT_ACT_QTE	Quantité de la Prestation Préfiltrée	100%	99,9%	99,9%	✓
FLT_PA1_MNT	Montant de la Dépense de la Prestation Préfiltrée	100%	99,9%	99,6%	✓
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltrée	100%	99,9%	99,9%	
PRS_REM_TYP	Type de Remboursement	100%	48,1%		
CPL_COD	Complément d'Acte	100%	100%	99,9%	
PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)	100%	100%		
PRS_PPU_SEC	Code Secteur Privé/Public	100%	100%		
BEN_QLT_COD	Qualité du Bénéficiaire	100%	100%		
ETE_TYP_SNDS	Type Etat Exécutant	100%			
PSE_ACT_CAT	Catégorie de l' Exécutant	100%	99,7%		
PSP_ACT_CAT	Catégorie du Prescripteur	100%	92,0%		
PSE_SPE_SNDS	Spécialité Médicale PS Exécutant	100%	32,0%		
PSP_SPE_SNDS	Spécialité Médicale PS Prescripteur	100%	80,8%		
PSE_ACT_SNDS	Nature d'Activité PS Exécutant	100%	63,3%		
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur	100%	5,5%		

FIGURE 2.3 – Résultats des contrôles de qualité des données réalisés sur la base de juin 2018

Les contrôles ont permis d'identifier différents traitements à effectuer afin d'améliorer la qualité des données. Ils seront détaillés dans la partie suivante. Le contrôle de vraisemblance effectué sur le "Type de remboursement" a mis en évidence la nécessité de fiabiliser cette variable avant de pouvoir l'utiliser. Une nouvelle variable relative au type de remboursement sera donc créée. Il convient de noter que certains contrôles nécessitent une distinction des différents types de remboursement en raison de la construction particulière de la base (par exemple, la base de remboursement n'est pas renseignée pour les prises en charges complémentaires). La nouvelle variable créée est alors utilisée pour réaliser les contrôles nécessitant cette distinction.

Pour rappel, les compléments d'acte et les prises en charge complémentaires sont inscrits sur des lignes indépendantes des prestations de référence auxquelles ils sont associés. Les montants de dépense liés à ces remboursements sont ceux présents sur les lignes de prestations de référence associées et il est le plus souvent impossible d'identifier individuellement la ligne de prestations de référence associée à une ligne de compléments d'acte ou de prises en charge complémentaires. Ainsi, il est nécessaire d'agrégier la base afin de pouvoir contrôler la cohérence des montants remboursés au titre de compléments d'acte ou de prises en charge complémentaires, puisque l'agrégation permettra de rassembler sur une même ligne les montants associés aux différents types de remboursements. Cependant, l'agrégation de la base ne peut être effectuée qu'après certaines étapes réalisées par la suite puisqu'elle nécessite notamment la suppression du taux de remboursement (qui n'est pas forcément identique pour les différents types de remboursements, comme l'illustre l'exemple 2.1). Les contrôles seront donc effectués lors d'une étape ultérieure.

2.1.4 Fiabilisation des données

L'analyse de la qualité des données a mis en évidence plusieurs anomalies. Des traitements sont effectués afin de fiabiliser les variables qui s'y prêtent.

Région de résidence du bénéficiaire

Afin de réduire le nombre d'observations manquantes de cette variable, les régions associées aux professionnels de santé et à l'organisme d'assurance maladie sont utilisées. L'ordre d'utilisation de ces variables est déterminé en observant le taux de correspondance de chaque variable avec la "Région de résidence du bénéficiaire". La variable avec le plus d'observations similaires à la "Région de résidence du bénéficiaire" est utilisée en premier, et ainsi de suite. L'ordre d'utilisation retenu est le suivant :

1. Région de l'organisme de liquidation,
2. Région du professionnel de santé exécutant,
3. Région du professionnel de santé prescripteur,
4. Région d'implantation de l'établissement exécutant,
5. Région d'implantation de l'établissement prescripteur.

A titre illustratif, le pourcentage d'observations manquantes avant ce traitement était de 4,4% sur la base de juin 2018. Il est porté à 0,2% après cette fiabilisation.

Taux de remboursement

Des anomalies relatives au taux de remboursement ont été observées lors du contrôle de cohérence. Certains taux de remboursements sont strictement positifs alors que le montant remboursé et la base de remboursement sont nuls. Ces taux sont alors forcés à 0. Cette modification concerne 0,5% des taux de la base de juin 2018.

Montant remboursé

Comme observé précédemment, les participations assuré de 18/24€ sont inscrites en montants négatifs. Seuls les montants remboursés sont conservés puisqu'ils correspondent à l'opposé des montants remboursés par l'assurance complémentaire santé. Leur signe est modifié en positif.

Dans le tableau des résultats des contrôles de qualité des données affiché précédemment, une anomalie de cohérence apparaît pour la variable "Montant remboursé". Elle met en évidence des observations telles que le montant remboursé est différent de la base de remboursement multiplié par le taux de remboursement. Cette anomalie concerne 1,1% des remboursements de la base mais 9% des remboursements liés à l'hospitalisation. Par hypothèse, le montant remboursé est considéré fiable et la base de remboursement pour les prestations d'hospitalisation ne sera pas utilisée.

Type de remboursement

Le contrôle de qualité des données réalisé sur la variable "Type de remboursement" montre que cette variable possède beaucoup d'observations manquantes. Cependant, la base possède deux types d'indicateurs. Les indicateurs préfiltrés concernent uniquement la part des prestations prise en charge par le régime obligatoire, c'est-à-dire les remboursements au titre des prestations de références et des compléments d'acte. Cette distinction est utilisée pour isoler les prises en charge complémentaires lorsque le "Type de remboursement" n'est pas renseigné. La variable "Complément d'acte" est ensuite utilisée pour distinguer les compléments d'acte des prestations de référence.

Professionnels de santé : exécutant et prescripteur

Les contrôles réalisés sur les variables liées aux exécutants et aux prescripteurs ont fait ressortir le fait que les variables "Nature d'activité" et "Spécialité médicale" se complètent : la spécialité n'est pas renseignée lorsque la nature d'activité l'est et inversement. Ces variables peuvent donc être regroupées. La "Catégorie" du professionnel de santé donne une information moins précise que les variables "Nature d'activité" et "Spécialité médicale". Elle n'est donc utilisée qu'à des fins de fiabilisation d'autres variables.

Une variable "Exécutant" est créée afin de combiner les variables "Spécialité médicale du professionnel de santé exécutant" et "Nature d'activité du professionnel de santé exécutant". Cette variable est complétée à l'aide de la "Catégorie de l'exécutant". Ainsi, seulement 0,3% des observations de cette nouvelle variable sont manquantes suite à ce traitement, contre 4,7% initialement.

De la même manière, les variables "Spécialité médicale du professionnel de santé prescripteur" et "Nature d'activité du professionnel de santé prescripteur" sont regroupées. La variable créée est complétée à l'aide de la "Catégorie du prescripteur". La variable obtenue possède 7,9% de valeurs non renseignées pour la base de juin 2018. Ce taux élevé s'explique par le fait que tous les actes ne sont pas prescrits par un professionnel de santé. Une modalité "Sans objet" est donc créée.

Incohérences restantes

Les observations incohérentes qui ne peuvent pas être fiabilisées, comme le sexe de valeur inconnue par exemple, sont supprimées. Seules les observations non renseignées relatives aux variables "Exécutant" et "Prescripteur" sont conservées.

Les contrôles de qualité sont appliqués aux données après ces retraitements et suppressions afin de vérifier que toutes les incohérences sont bien traitées. Sur la base de juin 2018, il reste ainsi 98,81% des lignes conservées après la sélection des données. Cette analyse de la qualité ainsi que les retraitements et suppressions réalisés ont permis d'augmenter le niveau de confiance dans la base de données.

2.2 Préparation des données pour une mesure de l'inflation comparable à celle du portefeuille de Generali

Deux objectifs sont poursuivis à travers cette étape de préparation des données : un portefeuille Generali doit être reconstitué à partir des données marché et la base d'étude obtenue doit permettre de calculer l'inflation.

2.2.1 Catégorisation des prestations

La base Open DAMIR comporte initialement 1080 codes de natures de prestation, d'après le lexique Open DAMIR. Après sélection du périmètre d'étude, il reste plus de 300 codes d'actes différents. Il est alors nécessaire de regrouper ces natures de prestation en catégories de manière à disposer d'une granularité cohérente afin de calculer l'évolution du coût moyen pour chacune d'entre elles et ainsi limiter l'impact du mix de prestations.

Les prestations sont d'abord regroupées en 8 postes : Soins courants, Pharmacie, Dentaire, Optique, Hospitalisation, Cures thermales, Appareillage-prothèses et Autre. Cette segmentation en postes est assez classique en assurance santé, même s'il n'existe pas de catégorisation officielle. En particulier, le poste Cures thermales est fréquemment regroupé avec le poste Autre. Cependant, les cures thermales sont souvent utilisées par les assureurs pour améliorer l'attractivité de leurs contrats. Ils sont alors particulièrement intéressés par l'inflation de la sinistralité de ce poste et par son suivi spécifique. Un découpage en familles de remboursements est ensuite effectué, comme le montre la figure ci-dessous.

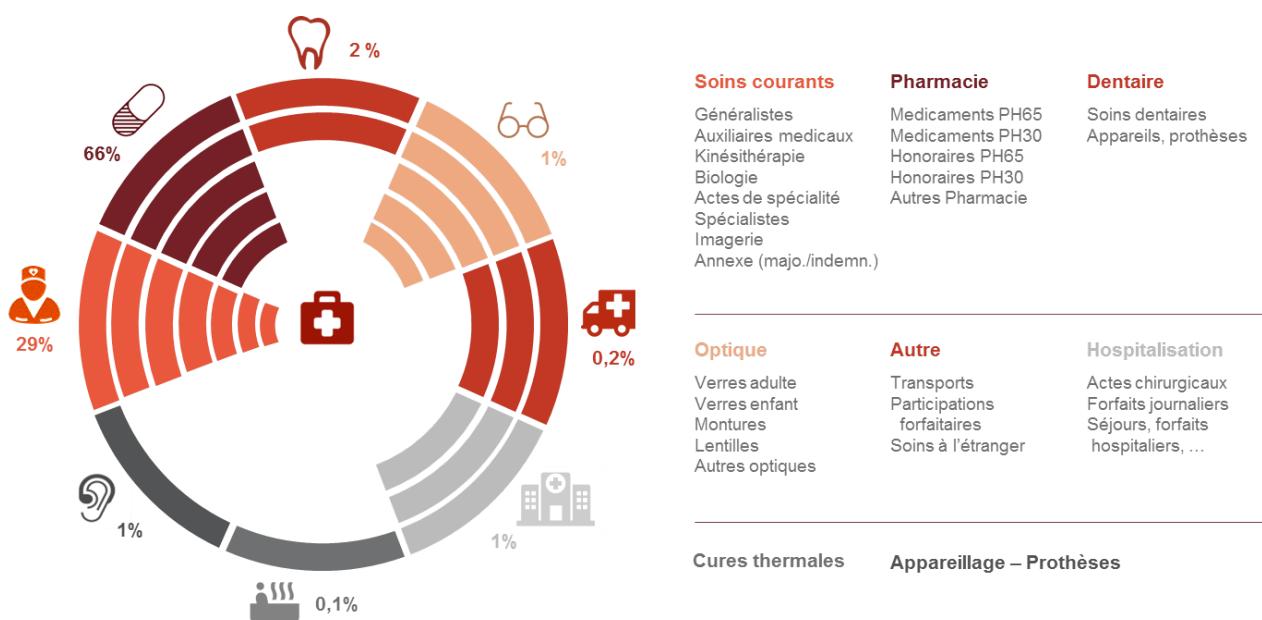


FIGURE 2.4 – Postes et familles de remboursements

Ces catégories sont créées en observant, pour chaque prestation, la nature de l'acte et les montants moyens associés pour chaque année (le montant remboursé moyen, le montant de frais réels moyen, le taux de remboursement moyen et la base de remboursement moyenne). Le nombre d'actes est aussi regardé afin de disposer d'un volume de données suffisant dans chaque regroupement.

Les pourcentages affichés sur la figure 2.4 correspondent à la proportion d'actes de chaque poste dans la base d'étude.

Concernant le poste Hospitalisation, il a été mentionné précédemment que la base Open DAMIR contient seulement les prestations hospitalières facturées directement à l'assurance maladie. Une grande partie des prestations hospitalières (frais de séjour, frais de salle d'opération, etc.) du secteur public sont donc absentes. L'inflation observée sur les données liées à l'hospitalisation de la base Open DAMIR ne sera donc pas représentative du marché. Ce poste est conservé dans la base d'étude à titre indicatif, l'inflation associée ne sera pas étudiée.

La segmentation retenue rassemble les consultations et visites de médecine générale dans la famille "Généralistes", tandis que la famille "Spécialistes" regroupe les consultations et visites hors médecine générale ainsi que les avis ponctuels de consultant. Cette segmentation permettra par la suite d'observer l'évolution de la base de remboursement de 23€ à 25€ survenue en 2017 pour la famille "Généralistes".

Il convient de souligner que les codes de natures de prestation ne permettent pas d'isoler les actes de médecine générale. Ils ne suffisent donc pas pour distinguer les catégories "Généralistes" et "Spécialistes". Quatre codes d'actes existent pour les consultations des médecins :

- C : Consultation par le médecin généraliste, le chirurgien-dentiste omnipraticien ou la sage-femme,
- G : Consultation de médecine générale,
- CS : Consultation par le médecin spécialiste qualifié, le médecin spécialiste qualifié en médecine générale ou le chirurgien-dentiste spécialiste qualifié,
- GS : Consultation de spécialiste qualifié en médecine générale.

Les équivalents de ces quatre codes existent pour les visites des médecins (V, VG, VS, VGS).

Comme l'indiquent les libellés des codes d'actes inscrits ci-dessus, les lettres clés G, GS, VG et VGS ne concernent que la médecine générale, contrairement aux codes C, CS, V et VS. Cela s'explique par le fait que les codes G, GS, VG et VGS ont été créés lors de l'évolution réglementaire de 2017 pour les consultations et visites des généralistes avec une base de remboursement de 25€. Ce point sera détaillé dans la partie 2.2.2.

Les prestations associées aux lettres clés G, GS, VG et VGS sont donc rassemblées dans la famille "Généralistes".

En revanche, pour les codes C, CS, V et VS, l'information relative au professionnel de santé exécutant est utilisée pour isoler les actes de médecine générale. La catégorie "Généralistes" contient ainsi les remboursements liés aux codes C, CS, V et VS pour lesquels le professionnel de santé exécutant a comme spécialité médicale "Médecine Générale". Les remboursements des codes C, CS, V et VS associés aux autres professionnels de santé exécutants sont regroupés dans la famille de remboursements "Spécialistes".

2.2.2 Retraitements des évolutions réglementaires

Dans le but de mesurer l'inflation, il est nécessaire de retraitier les montants impactés par une évolution réglementaire, puisque ces dernières créent une inflation artificielle des frais de santé. Chaque famille d'acte est alors inspectée afin d'identifier les modifications légales ayant eu lieu entre 2016 et 2019. Les évolutions réglementaires peuvent être de différents types :

- Création d'une prestation,
- Suppression d'une prestation,
- Changement de la base de remboursement de la Sécurité Sociale,
- Mise en place d'honoraires de facturation limités.

Des prestations ont été créées durant le périmètre temporel étudié. Il s'agit par exemple des trois nouveaux honoraires de dispensation entrés en vigueur au 1er janvier 2019 ou des majorations MUT (Majoration Urgence médecin Traitant) et MCU (Majoration Correspondant Urgence) créées au 1er janvier 2018. Elles sont exclues du calcul de l'inflation pour les années dont l'inflation serait biaisée par cette évolution. Par exemple, les MUT et MCU sont exclues du calcul de l'inflation 2017-2018 mais sont conservées pour le calcul de l'inflation 2018-2019 puisqu'elles sont entrées en vigueur au 1er janvier 2018. Les trois nouveaux honoraires de dispensation eux ne peuvent être pris en compte dans aucune des inflations puisqu'ils ne sont pas présents durant au moins deux années complètes du périmètre temporel étudié. Les prestations supprimées entre 2016 et 2019 sont exclues de la même manière. Au total, 28 prestations sont concernées par une ou plusieurs exclusions.

Certaines prestations ont subi une évolution de leur base de remboursement. Des montants *as-if* sont alors calculés. Ils correspondent aux montants qui auraient été observés si les conditions de remboursement de la Sécurité Sociale actuelles étaient en vigueur. Ainsi, les montants moyens sont observés sur les mêmes bases réglementaires au fil du temps. La liste des prestations retraitées est disponible en annexe A.3. Voici deux exemples permettant d'illustrer la démarche réalisée. Le premier exemple est simple et correspond au retraitement effectué sur la majorité des prestations. Le second exemple est particulier puisque des hypothèses supplémentaires sont nécessaires.

Exemple 1 : Majorations de coordination

Les majorations de coordination ont été revalorisées de 3€ à 5€ au 1er juillet 2017 pour les médecins généralistes et spécialistes et de 4€ à 5€ pour les psychiatres, neuropsychiatres et neurologues. Cette évolution réglementaire est bien visible sur les données de la base Open DAMIR et est observée sur les histogrammes ci-dessous.

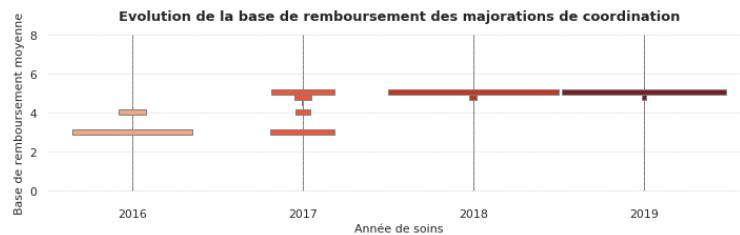


FIGURE 2.5 – Histogrammes annuels des bases de remboursement

L'évolution observée concerne la base de remboursement ainsi que le montant de frais réels. Pour rappel, la base Open DAMIR est agrégée. Pour chaque ligne, le retraitement *as-if* réalisé est le suivant :

- Lorsque la base de remboursement moyenne est comprise entre 2,50€ et 3,50€, la ligne est considérée comme une majoration de coordination associée à un médecin généraliste ou spécialiste. Le montant de frais réels et la base de remboursement sont augmentés de $2\text{€} \times \text{nombre d'actes}$.
- Lorsque la base de remboursement moyenne est comprise entre 3,50€ et 4,50€, la ligne est considérée comme une majoration de coordination associée à un psychiatre, neuropsychiatre ou neurologue. Le montant de frais réels et la base de remboursement sont augmentés de $1\text{€} \times \text{nombre d'actes}$.

Les montants *as-if* obtenus sont ainsi sur la même base réglementaire tout au long de la période étudiée.

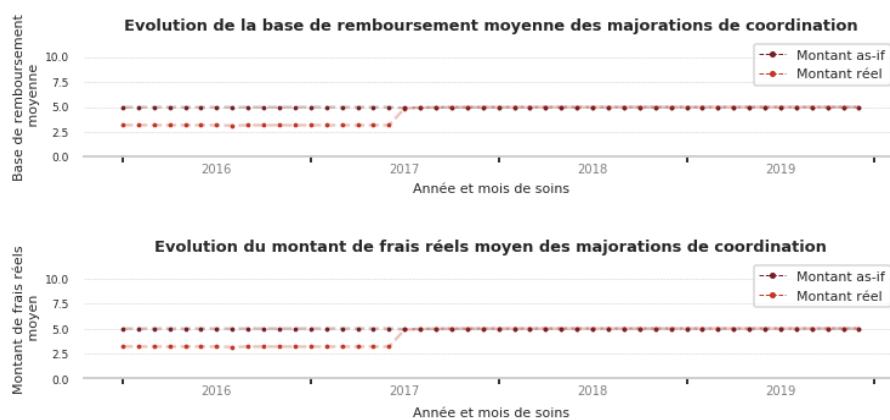


FIGURE 2.6 – Comparaison des montants moyens mensuels réels et *as-if*

Exemple 2 : Consultations et visites des médecins généralistes

Les consultations et visites des médecins généralistes ont été revalorisées de 23€ à 25€ au 1er mai 2017 pour les médecins généralistes de secteur 1 et de secteur 2 adhérents à l'OPTAM. Les consultations et visites des généralistes ont aussi été revalorisées de 2€ dans les DROM.

Pour les médecins concernés par cette évolution, deux nouveaux codes de consultations sont créés, G et GS, avec une base de remboursement de 25€. Ils remplacent C et CS. La signification de ces codes d'actes est inscrite dans la partie 2.2.1. De manière similaire, les codes d'actes VG et VGS sont créés pour les visites, avec une base de remboursement de 25€. Ils remplacent V et VS.

Les généralistes de secteur 2 hors OPTAM continuent à facturer les codes C/CS ou V/VS avec une base de remboursement de 23€. Ils ne peuvent coter G/GS ou VG/VGS (25€) que pour les patients bénéficiant de la CMU-C ou de l'ACS et pour les consultations facturées au tarif opposable.

Le retraitement doit concerter les actes réalisées avant le 1er mai 2017 afin de reproduire les conditions de remboursement de la sécurité sociale postérieures à cette date. Les actes à retraiter sont donc des prestations cotées C, CS, V ou VS associées à des généralistes. Plusieurs observations permettent d'identifier les retraitements à effectuer.

- Retraitements des actes des médecins de secteur 1 et de secteur 2 adhérents à l'OPTAM seulement :

Les médecins de secteur 2 hors OPTAM ne sont pas concernés par cette modification réglementaire. Le retraitement ne doit donc concerter qu'une partie des consultations et visites des généralistes réalisées avant le 1er mai 2017 et non la totalité des actes cotés C, CS, V et VS associés aux généralistes.

A partir du 1er mai 2017, la distinction entre les médecins généralistes de secteur 1 et de secteur 2 adhérents à l'OPTAM et les médecins généralistes de secteur 2 hors OPTAM peut être approchée grâce à la distinction entre les codes d'actes C/CS/V/VS et G/GS/VG/VGS.

La méthode retenue consiste alors à observer le pourcentage d'actes réalisés par les médecins de secteur 1 et de secteur 2 adhérents à l'OPTAM sur les données postérieures au 1er mai 2017, c'est-à-dire le pourcentage d'actes cotés en G, GS, VG ou VGS parmi tous les actes de la famille "Généralistes" postérieurs au 1er mai 2017. Ce pourcentage est ensuite appliqué aux données antérieures au 1er mai 2017 pour déterminer le nombre d'actes à retraiter.

98% des actes des généralistes sont cotés G, GS, VG ou VGS après le 1er mai 2017. Il est donc considéré que 98% des actes ayant eu lieu avant cette date ont été effectués par des généralistes de secteur 1 ou de secteur 2 adhérents à l'OPTAM. 98% des actes ayant eu lieu avant le 1er mai 2017 sont ainsi concernés par le retraitement réglementaire.

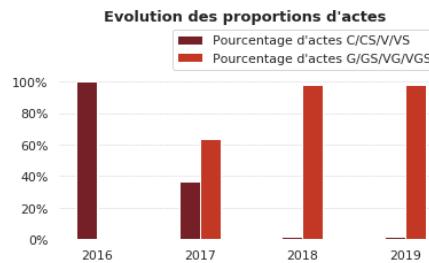


FIGURE 2.7 – Évolution des proportions d'actes C/CS/V/VS vs G/GS/VG/VGS

- Retraitements spécifiques des montants de frais réels :

De manière similaire à la base de remboursement, les montants de frais réels des médecins de secteur 1 ont évolué de 2€. Concernant les médecins de secteur 2 adhérents à l'OPTAM, il a été vu dans le paragraphe 1.2.4 que l'option repose sur la stabilité du montant de l'acte et non sur la stabilité du montant du dépassement. Il est donc considéré que les montants de frais réels de ces médecins n'ont pas augmenté suite à l'évolution de la base de remboursement.

En observant les données du portefeuille de Generali, il apparaît en effet que les montants de frais réels qui étaient supérieurs à 25€ avant la modification légale n'ont pas évolué. Cette observation semble se confirmer sur la base Open DAMIR. Le graphique ci-dessous montre que le montant de frais réels moyen a moins augmenté que la base de remboursement moyenne.

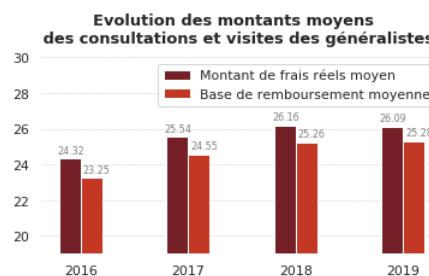


FIGURE 2.8 – Évolution des montants moyens annuels

Le problème lié à la mise en place de ce retraitement est que la base Open DAMIR est agrégée. Les nombres d'actes avec et sans dépassement d'honoraires ne sont pas disponibles. Les actes avec un montant de frais réels supérieur à 25€ ne peuvent pas être identifiés.

La méthode choisie consiste alors à définir un montant de dépassement moyen et en déduire, pour chaque ligne de la base agrégée, une estimation du nombre d'actes avec et sans dépassement.

Pour ce faire, le montant de dépassement moyen a été défini en observant, sur les données antérieures au 1er mai 2017 du portefeuille de Generali, le montant de dépassement moyen des prestations des généralistes qui possèdent un dépassement. Il est noté $dep_moyen_{Generali}$.

La proportion d'actes dont le montant de frais réels est à augmenter de 2€ est définie à partir de ce dépassement moyen et du dépassement moyen de chaque ligne de la base d'étude agrégée.

L'hypothèse retenue est ainsi la suivante : pour chaque ligne de la base d'étude, le montant de dépassement moyen est composé d'une part d'actes x sans dépassement et d'une part $(1 - x)$ ayant un montant de dépassement égal à $dep_moyen_{Generali}$.

$$dep_moyen_{Open DAMIR} = x * 0 + (1 - x) * dep_moyen_{Generali}$$

Où x correspond à la proportion d'actes de chaque ligne dont le montant de frais réels est à augmenter de 2€.

- Récapitulatif du retraitement :

Pour chaque ligne de la base d'étude concernant des actes de généralistes survenus avant le 1er mai 2017 :

- 98% des actes ont vu leur base de remboursement augmenter de 2€,
- Parmi les 98% d'actes retraités, $x\%$ des actes ont vu leur montant de frais réels augmenter de 2€ ; où x est propre à chaque ligne et définit le nombre d'actes sans dépassement.

Après ce retraitement, les montants *as-if* obtenus sont sur la même base réglementaire tout au long de la période étudiée.



FIGURE 2.9 – Comparaison des montants moyens mensuels réels et *as-if*

Il convient de mentionner qu'un retraitement différent est effectué pour le poste Dentaire, sur lequel une évolution des dépenses est observée en raison de l'entrée en vigueur des premières mesures de la réforme 100% santé au 1er avril 2019. Étant donné que la mise en place de cette réforme est très récente et progressive sur plusieurs années, les retraitements sont effectués comme si cette réforme n'était pas encore entrée en vigueur, ce qui signifie que les prestations postérieures au 1er avril sont retraitées.

Cette réforme engendre une revalorisation de soins bucco-dentaires fréquents et l'instauration progressive de plafonds tarifaires pour certains actes prothétiques. Une augmentation de la base de remboursement est bien observée dans la base Open DAMIR sur les soins dentaires, à l'exception de l'inlay-core qui voit sa base de remboursement diminuer comme le prévoit la réforme. Pour les prothèses dentaires, une diminution des montants apparaît à la suite de la mise en place de plafonds. La base Open DAMIR ne permet cependant pas de distinguer précisément les actes revalorisés. Les sauts de sinistralité observés sont alors neutralisés de façon globale pour chaque code d'acte ou ensemble de codes d'acte concernés par une évolution.

2.2.3 Création de la base d'étude

Sélection des indicateurs pertinents

Comme mentionné dans le paragraphe 2.1.1, il existe plusieurs types d'indicateurs dans la base Open DAMIR. Voici la liste des indicateurs utilisés. Ils sont renommés afin de ne plus porter l'indication "préfiltré".

- Nombre d'actes : Il s'agit de la variable "Quantité de la prestation préfiltrée". Elle permet de comptabiliser les remboursements au titre d'une prestation de référence.
- Montant de frais réels : Cette variable correspond au "Montant de la dépense préfiltré". Elle permet de ne prendre en compte qu'une seule fois le montant de frais réels de l'acte même si ce dernier est concerné par des remboursements complémentaires.
- Montant du dépassement : Il s'agit de la variable "Montant du dépassement préfiltré".
- Montant remboursé par la Sécurité Sociale : Il correspond au "Montant versé/ remboursé" non préfiltré. Tous les types de remboursements sont pris en compte afin de pouvoir déduire de ce montant remboursé le montant qu'il reste à la charge de l'assuré, sur lequel l'assurance complémentaire peut intervenir.

Certaines variables sont créées en appliquant un filtre sur une variable présente initialement dans la base Open DAMIR :

- Nombre de compléments d'acte : Il correspond à la variable "Quantité" de la base seulement pour les lignes associées à des remboursements de compléments d'acte.

- Nombre d'actes complémentaires : Il correspond à la variable "Quantité" pour les lignes associées à des remboursements complémentaires.
- Base de remboursement de la Sécurité Sociale : Il s'agit de la variable "Base de Remboursement" sur laquelle un filtre a été appliqué pour ne conserver que les montants associés aux remboursements au titre d'une prestation de référence ou d'un complément d'acte.

Création de variables

- Top établissements et code secteur Privé/Public :

Afin de réduire le volume de la base de données et de ne pas avoir de redondance dans les variables, le "Type d'Établissement Exécutant" et le "Code Secteur Privé/Public" sont regroupés en une seule variable. La nouvelle variable créée possède 3 modalités : "Établissement privé", "Établissement public" et "Soins de ville".

- Taux de changements de région :

La proportion d'actes associés à un changement de région pour aller consulter un professionnel de santé pourrait être un indicateur de difficulté d'accès au soins dans les régions de résidence des patients. Pour chaque région, le pourcentage d'actes réalisés hors de la région de résidence du bénéficiaire est alors calculé. Les variables "Région de résidence du bénéficiaire" et "Région du professionnel de santé exécutant" sont utilisées pour déterminer si le bénéficiaire des soins a changé de région pour se rendre chez le professionnel de santé exécutant.

Il convient de mentionner que la variable "Région du professionnel de santé exécutant" n'est pas renseignée pour 2,3% des actes. Ces valeurs manquantes ne sont pas prises en compte, afin de considérer que le même comportement est observé sur les valeurs manquantes que sur le reste des observations.

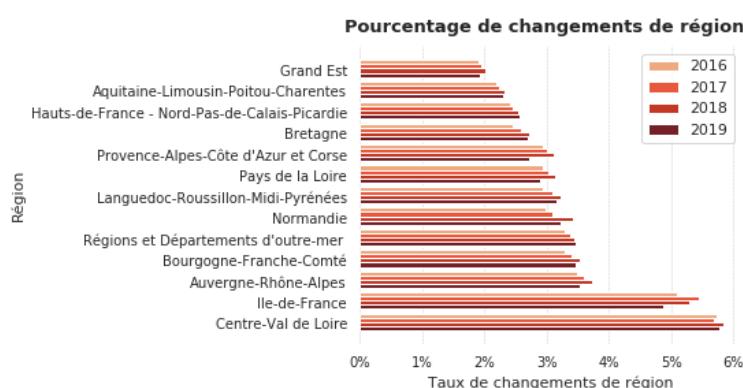


FIGURE 2.10 – Taux de changements de région observés

Traitement des remboursements hors du parcours de soins

Lorsque le patient est hors du parcours de soins, le taux de remboursement de la Sécurité sociale passe de 70% à 30%. L'assurance complémentaire ne couvre pas la pénalité financière appliquée par le régime obligatoire. Elle considère que l'assurance maladie a remboursé 70% de la base de remboursement même si le patient était hors du parcours de soins et qu'il n'a été remboursé qu'à hauteur de 30% de la base de remboursement.

Afin de ne pas inclure les pénalités dans le remboursement de Generali calculé par la suite, il est nécessaire de retraitre les actes hors parcours de soins. La variable "Code Qualificatif Parcours de Soins" est utilisée pour identifier les actes à retraiter. Lorsque le taux de remboursement est égal à 30% et que l'acte est un soin courant réalisé hors du parcours de soins, le taux de remboursement est forcé à 70% et le montant remboursé par la Sécurité Sociale est recalculé. Le montant remboursé *as-if* calculé contient ainsi la partie non remboursée par l'assurance maladie en raison du parcours de soins. Ce montant sera alors exclu du remboursement de l'assurance complémentaire simulé par la suite.

Agrégation de la base

A ce stade de la préparation de la base d'étude, toutes les variables qui avaient été conservées dans un objectif de fiabilisation ou de construction d'autres variables peuvent être supprimées. 18 variables sont conservées dont 7 indicateurs. Toutes les bases mensuelles sont rassemblées et agrégées.

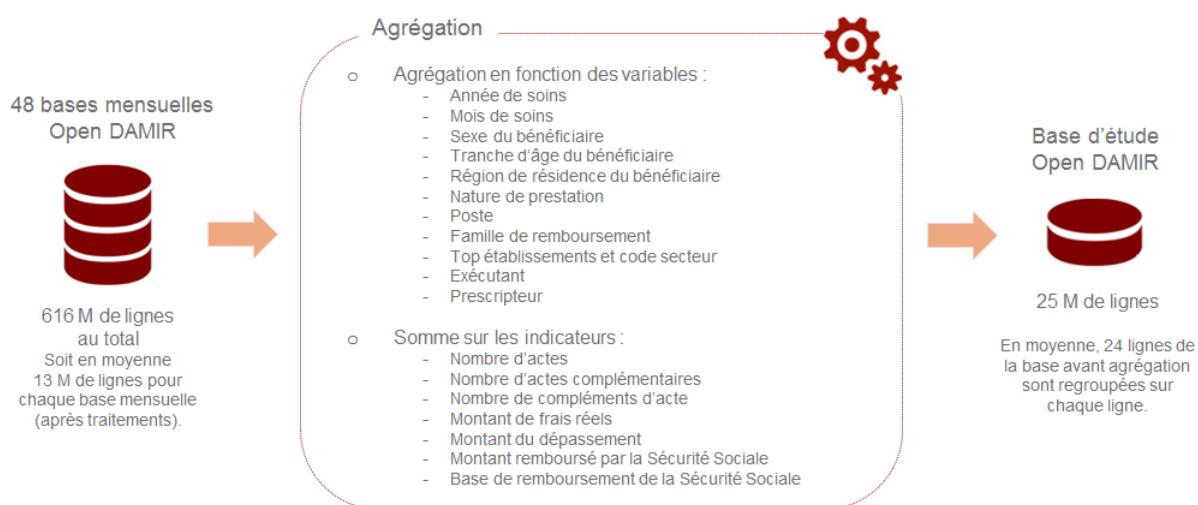


FIGURE 2.11 – Agrégation des bases

L'agrégation de la base a permis de sommer sur une même ligne les différents types de remboursement. Les remboursements au titre d'une prestation de référence, d'un complément d'acte ou d'une prise en charge complémentaire ne sont plus séparés.

De nouveaux contrôles de cohérence et de distribution sont effectués sur la base obtenue. Ils permettent de mettre en évidence certaines anomalies.

Traitements des anomalies

3% des lignes de la base agrégée possèdent une anomalie. Ces anomalies sont liées au rassemblement des différents types d'actes. Elles peuvent être engendrées par les suppressions réalisées lors de la fiabilisation de la base de données ou peuvent être dues à la qualité des données. Par exemple, la base étant agrégée en fonction de l'âge du bénéficiaire, si cette variable est mal renseignée pour la prestation de référence ou pour le complément d'acte alors ces deux types de remboursements ne sont pas rassemblés, ce qui peut engendrer une anomalie.

Afin de vérifier que toutes les anomalies ne sont pas dues aux suppressions réalisées lors de l'étape de qualité des données, la base contenant les lignes supprimées est agrégée de la même manière que la base d'étude. Les deux bases sont ensuite comparées. Seulement 16% des lignes de la base d'étude qui présentent une anomalie ont au moins une ligne qui a été supprimée lors de l'étape de fiabilisation des données. La majorité des anomalies est alors due à la qualité des données de la base et non pas aux traitements réalisés pour les besoins de l'étude.

Trois sortes d'anomalies sont détectées. La première correspond à des lignes qui ne contiennent que des compléments d'acte et des prises en charge complémentaires mais pas de prestation de référence associée. La deuxième anomalie est composée de lignes de la base de données qui possèdent une base de remboursement supérieure au montant de frais réels diminué du montant du dépassement. La dernière anomalie concerne des lignes qui possèdent un montant remboursé supérieur à la base de remboursement. Des traitements sont réalisés pour corriger ces anomalies. Ils sont détaillés en annexe A.4. Ces différents retraitements contribuent à la fiabilisation de la base d'étude.

2.2.4 Rapprochement de la base d'étude et du portefeuille de Generali

L'inflation calculée sur la base d'étude doit pouvoir être comparée à celle observée sur le portefeuille de Generali. Des ajustements sont réalisés afin de reconstituer un portefeuille Generali à partir des données de la base Open DAMIR.

Application de garanties Generali

La base d'étude contient des prestations avec leur montant de frais réels, leur base de remboursement et leur montant remboursé par la Sécurité Sociale. Il est alors possible de calculer le montant qui serait remboursé par Generali si ces prestations concernaient ses assurés.

Les garanties des formules proposées par Generali sont alors appliquées aux prestations contenues dans la base d'étude. Une étape préliminaire consiste à étudier le portefeuille de Generali afin de sélectionner les formules à appliquer et dans quelles proportions, puisque Generali ne propose pas une unique formule à ses assurés. Ces garanties sont ensuite utilisées pour calculer le montant que rembourserait Generali sur les prestations de la base d'étude. Étant donné la structure et la composition de la base Open DAMIR, des biais sont engendrés par l'application des garanties. La méthodologie de calcul des remboursements de Generali est alors testée sur les données du portefeuille avant d'être appliquée à la base d'étude afin de mesurer les différents biais.

Une erreur est calculée pour chaque famille de remboursements comme étant la différence entre le montant réel remboursé et le montant estimé. L'erreur totale correspond à la somme des valeurs absolues des erreurs associées aux différentes familles de remboursements. Elle est exprimée en pourcentage du montant réel remboursé par Generali.

- Choix des garanties appliquées

Le portefeuille de Generali est composé de plusieurs produits. Deux produits sont principalement présents et représentent une part importante du portefeuille. Afin de le simplifier, il est envisagé d'utiliser uniquement les 9 formules associées à ces deux produits pour reproduire l'ensemble des remboursements du portefeuille. Ces formules sont appelées les "formules principales".

Afin de vérifier que l'utilisation de ces formules principales uniquement permet d'approcher raisonnablement les remboursements réels, l'erreur d'estimation engendrée par l'utilisation de ces 9 formules est calculée. Pour ce faire, les montants remboursés associés aux prestations présentes dans le portefeuille sont recalculés en utilisant successivement les garanties de chaque formule principale. Pour chaque formule non principale, il est alors possible de déterminer quelle formule principale permet d'obtenir le montant remboursé estimé le plus proche du montant réel. Pour évaluer l'erreur, les montants remboursés par Generali sont recalculés en utilisant pour chaque formule du portefeuille les garanties de la formule principale la plus proche.

La différence entre le montant réel remboursé par Generali et le montant estimé est observée dans un premier temps pour chaque formule du portefeuille. Ces différences montrent que les 9 formules principales permettent d'approximer les remboursements du portefeuille, excepté pour les produits "Anciennes Générations" et pour certaines formules avec un niveau de couverture élevé. Il est choisi de se concentrer sur les garanties actuelles. Les produits "Anciennes Générations" sont alors exclus du portefeuille Generali utilisé pour la comparaison. Concernant les garanties plus élevées, elles sont plus spécifiques aux clients qui les souscrivent ce qui explique qu'elles ne soient pas proches des formules retenues. Elles seront à l'origine d'un biais dans les remboursements estimés. L'erreur totale obtenue est égale à 2% du montant réel remboursé. Il est considéré que les formules retenues approximent de manière satisfaisante les garanties du portefeuille. Ces 9 formules sont par conséquent utilisées.

L'objectif est ensuite d'appliquer ces 9 formules sur les prestations de la base d'étude dans des proportions représentatives du portefeuille. Au cours du temps, l'un des deux produits a été de plus en plus commercialisé. Afin de répliquer au mieux le portefeuille Generali, les proportions sont calculées annuellement. L'impact est cependant relativement faible puisque les proportions sont modifiées entre des garanties de même niveaux (garanties d'entrée de gamme, milieu de gamme ou haut de gamme).

- Biais dans l'application des garanties à la base Open DAMIR

La base Open DAMIR est une base agrégée composée des remboursements de l'assurance maladie. Elle ne contient pas les remboursements des complémentaires santé et ne comporte par conséquent pas de notion de niveau de couverture des assurés. Deux principaux biais sont alors engendrés par l'application des garanties.

- Corrélation du montant remboursé avec le niveau de couverture souscrit

En observant le portefeuille de Generali, il apparaît que le niveau de couverture est corrélé au montant remboursé, ce qui souligne la présence d'aléa moral. Un biais est par conséquent engendré lors de l'application des garanties à la base Open DAMIR.

Afin de l'observer, les garanties sont appliquées aléatoirement sur les prestations du portefeuille de Generali, en respectant les proportions de chaque formule observées. L'erreur moyenne obtenue en répétant plusieurs fois le tirage aléatoire est égale à 6% du montant remboursé réel. Cette sélection aléatoire répétée revient à calculer pour chaque prestation la moyenne des remboursements générés par les différentes garanties pondérée par les proportions des formules. Il s'agit de la méthode appliquée à la base d'étude pour que le remboursement estimé ne dépende pas d'un choix aléatoire.

- Application des garanties sur un montant agrégé

La base Open DAMIR est agrégée. A la suite de l'application des garanties, les remboursements obtenus sont alors plus élevés que si la base n'était pas agrégée. Par exemple, pour une consultation de médecin généraliste remboursée à hauteur de 125% de la base de remboursement, si une ligne regroupe deux consultations ayant des montants de frais réels égaux à 25€ et 35€, le montant affiché sur la ligne est égal à 60€ et la base de remboursement est égale à 50€. Le montant remboursé calculé sera égal à $125\% \times 50\text{€} = 62,5\text{€}$, dans la limite du montant de frais réels, diminué du montant remboursé par la Sécurité Sociale, soit $\min(125\% \times 50\text{€}; 60\text{€}) - 2 \times 17,50\text{€} = 25\text{€}$. En réalité, le remboursement aurait dû être égal à $[\min(125\% \times 25\text{€}, 25\text{€}) - 17,50\text{€}] + [\min(125\% \times 35\text{€}, 35\text{€}) - 17,50\text{€}] = 21,25\text{€}$. Le montant remboursé calculé est donc supérieur au montant réel.

Une nouvelle fois, le biais engendré est observé sur le portefeuille de Generali. Pour ce faire, la base est agrégée comme la base Open DAMIR et les garanties sont à nouveau appliquées. L'erreur finalement obtenue est égale à 5% du montant remboursé réel.

Il convient de noter que ce second biais réduit l'erreur engendrée par le premier biais souligné.

Rapprochement de la population étudiée à celle du portefeuille de Generali

La population assurée par Generali n'est pas représentative de l'ensemble de la population française, contrairement à la population présente dans la base Open DAMIR. Par exemple, il y a très peu d'assurés de Generali qui résident dans une région ou un département d'Outre-Mer. Des différences importantes dans la population couverte peuvent altérer la comparaison des inflations globales mesurées sur la base d'étude et sur le portefeuille de Generali.

Pour rendre comparables les populations des deux bases de données, des coefficients de pondération sont appliqués aux quantités d'actes et aux montants remboursés pour chaque tranche d'âges, sexe et région de résidence des bénéficiaires des soins de la base d'étude. Ils sont calculés pour chaque modalité des variables concernées en divisant le poids de la modalité dans le portefeuille de Generali par le poids de cette même modalité dans la base d'étude. Ces coefficients mettent notamment en évidence le fait que le portefeuille de Generali contient plus de prestations associées à des hommes que la base Open DAMIR. Il peut être intéressant de mentionner que la base Open DAMIR contient initialement plus de prestations associées à des femmes tandis que le portefeuille de Generali contient plus de prestations associées à des hommes. Cette proportion d'hommes supérieure est donc une caractéristique du portefeuille.

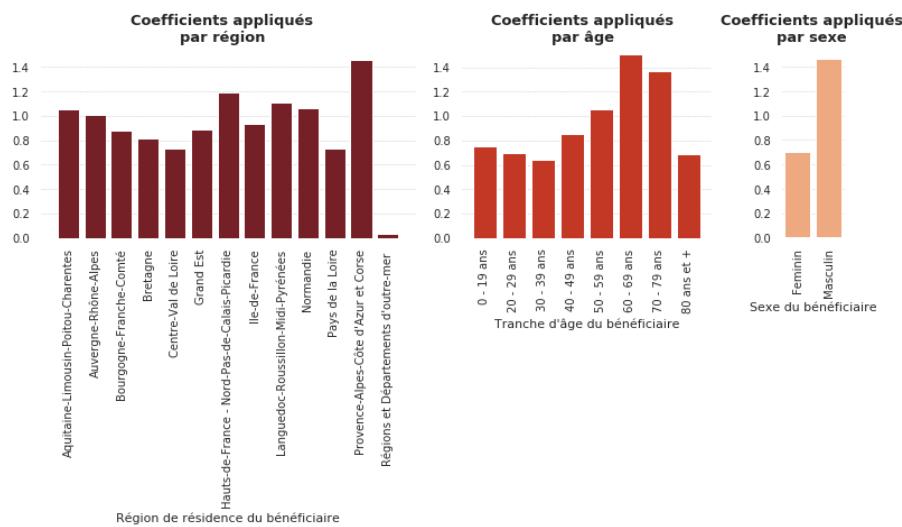


FIGURE 2.12 – Coefficients appliqués aux régions, tranches d'âges et sexes des bénéficiaires des soins la base d'étude

2.3 Statistiques descriptives

Les statistiques descriptives permettent de visualiser les données et peuvent être utiles afin de comprendre et d'interpréter certains résultats. Cette étude se place du point de vue de l'assureur. Le terme "coût" désigne alors le montant remboursé par Generali.

- Évolution des caractéristiques des bénéficiaires des soins

Les proportions d'actes associés à des hommes ou des femmes sont stables dans le temps, comme le montre le premier graphique ci-dessous. De même, la répartition des actes entre les régions est plutôt stable au cours du temps. Il convient de rappeler que le niveau des proportions correspond aux proportions observées sur le portefeuille de Generali.

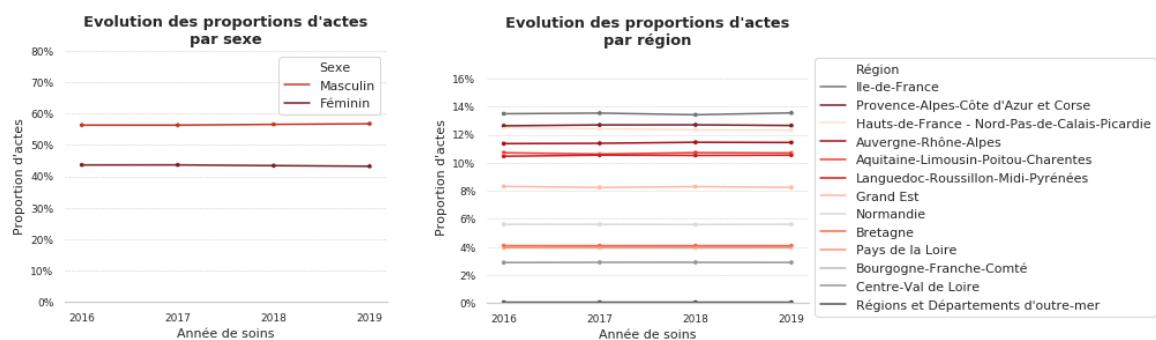


FIGURE 2.13 – Évolution des proportions d'actes par sexe et par région

Concernant l'âge des bénéficiaires des soins, l'âge moyen observé est environ égal à 57 ans. Cet âge est approximé étant donné que la base d'étude ne contient que des tranches d'âges. Le centre de chaque classe est utilisé comme âge associé à chaque bénéficiaire pour calculer l'âge moyen.

L'âge moyen de la base complète augmente de 6 mois entre 2016 et 2019.

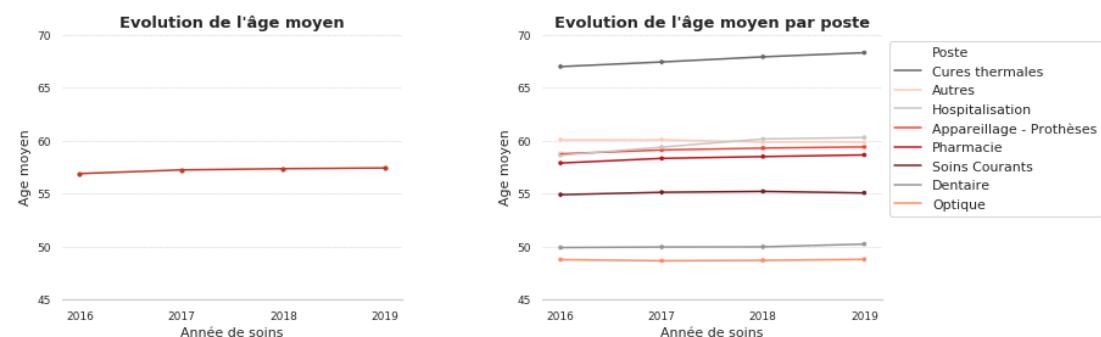


FIGURE 2.14 – Évolution de l'âge moyen

Le graphique de droite ci-dessus reflète la différence d'âge moyen en fonction du poste de soins étudié. Par exemple, les jeunes consomment plus de soins en optique et dentaire que les personnes âgées, ce qui fait baisser l'âge moyen de ces postes. Les prestations de cures thermales sont au contraire associées à des personnes plus âgées.

- Évolution du nombre d'actes

Le nombre d'actes de la base d'étude a diminué au fil des quatre années étudiées. Cette baisse est notamment portée par la pharmacie, qui a vu son nombre d'actes diminuer d'environ 2% chaque année.

La diminution observée entre 2018 et 2019 montre que les données sont incomplètes pour l'année 2019 puisque la baisse observée est de 4% tandis qu'elle était d'environ 1% entre 2016 et 2017 et entre 2017 et 2018. Le poste Soins courants, sur lequel une tendance à la hausse est observée entre 2016 et 2018, voit lui aussi son nombre de remboursements chuter en 2019.



FIGURE 2.15 – Évolution du nombre d'actes

Ce phénomène s'explique par le fait que, bien que les remboursements de l'assurance maladie soient en général effectués sous une semaine avec la carte vitale, un certain nombre de prestations réalisées en 2019 ne seront remboursées qu'en 2020. Le délai de remboursement est notamment plus élevé lorsque le professionnel de santé remet une feuille de soins papier au bénéficiaire. C'est ce que montre la cadence de règlement. En considérant uniquement les règlements de l'année N, 97% des actes réalisés durant l'année N sont présents. Ce pourcentage s'élève à 99% en considérant en plus les remboursements effectués en janvier N+1. Pour avoir plus de 99,9% des actes survenus l'année N, il faut prendre en compte les règlements de janvier N à octobre N+1.

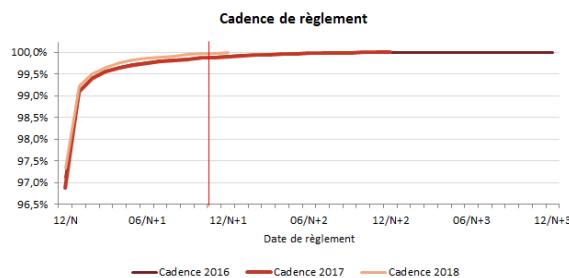


FIGURE 2.16 – Cadence de règlement

Les bases Open DAMIR du premier trimestre de 2020 ne sont pas disponibles. Elles ne seront mises en ligne qu'au début de l'année 2021. Les résultats obtenus en exploitant les données de 2019 sont donc à considérer avec prudence, en gardant à l'esprit qu'environ 3% des données sont manquantes.

- Évolution du coût moyen

Le coût moyen des prestations contenues dans la base d'étude a augmenté entre 2016 et 2019. Il s'agit du phénomène qui sera étudié par la suite. Le graphique de droite permet de mettre en évidence la différence de coût moyen existant entre les différents postes étudiés.

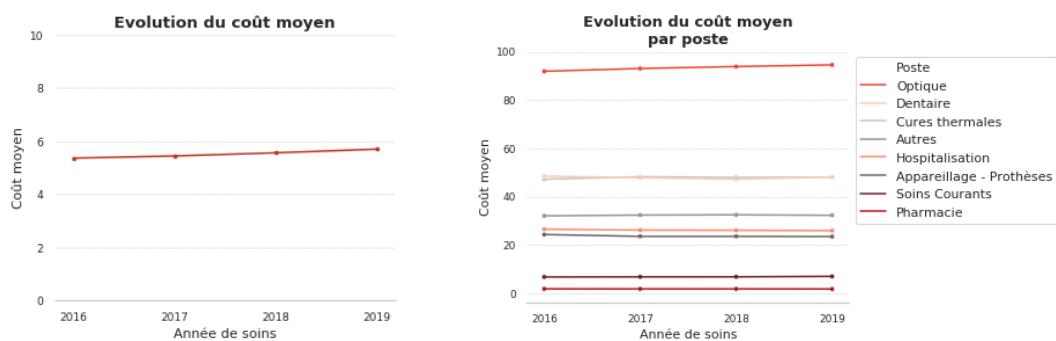


FIGURE 2.17 – Évolution du coût moyen

- Répartition des actes par poste

Le poste Pharmacie est le poste le plus représenté dans la base d'étude en termes de nombre d'actes. Si la répartition par poste est observée en termes de montant versé par Generali, le poste Soins courants est le poste le plus représenté. Cette différence de répartition s'explique simplement par le fait que les actes du poste Pharmacie ont une fréquence plus élevée et un coût moyen plus faible que les Soins courants.

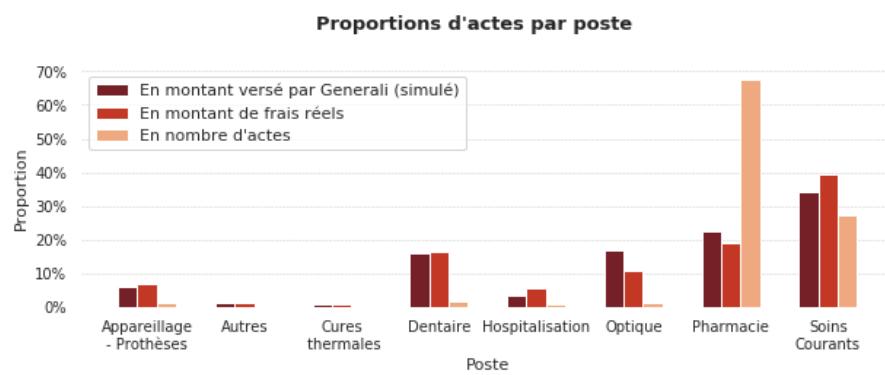


FIGURE 2.18 – Répartition des actes par poste

Chapitre 3

La mesure de l'inflation à partir de données du marché

La base de données construite permet de calculer l'inflation réelle, hors impact de la qualité des données ou des modifications réglementaires. L'utilisation de cet ensemble de données issues du marché permet de disposer d'un volume de prestations important et d'informations supplémentaires par rapport au portefeuille de santé individuelle.

L'inflation est calculée au global sur la base d'étude puis est observée de manière plus fine. Ce chapitre présente les inflations mesurées, la définition de l'inflation retenue pour l'étude ainsi que les différentes variables utilisées.

3.1 Mesure et observation de l'inflation globale

L'inflation est dans un premier temps calculée au global sur la base d'étude afin de permettre à Generali de positionner son portefeuille par rapport aux données nationales. Cette partie a pour but de présenter la méthode de calcul de l'inflation ainsi que les inflations observées et l'impact des retraitements réalisés sur l'inflation.

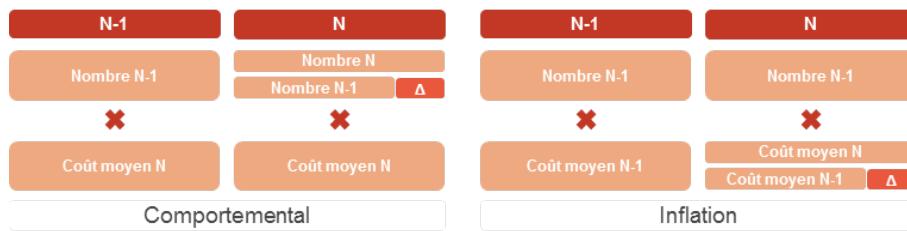
3.1.1 Définition de l'inflation

L'inflation est une composante de la dérive de la sinistralité. Cette dernière correspond à la variation du montant total versé par l'organisme d'assurance au titre des dépenses de santé entre deux années consécutives, à effectif constant. Elle est composée de deux effets : le comportement et l'inflation du coût unitaire.

$$\Delta \text{ coût} = \Delta \text{ comportemental} + \Delta \text{ inflation}$$

L'effet comportemental correspond à la part de la dérive résultant de la variation de la fréquence de consommation des assurés. Le montant de cet effet entre deux années N-1 et N est calculé en considérant le nombre de sinistres en delta au coût moyen de l'année N. Cet effet se mesure à effectif constant, en ne prenant en compte que les assurés présents dans le portefeuille les deux années (N-1 et N), afin de ne pas observer de dérive fictive engendrée par l'évolution du nombre d'assurés. Étant donné que la base Open DAMIR ne comporte pas d'informations relatives aux entrées et sorties des assurés, cet effet n'est pas étudié.

L'effet inflation correspond à la part de la dérive résultant de la variation du montant remboursé moyen. Le montant de cet effet entre N-1 et N est calculé en considérant le nombre de sinistre en N-1 et en leur appliquant la variation de coût moyen.



L'inflation globale peut être calculée en divisant l'effet inflation défini ci-dessus par le coût total de l'année N-1. Cette définition de l'inflation correspond à la formule suivante :

$$Inflation_{N/N-1} = \frac{(Coût moyen_N - Coût moyen_{N-1}) \times Nombre_{N-1}}{Coût total_{N-1}} \quad (3.1)$$

Ce qui est équivalent à écrire :

$$Inflation_{N/N-1} = \frac{(Coût moyen_N - Coût moyen_{N-1})}{Coût moyen_{N-1}} \quad (3.2)$$

Cette définition de l'inflation est cependant très dépendante de l'évolution de la composition du portefeuille puisque le coût moyen concerne toutes les prestations du portefeuille. Afin de limiter l'impact du mix du portefeuille, l'effet inflation est calculé par famille de remboursements. L'inflation globale correspond alors à la somme des effets inflations pour toutes les familles de remboursements, ramenée au coût total de l'année N-1. La formule exacte est donc la suivante : pour chaque famille de remboursements i appartenant à l'ensemble des familles de remboursements Ω :

$$Effet inflation_{i,N/N-1} = (Coût moyen_{i,N} - Coût moyen_{i,N-1}) \times Nombre_{i,N-1} \quad (3.3)$$

$$Inflation_{N/N-1} = \frac{\sum_{i \in \Omega} Effet inflation_{i,N/N-1}}{Coût total_{N-1}} \quad (3.4)$$

3.1.2 Observation de l'inflation sur la base d'étude

Étant donné que la base d'étude contient les données de 2016 à 2019, trois inflations globales peuvent être calculées. L'inflation entre 2016 et 2017 est égale à 0,16%, l'inflation entre 2017 et 2018 est égale à 0,53% et l'inflation entre 2018 et 2019 est égale à 0,10%. Les graphiques ci-dessous détaillent la composition des inflations en fonction des différents postes. Le nombre de sinistres en N-1 multiplié par le coût moyen en N est appelé montant N recalculé.

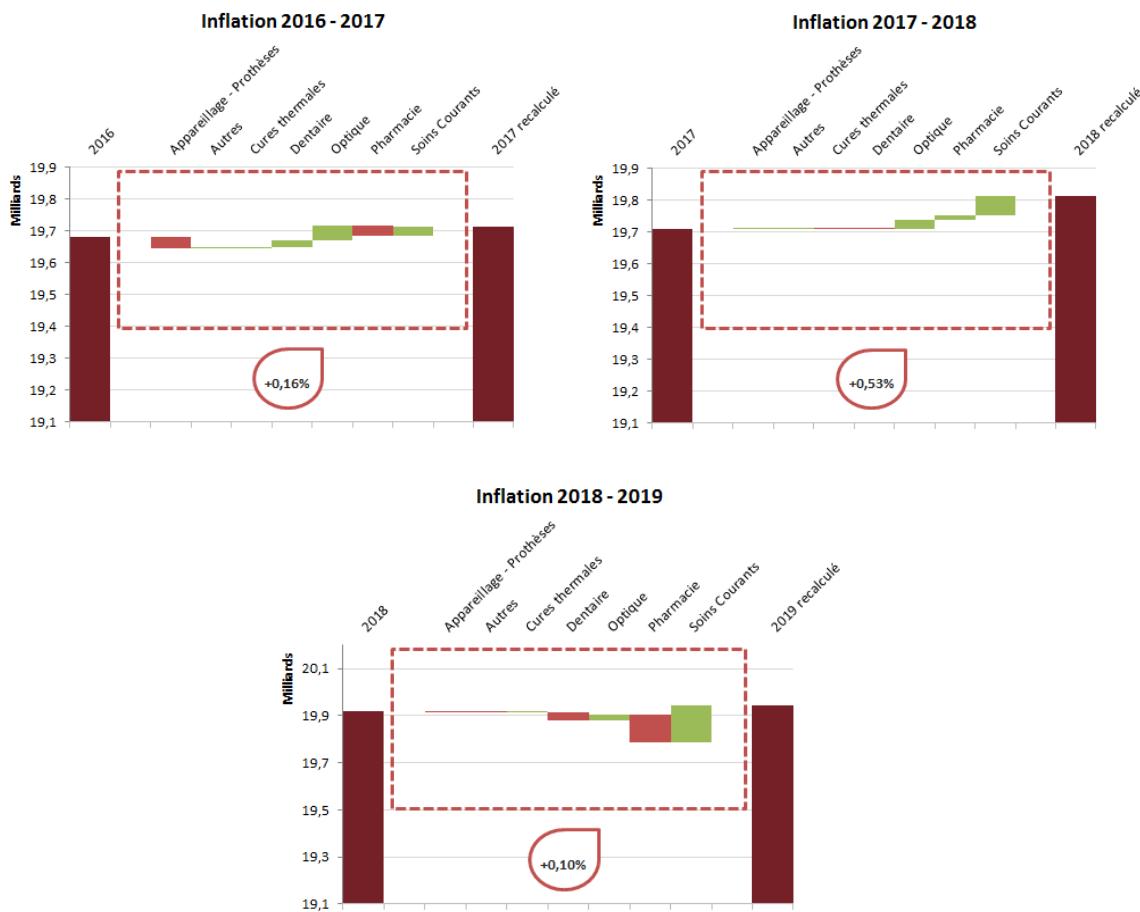


FIGURE 3.1 – Décomposition des effets inflation par poste

Ces résultats permettent dans un premier temps de positionner le portefeuille de santé individuelle de Generali par rapport aux données nationales. L'inflation issue du portefeuille de Generali ne sera cependant pas mentionnée dans ce mémoire par soucis de confidentialité.

Comme mentionné précédemment, environ 3% des données de l'année 2019 sont manquantes. Par prudence, l'inflation 2018-2019 ne sera donc pas étudiée par la suite.

Afin de mesurer l'impact des données manquantes sur la mesure de l'inflation, les inflations 2016-2017 et 2017-2018 sont calculées en se plaçant dans la situation du calcul de l'inflation 2018-2019, c'est-à-dire en ne considérant que les prestations réglées avant la fin de l'année N. L'inflation est ensuite calculée en prenant un mois de recul supplémentaire, puis deux mois de recul supplémentaires, etc., afin d'observer le recul à partir duquel l'inflation se stabilise. La courbe obtenue pour l'inflation 2017-2018 en fonction du recul est présentée ci-dessous. En ne considérant aucune donnée avec une date de règlement postérieure à 2018, l'inflation obtenue est égale à 0,50%. L'impact des données postérieures à 2018 correspond donc à une augmentation de 0,03% de l'inflation. L'inflation semble se stabiliser à partir de l'utilisation des données de mai 2019.

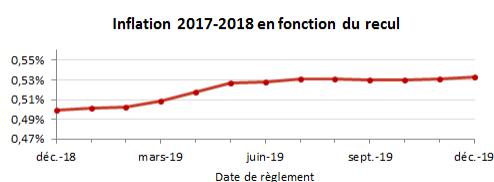


FIGURE 3.2 – Inflation 2017-2018 en fonction du recul

Concernant l'inflation 2016-2017 en fonction du recul, la courbe est inversée et l'inflation diminue de 0,05%. Il peut être supposé que l'inflation 2018-2019 est amenée à évoluer d'environ 0,05% à la hausse ou à la baisse en considérant les prestations survenues en 2019 et réglées en 2020.

3.1.3 Observation de l'impact des retraitements réalisés

Les différents retraitements effectués lors de l'étape de préparation des données ont des impacts sur l'inflation mesurée et ont contribué à sa fiabilité.

La création de familles de remboursements et leur utilisation dans le calcul de l'inflation ont permis de limiter l'impact du mix de prestations. Sans ces familles l'inflation mesurée serait bien plus élevée.



Les retraitements des évolutions réglementaires ont permis de neutraliser l'inflation artificielle engendrée par les modifications légales. Sans ces retraitements, les inflations 2016-2017 et 2017-2018 seraient plus élevées, notamment à cause du changement de la base de remboursement des consultations des médecins généralistes survenu en mai 2017.

L'inflation serait négative pour 2018-2019, principalement à cause de l'entrée en vigueur des premières mesures de la réforme 100% santé sur le poste Dentaire.



L'application des coefficients de pondération a permis de rapprocher les caractéristiques de la population présente dans la base d'étude de celles des assurés de Generali. L'inflation observée est alors plus adaptée à la population du portefeuille. En particulier, il sera observé par la suite que l'inflation observée sur les âges faibles est plus élevée. Comme la base Open DAMIR contient une proportion plus importante d'actes associés à des âges faibles que le portefeuille de Generali, il en résulte que l'inflation observée sur la base Open DAMIR sans l'application des coefficients de pondération serait plus élevée.



3.2 Définition des variables de l'étude

L'objectif de ce mémoire est d'étudier l'inflation sur la base nationale et d'identifier des variables qui l'influencent. L'inflation est alors calculée sur la base d'étude de manière plus fine. Il n'est cependant pas envisageable de la calculer pour chaque ligne de l'ensemble de données puisqu'il est nécessaire d'avoir un nombre d'actes assez élevé pour que l'inflation mesurée soit stable. Les lignes de la base sont alors agrégées selon une certaine granularité afin de calculer l'inflation. Des variables explicatives sont construites à l'aide des informations disponibles dans la base Open DAMIR ou introduites à partir de données en Open Data afin d'être en mesure d'identifier des phénomènes influençant l'inflation.

3.2.1 Le choix de la variable à expliquer et des données utilisées

Avant de présenter la manière dont l'inflation est étudiée, il peut être intéressant de donner quelques éléments permettant de comprendre le choix du phénomène observé, à savoir l'inflation elle-même, et de l'utilisation de données externes.

L'inflation de la sinistralité correspond à la variation du coût moyen des dépenses de santé des assurés. Cette définition peut laisser penser que le coût moyen pourrait être étudié puisqu'il permet de déduire l'inflation, d'autant plus que les modèles permettant d'expliquer le coût moyen sont plus classiques que les modèles permettant d'expliquer l'inflation. Cependant, un modèle basé sur le coût moyen permettrait d'identifier les variables ayant un impact sur le niveau de ce coût et pas sur sa variation. L'objectif de ce mémoire n'est pas d'identifier les dépenses de santé qui coûtent le plus cher mais d'observer les facteurs ayant un impact sur la variation du coût moyen. Un modèle permettant d'expliquer le coût moyen n'est par conséquent pas adapté. L'étude réalisée porte donc directement sur l'inflation.

La difficulté résidant dans l'explication de l'inflation provient du fait qu'elle peut dépendre d'une multitude de facteurs qui ne sont pas forcément observables. En particulier, l'inflation peut être liée au comportement des patients, que ce soit en raison de leur état de santé ou à cause des incitations de l'Etat. Par exemple, des dispositifs d'incitation ont été mis en place en faveur des médicaments génériques, comme notamment la campagne nationale d'information « Devenir générique ça se mérite » lancée en 2016 dans le but d'accentuer la prescription et l'utilisation des génériques, vendus moins cher que les principes. L'inflation peut également être influencée par le contexte économique, par l'offre de soins disponible ainsi que par le développement de nouvelles technologies innovantes et onéreuses.

L'utilisation de données en Open Data a été envisagée afin d'identifier des variables qui ne sont pas forcément contenues dans le portefeuille de Generali mais qui influencent l'inflation. Il faut cependant avoir conscience que de nombreuses variables influençant l'inflation ne sont pas disponibles. L'objectif est d'utiliser les variables à disposition afin de pouvoir améliorer le suivi de l'inflation.

3.2.2 La définition de l'inflation à étudier

L'inflation est calculée sur la base d'étude selon une granularité plus fine que l'inflation globale dans le but d'être analysée. L'inflation peut être observée à différents niveaux, tout en prenant en compte les familles de remboursements dans le calcul pour limiter l'impact de l'évolution de la composition du portefeuille. Afin d'avoir une stabilité, il est toutefois nécessaire de regrouper un certain nombre d'actes sur une ligne avant de calculer l'inflation.

La granularité de calcul de l'inflation retenue

- L'inflation en fonction des postes de soins

L'inflation est différente en fonction du poste de soins, comme le montre la figure ci-dessous. Elle diffère également en fonction de la famille de remboursements, c'est pourquoi la granularité retenue pour le calcul de l'inflation contient la famille de remboursements.

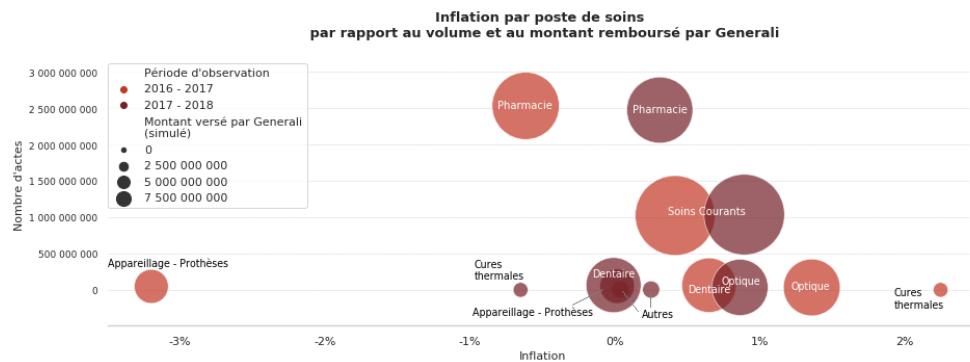


FIGURE 3.3 – Inflation par poste

- L'inflation en fonction des caractéristiques des bénéficiaires de soins

Concernant le sexe, la différence observée entre les hommes et les femmes n'est pas significative sur la base d'étude complète. Elle est cependant significative sur certains postes de soins, notamment sur le poste Soins courants. Concernant l'âge, l'inflation est plus élevée pour les âges faibles, que ce soit sur la base d'étude complète ou sur les différents postes de soins. La région est significative pour trois postes parmi les sept, en particulier pour les Soins courants. Ces observations conduisent à calculer l'inflation en conservant les âges, les sexes et les régions de résidence des bénéficiaires des soins.

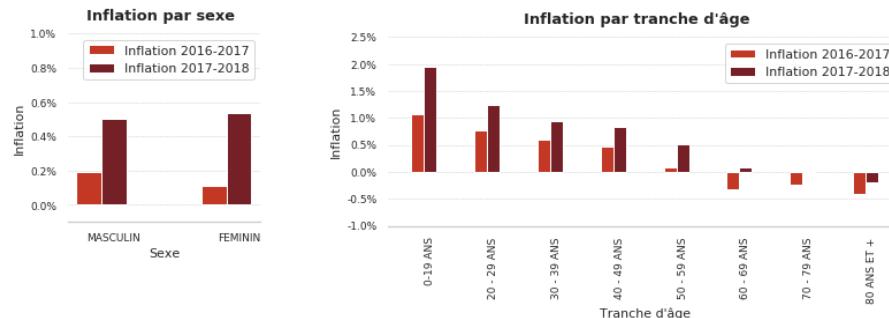


FIGURE 3.4 – Inflation en fonction de l'âge et du sexe



FIGURE 3.5 – Inflation 2016-2017 par région

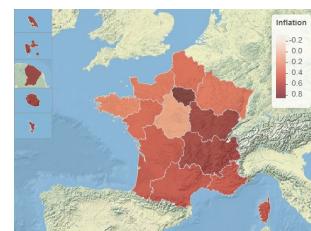


FIGURE 3.6 – Inflation 2017-2018 par région

Les inflations 2016-2017 et 2017-2018 sont donc calculées en fonction de l'âge, du sexe, de la région de résidence et de la famille de remboursements.

La stabilité de l'inflation obtenue avec la granularité âge, sexe, région et famille de remboursements

En calculant l'inflation avec la granularité retenue, des valeurs pouvant être considérées comme extrêmes apparaissent. Elles sont dues aux combinaisons des variables de la granularité qui possèdent peu d'actes, pour lesquelles le coût moyen varie fortement.

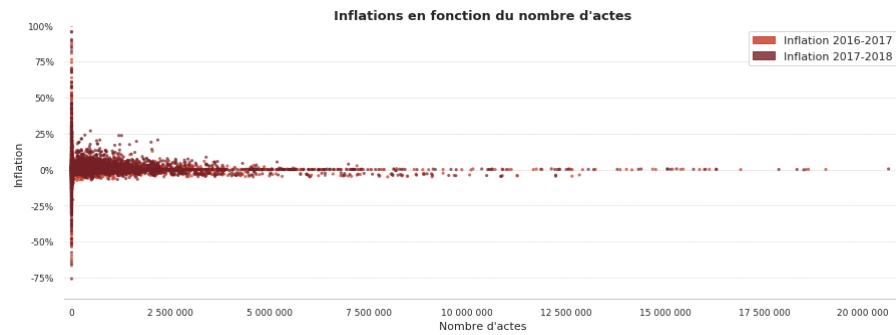


FIGURE 3.7 – Inflation, calculée avec la granularité retenue, en fonction du nombre d'actes

Ces valeurs concernent certaines familles de remboursements auxquelles peu d'actes sont associés.

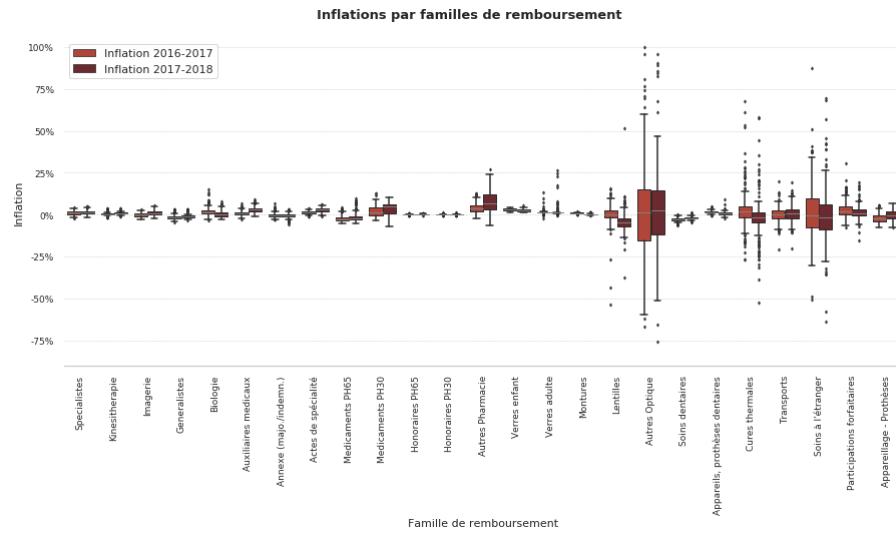


FIGURE 3.8 – Dispersion de l'inflation, calculée avec la granularité retenue, par famille de remboursements

La variable à expliquer

La granularité retenue contient les variables âge, sexe, région et famille de remboursements. Pour chacune des années appartenant au périmètre de l'étude, les lignes dont les modalités de ces variables sont identiques sont regroupées pour le calcul de l'inflation.

Le paragraphe précédent mentionne le fait que les variables en fonction desquelles l'inflation diffère ne sont pas les mêmes selon le poste de soins observé. Par exemple, la région semble être une variable importante seulement pour quelques postes de soins. Ce phénomène s'avère aussi vrai au niveau des familles de remboursements. En étudiant l'inflation sur la base complète, les variables les plus importantes conduisent à une discrimination des différentes familles. Ce résultat découle notamment du fait que les variables explicatives ne sont pas pertinentes pour chacune des familles de remboursements. Ce point sera détaillé lors de la présentation des variables explicatives. Il est alors décidé de réaliser une étude de l'inflation pour chaque famille de remboursements. L'objectif est ainsi d'expliquer l'inflation mesurée pour différents profils d'individus sur chaque famille de remboursements.

3.2.3 Les variables explicatives

Il est nécessaire de disposer de variables explicatives pour être en mesure d'analyser l'inflation observée. Les informations disponibles dans la base Open DAMIR sont utilisées et des variables externes sont ajoutées.

Utilisation des variables de la base Open DAMIR

La base de données construite grâce aux étapes du chapitre 2 est agrégée en fonction de la granularité choisie. Certaines variables n'appartenant pas à cette granularité sont modifiées afin d'être conservées pour l'étude de l'inflation.

- Professionnels de santé exécutants et prescripteurs, Top établissement et code secteur privé/public

La base d'étude contient les variables "Exécutant", "Prescripteur" et "Top établissement et code secteur" qui n'ont pas été retenues dans la granularité utilisée pour le calcul de l'inflation. En conservant ces variables dans la granularité, le nombre d'actes sur de nombreuses lignes serait relativement faible et le coût moyen serait instable. Afin d'utiliser ces variables, les proportions de chaque modalité sont associées à chaque ligne de la granularité. Par exemple, une variable "Proportion de prescripteurs qualifiés en Médecine générale" est créée à partir de la variable "Prescripteur" et de la modalité "Médecine générale" en calculant pour chaque ligne le ratio du nombre d'actes prescrits par des médecins généralistes sur le nombre d'actes total de la ligne.

Selon la famille de remboursements étudiée, les proportions associées à des modalités sur ou sous-représentées ne sont pas créées, puisque ces variables seraient difficilement interprétables. Cela permet également de ne pas disposer de multicolinéarité parfaite dans les données, c'est-à-dire de variables explicatives correspondant à une combinaison linéaire des autres variables.

Lors du calcul de l'inflation à partir de la base d'étude agrégée selon la granularité, deux lignes sont regroupées pour calculer une inflation, puisque l'inflation correspond à la variation du coût moyen des dépenses de santé des assurés entre deux années consécutives notées N-1 et N. Les variables utilisées comme granularité sont identiques pour les deux lignes rassemblées. Les variables ajoutées ne sont elles pas identiques. Pour chaque ligne, la variable relative à l'année N-1 est alors conservée et une seconde variable correspondant à la variation entre les deux années est créée.

- Taux de changements de région

La variable "Taux de changements de région", créée à l'aide des variables relatives aux régions de la base Open DAMIR, caractérise les régions et diffère pour chaque année. De la même manière que pour les variables précédentes, le taux relatif à l'année N-1 et la variation de ce taux entre les deux années sont utilisés comme variables explicatives.

- Nombre d'actes complémentaires et de compléments d'actes

Certaines lignes de la granularité possèdent plus ou moins de compléments d'actes (majorations de nuit, d'urgence, de dimanche et de jour férié) ou de remboursements complémentaires (prises en charge supplémentaires pour l'Alsace Moselle, suppléments liés à la prévention, etc.). Ces informations sont alors conservées en utilisant pour chaque ligne le ratio de compléments d'acte par rapport au nombre d'actes total et le ratio d'actes complémentaires par rapport au nombre d'actes. De la même manière que pour les variables précédentes, le ratio relatif à l'année N-1 et la variation entre les deux années sont utilisés.

Introduction de variables externes

Afin d'enrichir la base d'étude, des données externes sont ajoutées. Elles proviennent principalement de l'INSEE (Institut National de la Statistique et des Études Économiques) et de la DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques) et sont accessibles en ligne.

Ces données permettent de caractériser les années et les différents profils d'individus obtenus en agrégeant la base de données selon l'âge, le sexe et la région. Elles ne sont cependant pas toutes disponibles à la granularité utilisée. En fonction de leur disponibilité, ces données caractérisent alors une ou plusieurs des variables de la granularité.

En particulier, certaines données ne sont disponibles qu'en fonction de la région, c'est-à-dire sans distinction de l'âge ni du sexe et pour une unique année. Elle sont toutefois utilisées dans le but d'identifier les spécificités de chaque région et ainsi d'essayer de comprendre pourquoi les régions ont une influence sur l'inflation.

- Variables disponibles en fonction de l'année

Lorsque les données sont disponibles pour chacune des années étudiées, le niveau associé à l'année N-1 et la variation entre deux années sont utilisés, de la même manière que pour les variables issues de la base Open DAMIR.

- Densité de population

L'estimation de la population par région, sexe et âge quinquennal est téléchargée sur le site de l'INSEE. La superficie des régions en kilomètres carrés est également récupérée. La densité de population par région est ensuite calculée pour chaque sexe et tranche d'âges. Cette variable est disponible en fonction de chaque variable de la granularité, mais diffère principalement en fonction des tranches d'âges et des régions. La densité permet notamment de différencier les régions très peuplées et les régions peu peuplées.

- Indice de vieillissement

L'indice de vieillissement correspond au nombre de personnes âgées de plus de 65 ans pour 100 personnes âgées de moins de 20 ans. Il est calculé à l'aide des données démographiques par région, par sexe et par année d'étude. Un indice élevé indique une population plus vieillissante.

- Espérance de vie

L'espérance de vie à l'âge x représente la durée de survie moyenne à l'âge x, en nombre d'années.

L'espérance de vie à la naissance et à 60 ans sont disponibles sur le site de l'INSEE pour chacune des années étudiées et en fonction du sexe et de la région. Ces deux indicateurs sont cependant fortement corrélés, comme le montre la figure ci-contre. Seule l'espérance de vie à 60 ans est alors retenue.

Cette variable contribue à la caractérisation des régions, mais elle diffère principalement en fonction du sexe puisque les hommes ont une espérance de vie moyenne bien plus faible que celle des femmes.

- Taux de chômage

Le taux de chômage est le pourcentage de chômeurs dans la population active. Il est observé par région pour chacune des années étudiées. Cette variable permet de caractériser le marché du travail et peut donner une indication sur le niveau de vie de la population.

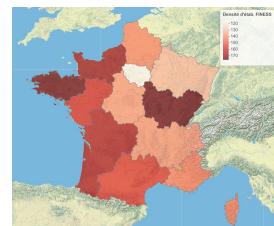


FIGURE 3.9 – Relation entre les espérances de vie à la naissance et à 60 ans

- Densité d'établissements recensés dans le FINESS, de centres de santé et de centres hospitaliers pour 100 000 habitants

Les informations issues du FINESS (Fichier National des Établissements Sanitaires et Sociaux) sont utilisées afin de caractériser l'offre de soins disponible dans les différents régions. Le FINESS est un répertoire national mis en œuvre par la DREES. Il recense l'ensemble des structures et équipements des domaines sanitaire, médico-social et social. Ce répertoire est disponible au téléchargement sur la plateforme ouverte des données publiques françaises du Ministère des Solidarités et de la Santé (sur *data.gouv.fr*). L'historique est disponible pour chaque année de 2004 à 2019.

Ce fichier contient la liste des établissements avec notamment leur adresse et leur catégorie d'établissement. Pour chacune des années étudiées, le nombre d'établissements contenus dans ce fichier est évalué par région. Afin de ne pas être biaisé par les écarts de population entre les régions, ce nombre d'établissements est calculé pour 100 000 habitants, en utilisant la population de l'INSEE. La densité d'établissements du FINESS par région est présentée sur la carte ci-contre à titre illustratif.



Elle peut refléter la capacité des habitants à assumer des frais de santé plus ou moins élevés.

— Part de la population vivant dans une aire urbaine

L'environnement dans lequel vit une personne peut fournir une information relative à l'accès au soins. Le pourcentage de la population vivant dans une aire urbaine est alors utilisé pour chaque région. Cette variable décrit le degré d'urbanisation de chaque région.

— Taux de bon niveau VQS et taux de maladies chroniques

Les données issues de l'enquête Vie Quotidienne et Santé (VQS) de 2014 sont utilisées afin de construire des indicateurs caractérisant les régions et les sexes. L'enquête VQS est un court questionnaire produit par la DREES s'adressant aux personnes de plus de 60 ans vivant en France. Son principal objectif est de mieux connaître l'état de santé des seniors. Les réponses à ce questionnaire sont différencierées en fonction de la région et du sexe des personnes âgées.

De nombreuses variables peuvent être construites à partir de ces réponses. Deux indicateurs sont retenus pour cette étude. Le premier est appelé "Taux de bon niveau VQS". Il est construit à partir du score de dépendance présumée calculé par la DREES à l'issue du questionnaire. Ce score est composé de quatre groupes caractérisant le niveau général d'état de santé des personnes ayant répondu à l'enquête. Les groupes VQS vont de I (bon niveau général d'état de santé) à IV (mauvais niveau général d'état de santé). Le taux de bon niveau VQS construit pour l'étude correspond à la proportion de personnes ayant répondu au questionnaire dont le groupe VQS attribué est le groupe I. Autrement dit, il s'agit du taux de personnes ayant un bon niveau général d'état de santé.

Le second indicateur créé est appelé "Taux de maladies chroniques VQS". Il est construit à partir des réponses à la question "La personne a-t-elle une maladie ou un problème de santé qui soit chronique ou de caractère durable?", en divisant le nombre de personnes ayant répondu "Oui" par le nombre de personnes ayant répondu à l'enquête.

— Taux de mortalité

Le taux de mortalité est le rapport du nombre de décès de l'année sur la population totale moyenne de l'année. Cet indicateur diffère en fonction du sexe et de la région, mais contribue principalement à la caractérisation des régions.

Corrélation entre les variables externes

Par la suite, ces variables seront utilisées dans des forêts aléatoires et des modèles linéaires. La corrélation entre les variables explicatives pose problème dans les modèles linéaires. Bien qu'elle n'influe pas sur les performances d'une forêt aléatoire, elle peut biaiser l'ordre d'importance des variables lors de l'interprétation du modèle. Ainsi, les variables externes ajoutées à la base d'étude sont sélectionnées de manière à ne pas être fortement corrélées entre elles. Le niveau de corrélation maximal retenu est égal à 0,80.

3.2.4 Les familles de remboursements retenues pour l'étude

L'étude de l'inflation est présentée dans ce mémoire pour deux familles de remboursements. Cette partie a pour but d'introduire ces familles et détaille les particularités des bases de données utilisées pour l'étude de chacune d'elles.

La famille des médicaments remboursés à 65% par la Sécurité Sociale

- Le choix de la famille "Médicaments PH65"

La pharmacie est un poste important dans le portefeuille de Generali et dans la base d'étude. Elle représente 67% des actes et 22% des montants remboursés par Generali simulés, comme observé dans la partie 2.3. Ce poste est composé de cinq familles de remboursements : les médicaments remboursés à 65% par la Sécurité Sociale, les honoraires de dispensation associés à ces médicaments, les médicaments remboursés à 30%, les honoraires de dispensation associés à ces médicaments et une famille appelée Autres qui contient les actes de pharmacie n'entrant pas dans les familles précédemment citées, comme les médicaments remboursés à 15% par exemple.

Les honoraires de dispensation sont la rémunération du pharmacien. Ils sont liés au conditionnement des médicaments ou à l'exécution d'une ordonnance. Ils ont pour but de valoriser le rôle de professionnel de santé du pharmacien et sa fonction de conseil lors de la délivrance de médicaments. Grâce à ces honoraires, la rémunération des pharmaciens dépend moins du prix des médicaments vendus. Le prix des honoraires de dispensation est opposable aux pharmaciens et aucun dépassement n'est possible. Ce prix n'a pas évolué depuis le 1er janvier 2016, c'est pourquoi aucune inflation n'est observée sur les deux familles d'honoraires créées.

La famille des médicaments remboursés à 65%, notée "Médicaments PH65", est la famille la plus représentée dans le poste Pharmacie. Elle représente une part importante des prestations de santé puisque 17% des actes de la base de données et 11% des montants remboursés par Generali simulés sont associés à cette famille.

- La base de données pour les "Médicaments PH65"

Une base de données spécifique à chaque famille de remboursements est construite, notamment pour les "Médicaments PH65". Certaines variables ne sont pas pertinentes pour l'étude de l'inflation associée à cette famille. Comme évoqué précédemment, des variables sont construites à partir des informations de la base Open DAMIR, en calculant des proportions de chaque modalité pour chaque ligne de données. Certaines variables sont alors associées à des modalités sur ou sous-représentées, ce qui les rend imprudentes ou difficilement interprétables. Il s'agit notamment du cas des proportions de professionnels de santé exécutants. Pour la pharmacie, l'exécutant est un pharmacien dans plus de 99% des cas. Les variables associées aux proportions d'exécutants ne sont donc pas conservées.

Certaines proportions de professionnels de santé prescripteurs se trouvent également dans cette situation. Elles sont alors soit supprimées ou soit regroupées avec les variables pour lesquelles l'inflation est semblable et le rassemblement semble cohérent. Par exemple, les prescripteurs anesthésistes sont regroupés avec les chirurgiens, tandis que la proportion de prescripteurs pédicures-podologues est supprimée.

- L'inflation globale sur cette famille

Au global, l'inflation observée sur cette famille est égale à -3,07% entre 2016 et 2017, et -2,56% sur la période 2017 - 2018.

La famille des consultations et visites des généralistes

- Le choix de la famille "Généralistes"

Le poste Soins courants est le deuxième poste le plus représenté dans la base d'étude en termes de nombre d'actes et le premier poste en terme de montant remboursé puisqu'il regroupe 39% des montants remboursés simulés. La famille "Généralistes" est la famille de remboursements la plus importante parmi les familles de ce poste. Elle contient les remboursements associés aux consultations et visites des médecins généralistes.

- La base de données pour les "Généralistes"

Comme pour l'étude de l'inflation relative aux médicaments, certaines variables ne sont pas pertinentes pour cette famille de remboursements. En particulier, les proportions de professionnels de santé prescripteurs et exécutants ne sont pas conservées au regard de la définition de cette famille. Concernant les densités de professionnels de santé par région, seule la densité de médecins généralistes est insérée parmi les variables explicatives.

Certaines données sont ensuite exclues de l'étude. L'inflation observée dans les régions et départements d'Outre-mer se démarque de manière extrême de celle mesurée en France métropolitaine. Étant donné que cette région représente moins de 1% du portefeuille de Generali, ces données ne sont pas conservées puisque des valeurs extrêmes pourraient biaiser les conclusions des études réalisées par la suite.

- L'inflation globale sur cette famille

Au global, l'inflation observée sur cette famille est égale à -1,71% entre 2016 et 2017, et -1,33% entre 2017 et 2018.

Chapitre 4

L'apport de la data science pour analyser l'inflation

L'objectif de cette partie est d'identifier sur la base d'étude, construite à partir de données du marché, des variables qui peuvent influencer l'inflation afin de suivre ces variables sur le portefeuille de Generali. L'inflation de la sinistralité est alors étudiée à l'aide de méthodes de Machine Learning pour chaque famille considérée.

4.1 Sélection des modèles

Afin de parvenir à cet objectif, deux modèles ont été sélectionnés. La méthode des forêts aléatoires a été retenue notamment pour les trois raisons suivantes : son caractère automatique puisqu'elle possède peu de paramètres à régler, sa performance notamment pour des problèmes complexes et sa quantification de l'importance des variables. La deuxième méthode retenue est un modèle linéaire. Il s'agit d'un modèle simple d'interprétation qui est régulièrement utilisé en actuariat. Sa mise en place nécessite cependant la vérification de plusieurs hypothèses. Les éléments théoriques sont rappelés afin de comprendre les modèles utilisés.

4.1.1 Le modèle linéaire gaussien

Le modèle linéaire gaussien [2] est une méthode de régression classique qui possède l'avantage d'être facilement interprétable. L'objectif général de la régression est d'expliquer une variable Y , dite variable à expliquer, en fonction de p variables X_1, \dots, X_p , dites variables explicatives. Ces variables sont observées sur n individus.

Le principe du modèle

Le modèle linéaire repose sur l'hypothèse de linéarité de la relation entre les variables explicatives et la variable à expliquer. Il s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + W_i$$

Avec :

- $(x_{i,1}, \dots, x_{i,n})$ les valeurs prises par la variable explicative X_i ,
- $(\beta_0, \beta_1, \dots, \beta_p)$ sont des paramètres inconnus à estimer,
- (W_1, \dots, W_n) des variables aléatoires indépendantes identiquement distribuées vérifiant $E[W_i] = 0$ et $\text{var}(W_i) = \sigma^2 < \infty$.

Le modèle linéaire se réécrit sous la forme matricielle :

$$Y = X\beta + W$$

avec $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$, $W = \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}$, et $X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix}$

Le modèle linéaire gaussien suppose que les résidus sont indépendants et suivent une loi normale de variance σ^2 constante (hypothèse d'homoscédasticité) :

$$(W_1, \dots, W_n) \sim_{i.i.d} \mathcal{N}(0, \sigma^2).$$

Les paramètres $(\beta_0, \dots, \beta_p)$ sont généralement estimés par la méthode des moindres carrés. Ils sont déterminés afin de minimiser la somme des carrés des résidus. L'estimateur de β est ainsi donné par :

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

Afin de pouvoir inverser la matrice $X'X$, il faut que les vecteurs colonnes de la matrice X soient libres. L'absence d'information redondante dans les prédicteurs est donc supposée.

Dans le cas où toutes les variables explicatives sont qualitatives, le modèle appliqué est appelé analyse de la variance. Il permet de comparer les moyennes empiriques de la variable à expliquer observées pour les différentes modalités des variables explicatives. L'analyse de covariance considère une situation plus générale et permet de modéliser une variable quantitative en fonction de variables quantitatives et qualitatives. Ces modèles se situent dans le cadre général du modèle linéaire.

La performance du modèle

La performance permet de comparer des modèles et de mesurer leur pertinence. Le modèle d'explication de Y recherché doit être à la fois performant (résidus les plus petits possibles) et économique (avec le moins possible de variables explicatives).

Deux critères usuels sont présentés.

- Le R^2 ajusté

Le R^2 représente la proportion de variation qui est expliquée par le modèle :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad \text{avec } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Le R^2 est compris entre 0 et 1, par construction. Il ne tient pas compte du nombre de prédicteurs du modèle et est systématiquement plus grand pour le modèle complet.

Une augmentation du R^2 n'est donc pas forcément synonyme d'amélioration du modèle. C'est pourquoi le R^2 ajusté est utilisé. Il s'agit d'une version modifiée du R^2 pour tenir compte du nombre de prédicteurs dans le modèle :

$$R^2_{aj} = \frac{(n-1) R^2 - p}{n-p-1}$$

où p est le nombre de variables explicatives.

Plus le R^2 ajusté est proche de 1, plus le modèle ajuste correctement les données.

- La RMSE : l'erreur quadratique moyenne

La RMSE (*Root Mean Square Error*) permet de mesurer l'erreur en prédiction d'un modèle. Pour la calculer, il est nécessaire de diviser l'échantillon de données en deux parties : un ensemble d'apprentissage et un ensemble de test. L'erreur quadratique moyenne est calculée sur l'ensemble de test comme suit :

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

où \hat{y}_i est la prédiction obtenue avec le modèle ajusté sur l'ensemble de test, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ et n est la taille de l'ensemble de test.

D'autres critères classiques proches de la RMSE existent, comme la MAE (*Mean Absolute Error*) correspondant à la moyenne arithmétique des valeurs absolues des écarts. La RMSE possède cependant l'avantage de pénaliser plus fortement les larges erreurs, ce qui s'avère être une particularité intéressante pour l'étude.

Ces indicateurs ne s'appliquent pas uniquement aux modèles linéaires. Ils peuvent alors permettre de comparer différents modèles.

4.1.2 Les forêts aléatoires

Les forêts aléatoires sont l'un des algorithmes de Machine Learning les plus populaires, notamment grâce au caractère automatique de la méthode qui possède peu de paramètres à régler. Cet algorithme est utilisable tant en régression qu'en classification et offre en général une bonne performance prédictive notamment pour des problèmes complexes comme des relations non linéaires, des interactions, etc. L'une de ses caractéristiques les plus importantes ayant contribué à sa notoriété est sa quantification de l'importance des variables.

L'algorithme est présenté dans le cadre de la régression dans ce mémoire. L'objectif est d'expliquer une variable quantitative Y , dite variable à expliquer, en fonction de p variables X_1, \dots, X_p , dites variables explicatives. Pour simplifier cette partie, les variables explicatives sont supposées quantitatives mais la méthode se généralise pour des variables qualitatives.

Les forêts aléatoires sont une méthode de statistique non-paramétrique introduite par Breiman [5] en 2001. Cette méthode est basée sur des agrégations d'arbres de décision dont les réponses sont combinées pour obtenir une estimation de la variable à expliquer.

Les arbres de décision

Les arbres de décisions permettant d'expliquer une variable quantitative sont appelés des arbres de régression. L'algorithme CART (*Classification And Regression Trees*) est une méthode introduite par Breiman et al. [15] permettant de construire des arbres de décisions. Le principe de cet algorithme [18] est de construire l'arbre de décision par partitionnements successifs de l'espace des entrées. La construction se fait en deux étapes : la construction de l'arbre maximal puis l'élagage.

La construction de l'arbre maximal consiste à découper successivement des parties de l'espace en deux sous-parties. L'algorithme découpe tout d'abord la racine de l'arbre notée η_1 , contenant toutes les observations de l'échantillon d'apprentissage L_n . Pour chaque indice $j \in \{1, \dots, p\}$ et toute valeur $d \in \mathbb{R}$, l'espace est partitionné en deux classes $\{X^j \leq d\}$ et $\{X^j > d\}$ et les éléments de L_n sont répartis dans les deux classes. L'algorithme sélectionne le couple (j, d) qui minimise une fonction de coût.

Dans le cas de la régression, la fonction à minimiser est la variance intra-groupes résultant de la découpe d'un noeud en deux :

$$C(j, d) = \sum_{i \in \eta_{1,-}(j, d)} (y_i - \bar{y}_-)^2 + \sum_{i \in \eta_{1,+}(j, d)} (y_i - \bar{y}_+)^2$$

Où :

$$\eta_{1,-}(j, d) = \left\{ i \in \{1, \dots, n\} : x_i^j \leq d \right\} ; \quad \eta_{1,+}(j, d) = \left\{ i \in \{1, \dots, n\} : x_i^j > d \right\}$$

et :

$$\bar{y}_- = \frac{1}{Card(\eta_{1,-}(j, d))} \sum_{i \in \eta_{1,-}(j, d)} y_i ; \quad \bar{y}_+ = \frac{1}{Card(\eta_{1,+}(j, d))} \sum_{i \in \eta_{1,+}(j, d)} y_i$$

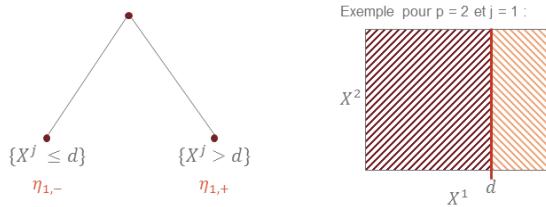


FIGURE 4.1 – Découpe de la racine de l’arbre de décision (à gauche) et exemple de partition associée dans l’espace des variables explicatives (à droite)

De la même manière, la meilleure découpe est recherchée dans chacun des noeuds fils, et ainsi de suite.

Les arbres sont développés jusqu’à atteindre une condition d’arrêt : soit quand le nombre d’individus dans une classe est inférieur à un seuil choisi, soit quand le noeud obtenu est un noeud pur, c’est-à-dire que la variable Y est la même pour tous les individus de la classe. L’arbre pleinement développé est appelé T_{max} et les noeuds terminaux sont appelés feuilles de l’arbre. La valeur de Y pour chaque famille η de T_{max} est estimée par :

$$\frac{1}{Card(\eta)} \sum_{i \in \eta} y_i.$$

Si l’arbre de décision est laissé grandir, il pourra contenir une feuille par donnée de sortie, ce qui signifie que la base d’apprentissage sera parfaitement décrite, le biais sera nul, mais l’arbre sera fortement dépendant des données d’apprentissage et la variance sera très forte. C’est pourquoi il est possible d’élagger l’arbre à posteriori ; c’est-à-dire de supprimer des feuilles en regroupant les données. L’arbre constitué uniquement de la racine aura au contraire une très petite variance mais un biais élevé.

L’élagger est la deuxième étape de l’algorithme CART. Il consiste à sélectionner le meilleur sous-arbre élagué à partir de T_{max} . Le but de l’élagger est d’obtenir un compromis biais-variance¹ acceptable. L’arbre optimal est l’arbre minimisant l’erreur de prédiction sur la base de validation.

Le principal inconvénient des arbres de décision est qu’ils peuvent dépendre fortement des données d’apprentissage choisies. Autrement dit, ils sont sensibles à l’ajout de nouvelles valeurs dans les modèles. C’est la raison pour laquelle la méthode du *Bagging* est utilisée.

La méthode de *Bagging*

Le *Bagging*, contraction des mots *Bootstrap* et *Aggregating*, a été introduit par Breiman en 1994 [4]. Le principe de cette méthode est de tirer B échantillons *bootstrap*, d’entraîner un arbre de décision sur chacun d’eux pour obtenir une collection de prédicteurs, puis d’agréger ces prédicteurs.

1. Une description du compromis biais-variance est donnée dans la source suivante [10].

Chaque arbre construit n'est pas élagué, puisque les arbres CART non élagués ont un biais faible. Dans le cas d'une régression, la valeur prédite est la moyenne de la valeur prédite par chacun des B arbres. Pour chaque tirage, les individus qui ne sont pas dans l'échantillon sont utilisés pour mesurer l'erreur de prédiction du modèle. Le but de cet algorithme est d'augmenter la précision des modèles tout en réduisant leur variance.

La méthode des forêts aléatoires

En 2001, Breiman [5] a proposé une amélioration du *Bagging* en ajoutant une randomisation dans le but de rendre les prédicteurs moins corrélés et ainsi de réduire la variance du modèle.

Le principe de la construction de forêts aléatoires est de générer plusieurs échantillons *bootstrap* (comme dans le *Bagging*) puis d'appliquer une variante de l'algorithme CART. Pour cette variante, des variables explicatives sont sélectionnées aléatoirement pour générer les arbres individuels. La meilleure découpe est recherchée uniquement parmi les variables sélectionnées. Le prédicteur est finalement obtenu en agrégant la collection d'arbres obtenus.

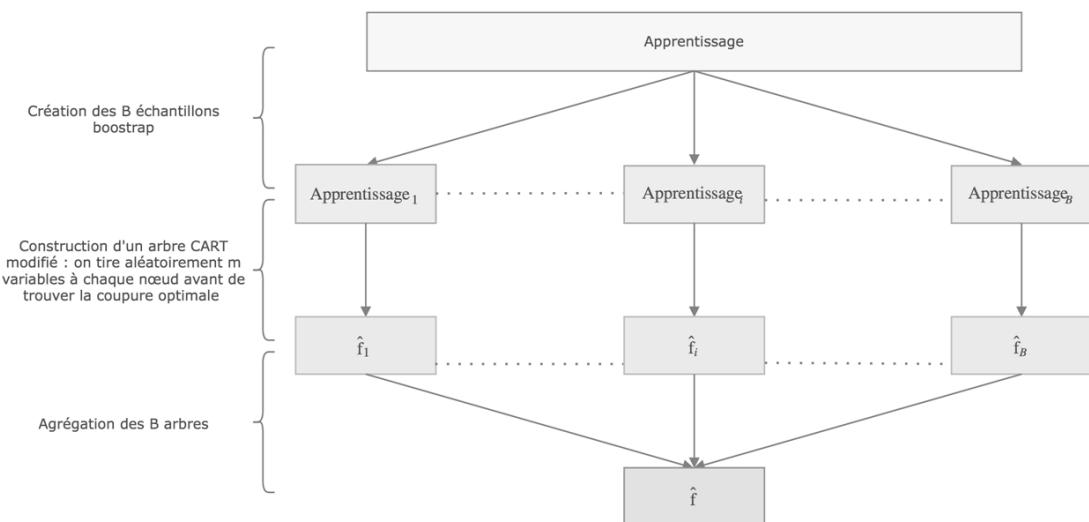


FIGURE 4.2 – Principe des forêts aléatoires - Source [3]

L'algorithme des forêts aléatoires a été importé dans le logiciel R via le package *randomForest*. Il existe deux paramètres principaux pour cette méthode.

- **Le nombre m de variables choisies aléatoirement à chacun des nœuds des arbres :**
Ce nombre est fixé au début de la construction de la forêt. Il peut varier de 1 à p .

Lorsque $m = p$, la méthode des forêts aléatoires correspond à la méthode de *Bagging*. Lorsque $m = 1$, les deux méthodes sont très différentes puisque chaque nœud est découpé en fonction d'une seule variable qui est choisie aléatoirement. Ainsi, plus m est petit, plus la création des arbres est aléatoire. La corrélation entre les arbres et la variance du modèle sont donc plus faibles mais le biais de chaque arbre est fort. Ce paramètre est par conséquent le plus important. Il est nommé *mtry* dans le package *randomForest* de R et possède une valeur par défaut égale à $p/3$ en régression.

— **Le nombre d'arbres B de la forêt :**

Plus ce nombre est grand, plus le modèle est performant. Il faut donc retenir un nombre d'arbres suffisant pour obtenir un bon modèle avec un temps de calcul raisonnable. Ce paramètre est appelé *ntree* dans le package *randomForest* et sa valeur par défaut est égale à 500.

D'autres paramètres existent pour cette méthode comme le nombre minimum d'individus par feuille pour qu'il n'y ait plus de scissions possibles avec la création d'un nouveau noeud, par défaut égal à 5 en régression. Seuls les paramètres *mtry* et *ntree* sont optimisés dans ce mémoire. Ils sont optimisés de manière à minimiser l'erreur *Out-of-Bag*.

L'erreur OOB : *Out-of-Bag*

Out-Of-Bag signifie "en dehors du *bootstrap*" [16]. Soit une observation (X_i, Y_i) de l'ensemble d'apprentissage L_n et soit I_i l'ensemble des arbres de la forêt qui ne contiennent pas cette observation dans leur échantillon *bootstrap*, c'est-à-dire pour lesquels cette observation est *Out-Of-Bag*. Pour estimer la prévision de la forêt \hat{Y}_i de Y_i , uniquement les arbres de I_i sont agrégés. Après avoir effectué cette opération pour toutes les données de L_n , l'erreur *Out-Of-Bag* est alors définie par :

$$Err_{OOB} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

L'avantage de l'erreur OOB par rapport aux autres méthodes classiques d'estimation de l'erreur de généralisation est qu'il n'est pas nécessaire de découper la base d'apprentissage pour créer une base de validation puisqu'un découpage est déjà effectué en tirant l'échantillon *bootstrap* lors de la construction de chaque arbre. Il convient cependant de remarquer que cette erreur n'utilise pas les prédictions de la forêt aléatoire construite. Elle utilise celles de prédicteurs qui sont des agrégations d'arbres de la forêt, mais pour chaque observation ce n'est pas le même ensemble d'arbres qui est agrégé.

4.2 Méthodes d'extraction des informations des modèles

Après avoir construit un modèle, il est nécessaire de pouvoir en extraire des informations. Les modèles linéaires sont qualifiés de modèles "interprétables" puisque les coefficients associés aux variables de ces modèles sont comparables. Les forêts aléatoires appartiennent en revanche à la catégorie de modèles dits "boîtes noires". Différentes solutions existent néanmoins afin de permettre leur explicabilité.

4.2.1 L'importance des variables

Disposer d'une mesure de l'importance des variables dans les modèles permet de construire une hiérarchie des variables explicatives et participe ainsi à la compréhension de l'impact des variables explicatives sur la variable à expliquer. Certains modèles disposent d'indices intégrés qui permettent de mesurer les contributions des variables, comme les modèles linéaires ou les forêts aléatoires.

L'importance des variables dans le cadre des modèles linéaire

Dans le cas de la régression linéaire, le coefficient de régression β_j associé à une variable explicative X_j , $j \in \{1, \dots, p\}$, indique la variation de la variable à expliquer Y lorsque X_j augmente d'une unité. L'importance de la variable X_j est définie par la valeur absolue de la t-statistique :

$$t_{\beta_j} = \left| \frac{\beta_j}{\sigma_{\beta_j}} \right|$$

Plus cette valeur est élevée, plus la variable est importante et inversement.

L'importance des variables dans le cadre des forêts aléatoires

L'indice d'importance des variables dans le cadre des forêts aléatoires a été introduit par Breiman. Il est présenté ici tel qu'il est implémenté dans le package *randomForest* de R. Il s'agit d'un indice d'importance par permutation. L'idée est qu'une variable explicative peut être considérée comme importante si briser le lien entre cette variable et la variable à expliquer Y dégrade l'erreur de prédiction du modèle. L'importance de chaque variable est mesurée par la diminution de la précision des prévisions lorsqu'une variable est permutée.

L'importance par permutation pour une variable X_j se calcule de la façon suivante. Soient B_l un échantillon *bootstrap*, OOB_l l'échantillon *Out-of-bag* associé et Err_{OOB_l} l'erreur OOB de l'arbre construit sur B_l . Un échantillon perturbé \widetilde{OOB}_l^j est créé en permutant aléatoirement les valeurs de la variable X_j dans l'échantillon OOB_l . L'erreur de l'arbre est calculée en utilisant l'échantillon permué \widetilde{OOB}_l^j . Elle est notée $Err_{\widetilde{OOB}_l^j}$.

Ces étapes sont répétées sur tous les échantillons *bootstrap* de la forêt, i.e. sur tous les arbres. L'indice d'importance est alors égal à la moyenne sur les B arbres de l'augmentation de l'erreur :

$$VI(X^j) = \frac{1}{B} \sum_{l=1}^B (Err_{\widetilde{OOB}_l^j} - Err_{OOB_l})$$

Si la permutation aléatoire de la variable X_j engendre une forte augmentation de l'erreur alors $VI(X^j)$ est élevé. Cela signifie que le modèle est sensible aux variations de la variable et donc qu'elle joue un rôle important dans le modèle, et inversement.

Une version normalisée de cet indice d'importance existe, le Z-score, où $VI(X^j)$ est divisé par son écart type. Pour que l'évaluation de ces indices d'importances soit stable, les forêts doivent être composées de beaucoup d'arbres et plusieurs répétitions des forêts sont utiles pour vérifier la variabilité. Il convient de noter que l'indice d'importance des variables peut être biaisé en présence de variables explicatives corrélées. Ce biais s'explique notamment par le fait que la redondance des informations dilue les scores d'importance. Les permutations peuvent également fournir des instances irréelles, ce qui introduit un biais dans l'importance estimée des variables.

4.2.2 Le graphique des effets locaux accumulés (ALE)

Il existe plusieurs méthodes visuelles d'interprétabilité des modèles de Machine Learning. Les graphiques des effets locaux accumulés (ALE) font partie de ces méthodes et ont l'avantage de ne pas être biaisés par la corrélation entre les variables explicatives.

Un graphique des effets locaux accumulés [8], généralement noté ALE (*Accumulated Local Effects Plot*), décrit comment une variable influence en moyenne la prédiction d'un modèle. Elle repose sur la distribution conditionnelle d'une variable. Le calcul de l'ALE pour une variable X se déroule de la manière suivante. La variable X est divisée en intervalles. Les quantiles de la distribution de X sont souvent utilisés pour définir les intervalles. Pour chaque instance d'un intervalle, la différence en prédiction est calculée lorsque la valeur de X est remplacée par la borne supérieure et par la borne inférieure de l'intervalle considéré. Pour chaque intervalle, les différences obtenues pour toutes les instances sont ensuite additionnées et la somme est divisée par le nombre d'observations de l'intervalle pour obtenir la différence moyenne en prédiction. Les différences moyennes de tous les intervalles sont ensuite accumulées et centrées afin d'obtenir la courbe d'ALE.

L'étape de centrage permet d'interpréter l'ALE comme l'effet d'une variable X sur la prédiction par rapport à la prédiction moyenne sur l'ensemble des données. Par exemple, si la valeur de l'ALE est égale à -2 lorsque la variable X vaut 3, alors la prédiction est inférieure de 2 à la prédiction moyenne lorsque X vaut 3.

4.2.3 L'approche SHAP

La méthode SHAP (*SHapley Additive exPlanations*) est une méthode d'interprétabilité permettant d'expliquer les prédictions individuelles faites par un modèle de Machine Learning. Elle repose sur un fondement mathématique solide et est la seule méthode actuelle qui donne une explication complète des prédictions selon toutes les variables, ce qui la rend compatible avec le droit à l'explication du RGPD.

L'objectif de SHAP est de quantifier le rôle de chaque variable dans la décision finale d'un modèle. Cette approche repose sur l'utilisation de la valeur de Shapley, provenant de la théorie des jeux et fournissant une réponse à la question fondamentale suivante : "Dans une coalition composée de plusieurs joueurs ayant des compétences différentes, qui se traduit par un gain collectif, quelle est la manière la plus équitable de répartir ce gain entre les joueurs ?". Dans le contexte d'interprétabilité de modèles, les valeurs de Shapley sont calculées pour chacune des observations de l'ensemble de données. SHAP attribue une valeur de contribution à chaque variable utilisée. Pour cela, la différence entre la prédiction obtenue et la prédiction moyenne est distribuée entre les différentes variables utilisées par le modèle. Cette approche explique ainsi la sortie d'un modèle par la somme des effets ϕ_j de chaque variable X_j .

Le principe SHAP est illustré à l'aide d'un exemple (cf. [8]). Soit Y une variable à expliquer représentant le prix d'une voiture en euros. Soient x_1 et x_2 deux variables explicatives correspondant respectivement au nombre de chevaux de la voiture et au nombre de portes. Il est supposé que la prédiction moyenne sur l'ensemble de données complet est égale à 170 000€ et que pour $x_1 = 150$ et $x_2 = 4$ le prix estimé par le modèle est $y = 150 000$. L'objectif est alors d'expliquer la différence de -20 000€ entre la prédiction faite par le modèle et la prédiction moyenne. Le résultat obtenu peut par exemple être : x_1 a contribué de +10 000€ et x_2 de -30 000€ par rapport à la valeur moyenne prédite. La valeur de Shapley correspond donc à la contribution marginale moyenne d'une variable explicative sur toutes les coalitions possibles.

Dans le cas de la régression linéaire, la valeur de Shapley s'exprime simplement. Soit le modèle linéaire $\hat{y} = \beta_0 + \sum_{j=1}^p \beta_j x_j = \hat{f}(x)$. La valeur de Shapley de la variable X_j , $j \in \{1, \dots, p\}$, associée à la prédiction \hat{y} est :

$$\phi_j(\hat{f}) = \beta_j x_j - E[\beta_j X_j],$$

où $E[\beta_j X_j]$ est l'effet moyen de la variable X_j . La valeur de Shapley peut être interprétée comme la contribution de x_j dans la prédiction de \hat{y} , car il s'agit de la différence entre l'effet de la variable et l'effet moyen. La somme des contributions de toutes les variables explicatives est égale à la différence entre la valeur prédite et la valeur de prédiction moyenne.

$$\sum_{j=1}^p \phi_j(\hat{f}) = \sum_{j=1}^p (\beta_j x_j - E[\beta_j X_j]) = (\beta_0 + \sum_{j=1}^p \beta_j x_j) - (\beta_0 + \sum_{j=1}^p E[\beta_j X_j]) = \hat{y} - E[\hat{Y}]$$

Cette méthode se généralise pour tout modèle. L'expression générale de la valeur de Shapley est :

$$\phi_i(f) = \sum_{S \subseteq N \setminus i} \frac{|S|! (M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

Où N est l'ensemble de toutes les variables explicatives, M est le nombre de variables, S est un ensemble de variables, i est la i^{eme} variable, f est le modèle et $f_x(S) = E[f(x)|x_S]$.

En pratique, le temps nécessaire à l'exécution du calcul de la valeur de Shapley est très important. Une approximation basée sur le principe de Monte-Carlo est alors souvent utilisée. L'approche SHAP permet également de mesurer l'importance globale des variables en moyennant les valeurs absolues des valeurs de Shapley pour chaque variable.

4.3 Méthodes de sélection de variables

Machine Learning et Big Data sont deux termes souvent associés. Bien que les algorithmes de Machine Learning peuvent gérer de nombreuses variables, le modèle avec le plus grand nombre de variables explicatives n'est pas forcément le meilleur modèle et toutes les variables ne sont pas forcément pertinentes. Un tel modèle est notamment difficile à interpréter. L'objectif de la sélection de variables est de simplifier le problème en supprimant les variables inutiles et en sélectionnant uniquement celles qui sont pertinentes.

4.3.1 La sélection de variables basée sur la régression linéaire

La sélection de variables permet de trouver le meilleur sous-modèle du modèle complet qui contient seulement les variables pertinentes. Pour comparer des modèles et être en mesure de sélectionner le meilleur, différents critères peuvent être utilisés dans le cadre de la régression linéaire, comme le R^2_{aj} , le BIC (Critère d'Information Bayésien) ou l'AIC (Critère d'Information d'Akaike). Deux types de méthodes sont souvent utilisés pour réaliser la sélection : la méthode exhaustive et les méthodes pas-à-pas.

La méthode exhaustive consiste à construire tous les modèles possibles à partir des différentes variables, puis à les comparer à l'aide d'un critère. Cette technique n'est cependant pas envisageable en présence de nombreuses variables.

Les méthodes pas-à-pas permettent de trouver un "bon" modèle avec un temps de calcul raisonnable. Il existe trois algorithmes classiques : l'élimination en arrière (ou méthode *backward*), la sélection en avant (ou méthode *forward*) et la sélection pas à pas (ou méthode *stepwise*). L'élimination en arrière part du modèle complet et retire les variables les moins significatives les unes après les autres. La sélection en avant part du modèle réduit contenant uniquement l'intercept et ajoute les variables les plus significatives successivement. La sélection pas à pas utilise l'algorithme *forward* en intercalant à chaque itération une étape *backward* afin de retirer les variables non significatives. Ces algorithmes s'arrêtent lorsque l'ajout ou la suppression de variables n'améliore plus le modèle.

4.3.2 La sélection de variables basée sur les forêts aléatoires

Il existe de nombreux algorithmes de sélection de variables dans les forêts aléatoires. Trois méthodes sont présentées ici. Leur fonctionnement est détaillé en annexe B.1.

- L'algorithme RFE (*Recursive Feature Elimination*) avec les forêts aléatoires :

La méthode RFE est probablement la méthode la plus utilisée pour la sélection de variables. Il s'agit d'une méthode de type *backward* qui élimine itérativement les variables les moins importantes et recalcule l'importance des variables dans le modèle à chaque itération.

Il convient de noter que cet algorithme ne retient pas toutes les variables liées à la variable à expliquer lorsque celles-ci sont corrélées. Cet algorithme possède cependant l'avantage de fournir un classement des variables corrigé des effets de la corrélation. L'importance est en effet recalculée à chaque étape, ce qui permet aux variables corrélées non supprimées de se positionner à leur vraie place dans le classement.

- La méthode de Boruta :

La méthode de Boruta est une méthode de sélection de variables souvent citée dans la littérature. Elle permet de sélectionner toutes les variables liées à la variable à expliquer en jugeant de la pertinence des variables. Elle utilise pour cela des copies randomisées de toutes les variables explicatives et repose sur l'idée qu'une variable n'est utile que si elle est capable de faire mieux que la meilleure variable randomisée. Cette méthode est utilisée lorsque l'objectif recherché est la compréhension des mécanismes liés à la variable d'intérêt, plutôt que la seule construction d'un modèle prédictif de type "boîte noire" avec une bonne précision de prédiction.

- La méthode VSURF :

La méthode VSURF (*Variable Selection Using Random Forests*) fonctionne en plusieurs étapes et permet de créer deux sous-ensembles de variables, répondant à deux objectifs différents. Le premier sous-ensemble est construit dans un but d'interprétation. Les variables sélectionnées sont celles qui sont fortement reliées à la variable à expliquer, même si elles sont corrélées entre elles. Le second sous-ensemble répond à un objectif de prédiction. Le plus petit sous-ensemble de variables suffisant pour bien prédire la variable à expliquer est recherché. Cet ensemble contient peu de variables avec très peu de corrélation entre elles.

Le principal inconvénient de cette méthode est qu'elle ne permet pas toujours de détecter toutes les variables liées à la variable à expliquer dans le cas de nombreux prédicteurs corrélés. La stratégie proposée par Boruta est plus précise que VSURF pour récupérer les faibles corrélations redondantes entre les prédicteurs et la variable à expliquer. Le risque avec la méthode de Boruta est de sélectionner des "faux positifs". La méthode VSURF est plus parcimonieuse que la méthode de Boruta, dans le sens où elle sélectionne moins de variables.

4.4 Étude de l'inflation associée aux médicaments PH65

La première famille de remboursements étudiée est la famille "Médicaments PH65". Les modèles présentés précédemment sont utilisés afin d'établir une relation entre l'inflation et les différentes variables explicatives. L'objectif est d'identifier des variables qui influencent l'inflation et de déterminer de leur participation à la variabilité de l'inflation existant sur cette famille. Une étude préliminaire des associations entre les variables est réalisée afin de comprendre les données.

4.4.1 Les associations entre les variables

Les relations entre les différentes variables sont étudiées à l'aide de tests statistiques et de coefficients de corrélation. Afin de ne pas se fier uniquement à ces résultats, des représentations graphiques sont également utilisées.

Dépendance entre les variables explicatives et l'inflation

L'existence d'une relation entre l'inflation et chacune des variables explicatives est testée sous une forme univariée. Des analyses préliminaires révèlent l'absence de distribution gaussienne pour la plupart des variables quantitatives (diagrammes quantile-quantile et tests de Kolgomorov-Smirnov), en particulier pour l'inflation. Il est alors décidé de procéder à la réalisation de tests non paramétriques.

- Lien entre les variables qualitatives et l'inflation :

Le test de Kruskal-Wallis est utilisé pour déterminer si les variables explicatives qualitatives sont liées à l'inflation. Il s'agit d'un test non paramétrique qui utilise les rangs des données de k échantillons indépendants afin de déterminer si ces échantillons proviennent de la même population ou de populations ayant des caractéristiques identiques. Les hypothèses testées sont :

- H_0 : les échantillons proviennent de populations avec des médianes égales, contre
- H_1 : au moins une des médianes est différente des autres.

Ce test donne les résultats suivants :

Variable	Statistique du test	P-value du test
Âge	312,29	< 0,01%
Sexe	13,44	4%
Région	4,20	34%

TABLE 4.1 – Résultats du test de Kruskal-Wallis

Seule l'hypothèse nulle du test associé à l'âge est rejetée au seuil de significativité $\alpha = 1\%$.

Pour le sexe et la région, l'hypothèse nulle ne peut pas être rejetée au seuil de 1%. Elle pourrait toutefois être rejetée pour le sexe en considérant un seuil de 5%. Bien que ces trois variables aient été utilisées pour la construction de l'inflation étudiée, il semble que cette dernière diffère peu en fonction de la région de résidence des bénéficiaires de soins.

L'association entre l'âge et l'inflation est quant à elle largement significative, ce qui signifie qu'au moins un échantillon associé à une tranche d'âges est différent d'un autre. Afin d'identifier quelles sont les catégories qui diffèrent, le test de Dunn est utilisé pour réaliser des comparaisons par paires. Il montre que de nombreuses différences entre les tranches d'âges sont significatives (cf. Annexe B.2.1). L'inflation diffère donc significativement en fonction de l'âge.

- Lien entre les variables quantitatives et l'inflation :

En observant graphiquement les relations entre l'inflation et les variables quantitatives, il apparaît que la plupart d'entre elles ne sont pas linéaires. Le test basé sur les coefficients de corrélation de Spearman est alors utilisé. Il s'agit d'un test non paramétrique ayant pour objet de tester l'existence d'une relation monotone (croissante ou décroissante) non nécessairement linéaire entre deux variables. Les hypothèses testées sont :

- H_0 : les variables sont indépendantes, contre
- H_1 : les variables sont liées de façon monotone.

La base de données contient 40 variables quantitatives. L'hypothèse nulle est rejetée pour 21 variables au seuil de significativité de 1% (p-value << 1%), ce qui signifie qu'il existe une relation monotone significative entre ces variables et l'inflation. Les variables les plus corrélées avec l'inflation au sens de la corrélation de Spearman sont des proportions de professionnels de santé prescripteurs. En particulier, parmi les 9 variables relatives aux proportions de professionnels de santé prescripteurs, 7 variables apparaissent significativement liées à l'inflation. Les résultats associés aux tests de ces variables sont présentés en annexe B.2.2.

Il convient de mentionner que la corrélation de Spearman examine l'existence d'une relation monotone entre le rang des observations mais n'est pas opérante lorsque la liaison est non monotone. Des nuages de points doivent être utilisés en complément. En particulier, l'observation graphique des relations entre les variables explicatives et l'inflation montre que les relations semblent non monotones pour les proportions de prescripteurs psychiatres/neurologues, chirurgiens/anesthésistes et dentistes. Les résultats des tests associés à ces variables sont en accord avec cette observation puisque l'hypothèse d'absence de relation monotone n'est pas rejetée pour les psychiatres/neurologues et les chirurgiens/anesthésistes. Il ne faut cependant pas conclure à l'absence de relation entre ces variables et l'inflation en observant les résultats de ce test.

Corrélation entre les variables explicatives

Dans un modèle, lorsque plusieurs variables explicatives sont corrélées, leur influence estimée a toutes les chances d'être diluée entre ces différentes variables. La corrélation peut également perturber les estimations des paramètres de certains modèles. Elle est alors étudiée dans le but d'identifier les groupes de variables interdépendantes afin d'en tenir compte dans les interprétations futures.

Étant donné la nature des variables associées aux professionnels de santé prescripteurs, il semble évident que certaines variables sont corrélées. L'augmentation de la proportion d'un certain type de prescripteur est associée à la diminution d'un ou plusieurs autres types. La matrice des corrélations entre ces variables est alors observée.

Il apparaît principalement que les proportions de prescripteurs dentistes, chirurgiens/anesthésistes et psychiatres/neurologues sont fortement corrélées. Cette observation est notamment confirmée en constatant que les tranches d'âges pour lesquelles ces proportions sont élevées sont similaires. Il convient de noter que la proportion de prescripteurs pédiatriques, associée à la tranche d'âges 0-19 ans, n'est pas intégrée dans cette matrice étant donné sa nature particulière.

Bien que certaines modalités de la variable relative au professionnel de santé prescripteur n'aient pas été utilisées pour construire des variables, les données présentent une multicolinéarité élevée. En effet, les modalités exclues de l'étude représentent de faibles proportions, en particulier pour certaines tranches d'âges. Il est alors décidé de ne pas conserver la proportion de prescripteurs médecins généralistes dans les données, étant donné que cette proportion concerne toutes les tranches d'âges et peut être déduite de la somme des proportions des autres prescripteurs.

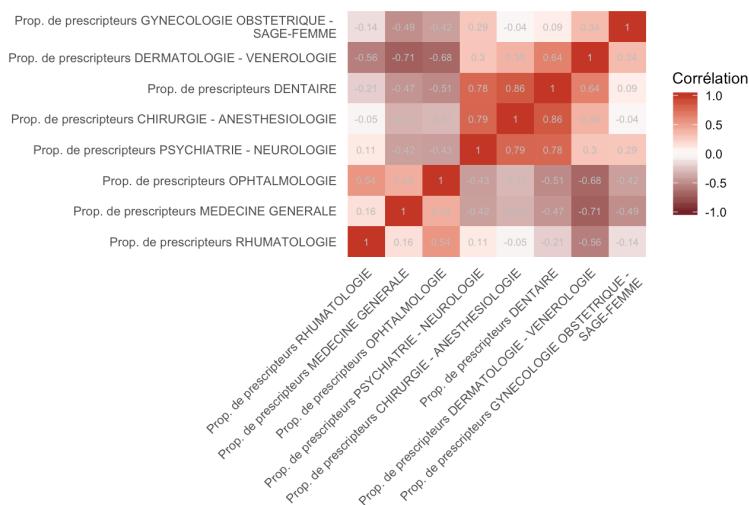


FIGURE 4.3 – Matrice des corrélations entre les proportions de professionnels de santé prescripteurs

Les corrélations entre les autres variables ne sont pas présentées ici étant donné que les proportions de professionnels de santé prescripteurs sont les seules variables quantitatives fortement corrélées entre elles.

Le rapport de corrélation est ensuite utilisé pour évaluer la corrélation entre les variables qualitatives et les variables numériques. Il s'agit d'un indicateur permettant de mesurer le pourcentage de variabilité d'une variable quantitative Y dû aux différences entre les classes d'une variable qualitative X. Il se calcule de la manière suivante :

$$\eta_{Y/X}^2 = \frac{\text{Variation interclasse}}{\text{Variation totale}}$$

Cet indicateur prend des valeurs allant de 0 à 1. Lorsqu'il vaut 0, les moyennes par classes sont toutes égales et les variables ne sont pas liées. Lorsqu'il vaut 1, les moyennes par classes sont très différentes.

Les rapports de corrélation sont affichés pour les variables associées aux proportions de professionnels de santé prescripteurs et pour quelques variables d'intérêt dans la suite de l'étude.

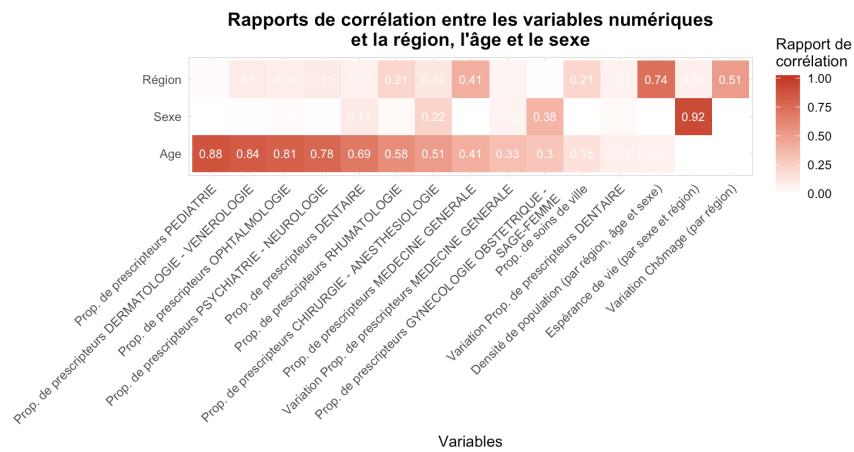
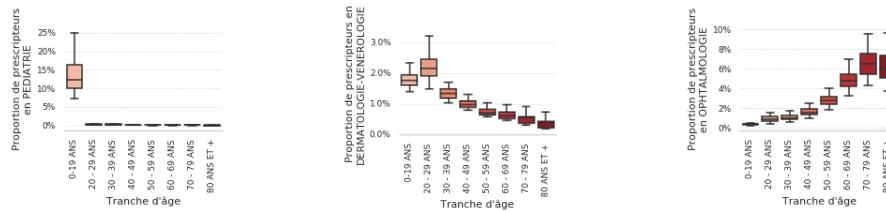


FIGURE 4.4 – Corrélation entre les variables qualitatives et les variables numériques d'intérêt

Ces indicateurs permettent de remarquer que certaines proportions de prescripteurs sont fortement liées à l'âge. Des diagrammes en boîte sont utilisés afin de visualiser le lien entre ces variables.



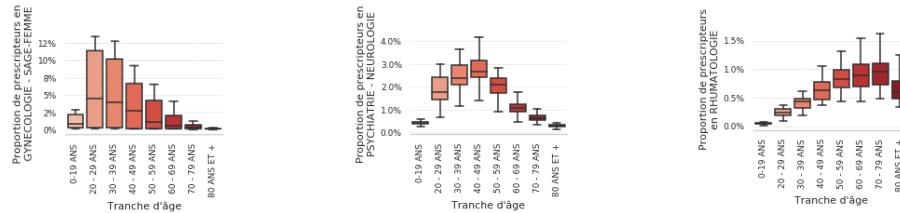


FIGURE 4.5 – Proportions de professionnels de santé prescripteurs en fonction de l'âge

Ces graphiques confirment que l'âge est fortement lié à certaines proportions de prescripteurs. Ces dernières semblent caractériser les tranches d'âge. Certains professionnels de santé sont consultés différemment en fonction de l'âge du bénéficiaire, ce qui explique que les médicaments prescrits sont associés à seulement certaines tranches d'âge. Par exemple, les pédiatres sont consultés par des enfants tandis que les rhumatologues sont peu consultés par des enfants. Cette remarque semble d'ailleurs mettre en avant une certaine relation entre les proportions de prescripteurs pédiatres et rhumatologues. D'autres professionnels de santé sont consultés par des patients de tous âges mais les médicaments qu'ils prescrivent concernent certaines affections qui ne touchent pas forcément toutes les tranches d'âge. C'est notamment le cas des ophtalmologues qui prescrivent des médicaments pour les personnes plutôt âgées, comme par exemple des traitements post-opération de la cataracte.

4.4.2 Les forêts aléatoires

L'étude des relations entre les variables a permis de comprendre les caractéristiques principales des données. L'approche retenue consiste ensuite à utiliser une méthode de Machine Learning afin d'établir une relation entre l'inflation et les variables explicatives et d'identifier les variables les plus influentes. La modélisation de l'inflation permet également de renseigner sur la participation de chaque variable à la variabilité de l'inflation existant sur la famille "Médicaments PH65".

La méthode des forêts aléatoires a été retenue principalement en raison de sa capacité à prendre en compte efficacement les relations non linéaires, compte tenu des observations réalisées sur les données. D'autres éléments ont également contribué à son choix, à savoir son caractère automatique puisqu'elle possède peu de paramètres à régler, sa propriété de modèle non paramétrique, ainsi que sa capacité à gérer les interactions entre les variables.

Importance des variables de la granularité retenue

Un premier modèle est ajusté en ne prenant en considération que les variables retenues dans la granularité de calcul de l'inflation. L'objectif est d'observer si l'une de ces variables est plus importante qu'une autre pour l'explication de l'inflation.

Il convient de rappeler que l'inflation a été mesurée pour les périodes d'observation 2016-2017 et 2017-2018 en fonction de l'âge, du sexe et de la région de résidence du bénéficiaire des soins.

L'âge apparaît comme la variable la plus importante dans le modèle (cf. figure 4.6). Cette observation est en accord avec le résultat du test de Kruskal-Wallis, identifiant l'âge comme variable qualitative significativement liée à l'inflation. Ce résultat peut également être observé à l'aide de diagrammes en boîte (cf. figure 4.7) qui montrent bien que l'inflation diffère en fonction de l'âge.

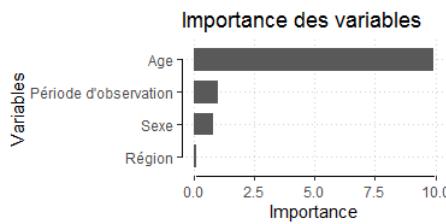


FIGURE 4.6 – Importance des variables de la granularité - Médicaments PH65

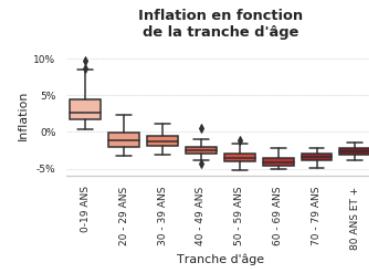


FIGURE 4.7 – Inflation en fonction de l'âge

Ajout des variables de la base d'étude

Un modèle est ensuite ajusté sur l'ensemble des variables de la base d'étude. L'objectif étant d'identifier les variables influentes afin d'envisager de les suivre à l'avenir, la période d'observation n'est pas incluse parmi les variables explicatives.

Afin de vérifier la pertinence du modèle, ce dernier est ajusté sur une base d'apprentissage composée de 75% des données, puis le R^2 ajusté et la RMSE sont calculés sur une base de test contenant les 25% restants.

Les paramètres $mtry$ et $ntree$ sont optimisés indépendamment de la base de test. Un nombre d'arbre élevé est retenu de manière à obtenir une importance des variables stable. Le nombre de variables choisies aléatoirement à chacun des noeuds des arbres ($mtry$) est quand à lui optimisé en minimisant l'erreur *Out-Of-Bag*.

Les valeurs prises par les indicateurs de performance sont présentées dans le tableau ci-dessous. La valeur de la RMSE obtenue après l'optimisation du paramètre $mtry$ est également affichée.

$RMSE_{oob}$	$RMSE_{test}$	$R^2_{aj,test}$
0,96	0,93	73,0%

La performance sur la base de test est jugée satisfaisante pour permettre l'étude de l'importance des variables dans le modèle. Étant donné que ce modèle contient un nombre important de variables explicatives, une sélection des variables est effectuée afin de ne conserver que les variables pertinentes.

Sélection des variables liées à l'inflation

La présence de variables explicatives corrélées ne dégrade pas la performance d'une forêt aléatoire. La corrélation peut cependant influer sur l'interprétation du modèle puisqu'elle peut biaiser l'importance des variables.

Il a été observé que l'âge est fortement lié à différentes proportions de professionnels de santé prescripteurs. En excluant les proportions de professionnels de santé prescripteurs dans le modèle, ce dernier est dégradé puisque sa RMSE augmente de 0,93 à 1,05. Cette dégradation semble indiquer que certaines proportions de professionnels de santé prescripteurs apportent des informations supplémentaires à l'âge pour expliquer l'inflation.

En retirant l'âge du modèle complet, sa performance ne diminue pas. La RMSE reste stable (0,93 sur la base de test). Cette observation semble indiquer que les tranches d'âge sont bien caractérisées par les proportions de professionnels de santé prescripteurs. Il est alors décidé de retirer l'âge du modèle afin de se focaliser sur l'influence des proportions de professionnels de santé prescripteurs, puisque l'influence des tranches d'âge est déjà observée et manifeste. Les interprétations tiendront toutefois compte du lien existant entre ces variables et l'âge.

Dans le but d'identifier les variables pertinentes, les trois méthodes énoncées dans la partie 4.3.2 sont ensuite testées. Les méthodes de Boruta et VSURF sont intéressantes puisqu'elles tentent de sélectionner toutes ou la plupart des variables liées à l'inflation. En particulier pour la méthode VSURF, le sous-ensemble de variables retenu dans un but d'interprétation est considéré. La méthode RFE semble quant à elle adaptée aux données compte tenu du niveau corrélation élevé identifié entre les variables associées aux proportions de professionnels de santé prescripteurs. Cette méthode permet en effet d'assurer qu'au moins une des variables de chaque groupe de variables interdépendantes liées à l'inflation soit sélectionnée.

La méthode de Boruta sélectionne 21 variables, la méthode VSURF en retient 14 et la méthode RFE en conserve 10. Il convient de mentionner que la méthode RFE est incorporée dans une boucle de validation croisée répétée 10 fois.

Une comparaison des ensembles sélectionnés est alors réalisée. Les variables retenues par les méthodes VSURF et RFE sont affichées.

- Les 14 variables sélectionnées par VSURF se trouvent dans la sélection de Boruta. En particulier, les 13 variables les plus importantes sélectionnées par VSURF correspondent aux 13 variables les plus importantes sélectionnées par Boruta.
- Les 10 variables sélectionnées par RFE se trouvent dans l'ensemble sélectionné par VSURF. Les proportions de prescripteurs chirurgiens/anesthésiologistes et dentistes font partie des variables retenues par VSURF qui ne sont pas retenues par RFE. Cette sélection est en adéquation avec les observations issues de la matrice des corrélations puisque ces deux variables sont fortement corrélées avec la proportion de prescripteurs psychiatres/neurologues.

Le classement des variables obtenu avec la méthode RFE est corrigé des effets de la corrélation (cf. figure 4.8). En effet, l'importance des variables est recalculée à chaque itération de l'algorithme. Les variables corrélées non supprimées sont alors positionnées à leur vraie place dans le classement. Ce dernier montre par exemple la présence de corrélation entre la proportion de prescripteurs dermatologues/vénérologues et les autres variables puisque cette variable est moins importante dans le classement obtenu avec l'algorithme RFE.

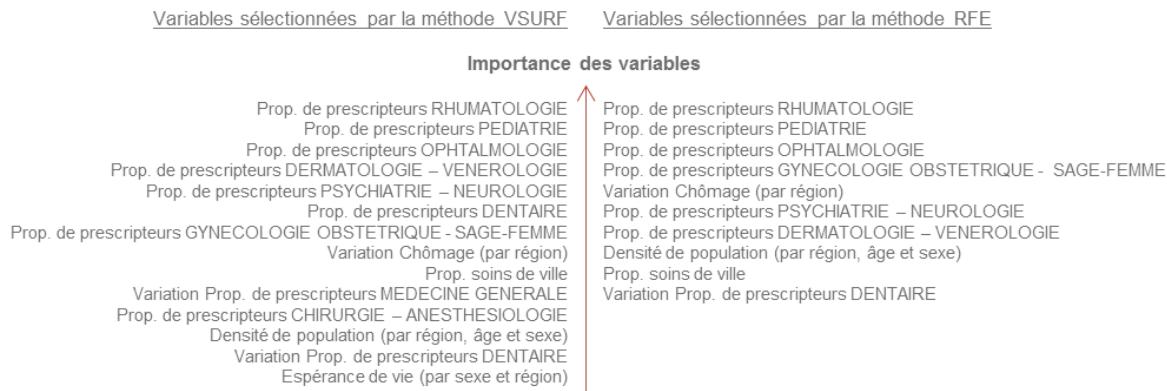


FIGURE 4.8 – Ordre d'importance des variables - méthodes VSURF et RFE

La méthode de Boruta sélectionne toutes les variables liées à l'inflation mais conduit à conserver un nombre élevé de variables, rendant le modèle difficile à interpréter. Les méthodes RFE et VSURF retiennent des ensembles de variables comparables, qui se diffèrent en raison des corrélations. L'observation des effets des variables dans le modèle est ensuite envisagée. Dans cet objectif, les variables sélectionnées par la méthode RFE sont retenues afin de faciliter les interprétations relatives aux variables corrélées. Les interprétations devront être réalisées en gardant à l'esprit l'existence des corrélations mentionnées. Les variables sélectionnées par VSURF pourront être étudiées afin de poursuivre les observations. Il convient toutefois de noter qu'il n'y a pas une absence de corrélation au sein des variables sélectionnées par RFE. Les corrélations sont cependant limitées par rapport au modèle complet et les variables retenues contribuent à l'explication de l'inflation.

La performance du modèle est observée sur le modèle réduit afin de vérifier sa pertinence. Il apparaît alors que la sélection réalisée permet d'améliorer légèrement sa performance.

$RMSE_{oob}$	$RMSE_{test}$	$R^2_{aj,test}$
0,93	0,92	81,9%

L'importance de variables est ensuite observée à l'aide de l'indice d'importance par permutation propre aux forêts aléatoires. Les variables sont énumérées par ordre décroissant d'importance, ce qui signifie que les variables du haut contribuent davantage au modèle que celles du bas.

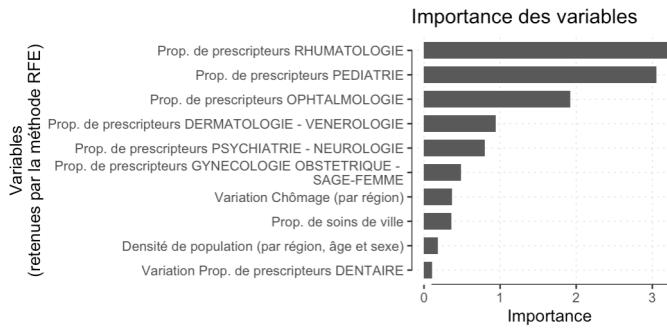


FIGURE 4.9 – Importance des variables du modèle après sélection de variables

L'importance des variables indique principalement que les proportions de certains professionnels de santé prescripteurs contribuent fortement au modèle. L'effet de ces variables sur l'inflation sera observé par la suite. Il peut être intéressant de mentionner que lorsque l'âge est ajouté dans ce modèle, il correspond à la variable la plus importante et l'importance des autres variables est légèrement réduite.

Zoom sur les variables liées aux professionnels de santé prescripteurs

En observant la liste des variables sélectionnées, notamment par VSURF et Boruta qui tentent de conserver toutes les variables liées à l'inflation, il apparaît que ce n'est pas la variation de la proportion de prescripteurs entre deux années qui ressort principalement du modèle mais le niveau de la proportion de prescripteurs.

Ces deux types de variables ont une signification différente par rapport à l'inflation. Si la variation de la proportion de prescripteurs entre deux années a un impact sur l'inflation, cela signifie alors que l'inflation est due à une modification de la composition du panier moyen de médicaments. Si le niveau de la proportion de prescripteurs a un impact sur l'inflation, il s'agit cette fois-ci du poids d'un prescripteur dans le panier moyen qui a un impact sur l'inflation. L'inflation est dans ce cas particulièrement liée à certains types de médicaments.

Afin d'observer ces deux effets, l'inflation est observée en fonction de la proportion de prescripteurs et de la variation de la proportion de prescripteurs à l'aide de nuages de points. Deux exemples de prescripteurs sont présentés ici.

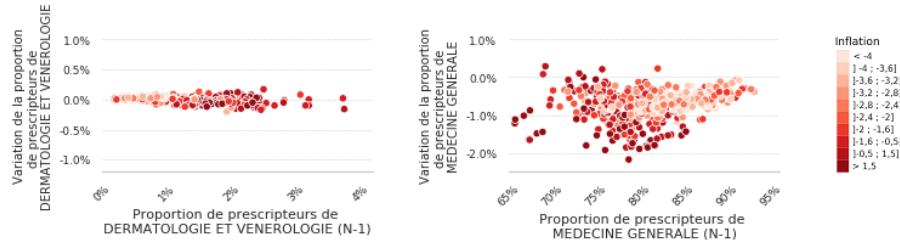


FIGURE 4.10 – Inflation par rapport aux proportions et aux variations des proportions de certains prescripteurs

- Le premier exemple considéré concerne les dermatologues/vénérologues. Leur proportion a peu évolué entre deux années. La variation affichée sur l'axe des ordonnées n'a par conséquent pas d'impact sur l'inflation. Les inflations élevées sont par contre associées à une proportion de prescripteurs dermatologues/vénérologues élevée, et inversement. Le niveau de la proportion de ces professionnels de santé prescripteurs semble donc corrélé à l'inflation.
- Le second exemple présenté concerne les prescripteurs qualifiés en médecine générale. L'inflation s'avère plus élevée lorsque la proportion de ces prescripteurs est plus faible mais aussi lorsque la variation de la proportion de prescripteurs est plus négative. La proportion et la variation de la proportion de prescripteurs qualifiés en médecine générale semblent donc corrélées à l'inflation.

Pour la majorité des professionnels de santé prescripteurs, la proportion a assez peu évolué d'une année à l'autre. L'inflation est alors peu liée à la variation de la composition du panier moyen qui serait engendrée par l'évolution des prescripteurs. L'inflation semble en revanche corrélée aux proportions des différents professionnels de santé prescripteurs, indiquant que l'inflation diffère en fonction du type de médicament consommé. Il s'agit bien du phénomène mis en avant par la forêt aléatoire.

Effet des variables sur l'inflation

Les effets des variables dans le modèle sont étudiés afin de comprendre comment elles contribuent à la variabilité de l'inflation modélisée. Seuls les effets des variables les plus importantes du modèle sont présentés dans ce mémoire.

Le diagramme de synthèse de SHAP est utilisé pour observer conjointement l'importance des variables explicatives et leur effet en fonction de leur valeur. Les variables sont classées par ordre décroissant d'importance, en considérant la moyenne des valeurs absolues des valeurs de Shapley pour chaque variable comme mesure de l'importance.

Les valeurs de Shapley étant calculées pour chaque observation de l'ensemble de données, il est possible de représenter chacune d'elles par un point. Pour rappel, la valeur de Shapley traduit, pour chacune des observations de l'ensemble de données, la contribution de chaque variable à la différence entre la prédiction obtenue et la prédiction moyenne. L'emplacement horizontal de chaque point sur le diagramme de synthèse de SHAP indique si l'effet de cette valeur est supérieur ou inférieur à la prédiction moyenne. Les points foncés correspondent à des valeurs élevées de la variable observée tandis que les points clairs représentent des valeurs faibles. Le graphique obtenu permet ainsi de visualiser facilement l'effet des variables en fonction de leur valeur.

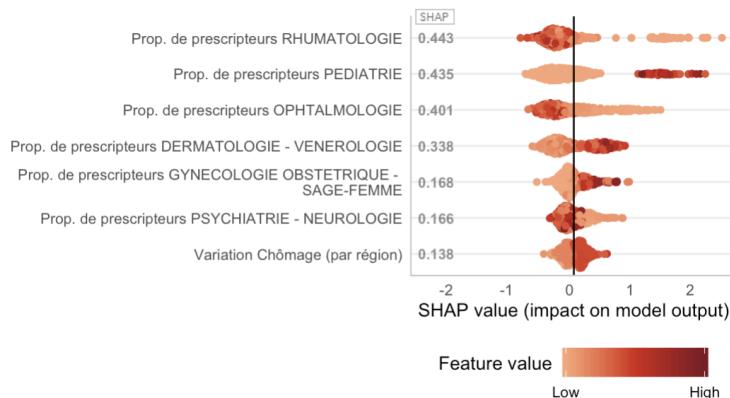


FIGURE 4.11 – Diagramme de synthèse de SHAP

Plusieurs observations peuvent être tirées de ce diagramme de synthèse de SHAP. Une proportion très faible de prescripteurs rhumatologues a un impact positif (inflation en hausse). Au contraire, une proportion élevée de prescripteurs pédiatres a un impact positif. Cela semble cohérent avec l'inflation élevée observée sur la tranche d'âges 0-19 ans (cf. figure 4.7) puisque les proportions élevées de prescripteurs pédiatres sont naturellement associées à la tranche d'âges contenant les enfants. Une proportion élevée de prescripteurs ophtalmologues semble ensuite avoir un impact négatif. Concernant les prescripteurs dermatologues/vénérologues, les proportions élevées ont un impact positif. Pour les prescripteurs gynécologues/sage-femmes ainsi que les prescripteurs psychiatres/neurologues, l'impact semble moins évident en observant ce diagramme.

Enfin, la variation du chômage par région semble avoir une influence positive sur l'inflation en sortie du modèle. Il convient de mentionner que cette variable diffère en fonction de la période d'observation (2016-2017 ou 2017-2018) et permet de caractériser ces deux périodes. Il est alors difficile d'imaginer une interprétation de la corrélation observée entre l'inflation et la variation du taux de chômage sur l'inflation.

Dans la continuité de ces observations, les valeurs de Shapley sont représentées sous forme de nuages de points, permettant de visualiser l'effet des variables sur l'inflation. Ils permettent de valider les observations faites à l'aide du diagramme de SHAP, tout en donnant quelques précisions supplémentaires.

Alors que le diagramme de synthèse SHAP affiché précédemment donne un aperçu général de chaque variable, les graphiques qui suivent représentent les valeurs observées d'une variable en abscisse et les valeurs de Shapley observées pour cette variable en ordonnées. La dispersion verticale résulte des effets d'interaction dans le modèle.

Ici, les graphiques sont colorés en fonction de l'âge dans le but de tenir compte du lien existant entre l'âge et les professionnels de santé prescripteurs lors de l'interprétation. Des graphiques montrant les interactions pourraient être présentés mais ces dernières s'avèrent relativement évidentes en colorant les points en fonction de l'âge.

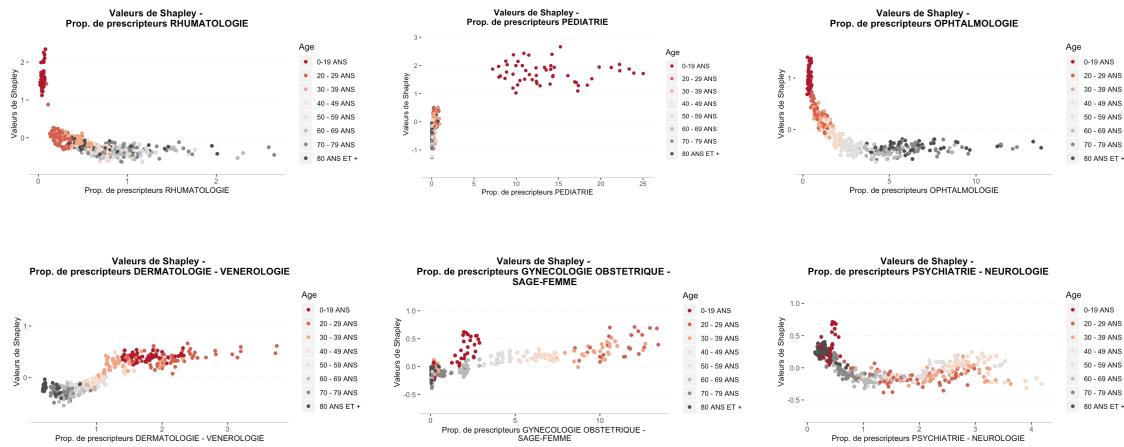


FIGURE 4.12 – Valeurs de Shapley associées aux proportions de prescripteurs

Il apparaît alors clairement que la proportion de prescripteurs pédiatres permet de distinguer l'inflation élevée associée à la tranche d'âges 0-19 ans. Les rhumatologues apparaissent également principalement utiles pour la différenciation de cette tranche d'âges. En plus de caractériser les tranches d'âge, les autres proportions de professionnels de santé prescripteurs semblent apporter quelques informations supplémentaires puisque de légères tendances se dégagent au sein de certaines tranches d'âge. Ces observations permettent ainsi d'affiner la caractérisation des profils d'individus présentant une inflation plus ou moins élevée.

Une limite à l'utilisation des ces méthodes d'interprétation des modèles de Machine Learning est qu'elles peuvent générer des instances irréelles pendant leur calcul en raison de la corrélation entre les variables explicatives [8]. Ces instances peuvent alors entraîner un biais dans l'importance et l'effet estimé des variables.

Des graphiques des effets locaux accumulés (ALE) sont utilisés en comparaison puisque cette méthode tient compte des corrélations et ne crée pas de données irréelles. Il s'agit d'une alternative à la méthode de SHAP montrant les effets moyens des variables.

Les effets observés sont similaires aux observations issues des méthodes précédentes. Il faut toutefois être vigilant en interprétant ces graphiques puisque certaines zones ne contiennent pas ou peu de données. Par exemple, l'intervalle [2% - 7%] de la variable proportion de prescripteurs pédiatriques ne contient pas d'observations.

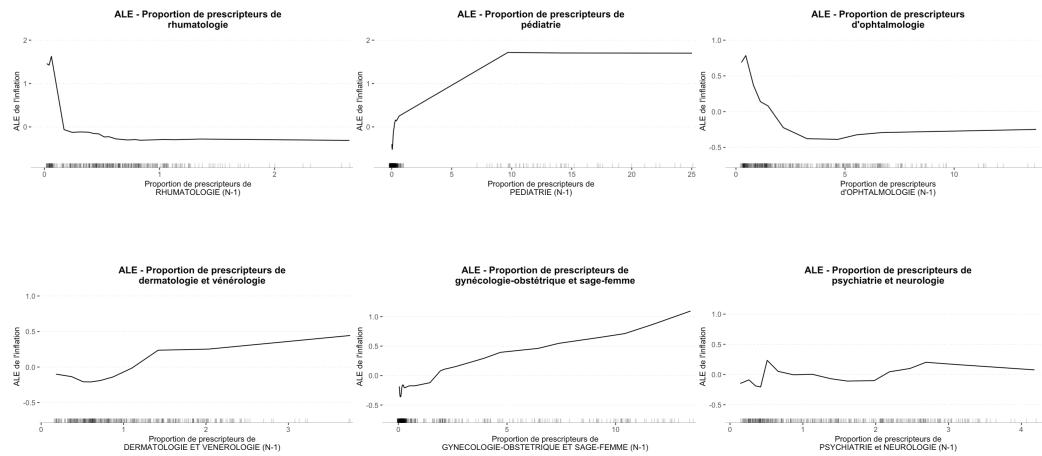


FIGURE 4.13 – Graphiques ALE associés aux proportions de prescripteurs

A la suite de ces observations, un modèle est implémenté sans les données relatives à la tranche d'âges 0-19 ans, puisque les forêts aléatoires sont sensibles aux valeurs extrêmes. Étant donné que cette tranche d'âges représente moins de 10% du portefeuille de Generali, il semble pertinent de vérifier que les résultats obtenus ne sont pas influencés uniquement par cette catégorie d'âge. Le graphique d'importance et les effets des variables les plus importantes sont présentés en annexe B.2. Les variables les plus importantes dans ce modèle correspondent à celles du modèle construit à partir des données complètes, à l'exception des proportions de prescripteurs rhumatologues et pédiatriques servant à différencier la tranche d'âges 0-19 ans. Les effets des variables sont également conformes à ceux issus du modèle contenant la tranche d'âges 0-19 ans.

4.4.3 Le modèle linéaire

Un modèle linéaire est implémenté afin de vérifier que les effets mis en évidence par la méthode des forêts aléatoires sont cohérents avec ceux observés à l'aide d'un modèle plus classique et usuel. Ce type de modèle est très utilisé pour sa simplicité d'interprétation. Il suppose cependant que certaines hypothèses soient vérifiées.

Hypothèses du modèle linéaire

Le modèle linéaire appartient à l'ensemble des modèles linéaires généralisés. Plusieurs familles de distribution peuvent être modélisées dans ce cadre. Ce choix est généralement guidé par la nature de la variable à expliquer. La loi retenue doit cependant appartenir à l'une des familles de distributions exponentielles. Dans le cas de l'inflation, la loi normale semble la plus adaptée, notamment puisque l'inflation prend des valeurs réelles positives et négatives.

Un modèle linéaire gaussien repose sur certaines hypothèses relatives à ses résidus : l'hypothèse de normalité des résidus et l'hypothèse d'homoscédasticité des résidus. Elles sont observées à l'aide d'outils graphiques. En particulier, deux graphiques sont présentés ici. Le diagramme Quantile-Quantile permet d'observer si les résidus sont normalement distribués. Le graphique appelé "Scale-Location" permet de vérifier l'hypothèse d'homoscédasticité. Les graphiques de validation des hypothèses relatives aux résidus affichés ci-dessous sont associés au modèle implémenté par la suite avec des variables discrètes. Ils indiquent que les hypothèses du modèle linéaire gaussien ne sont pas vérifiées, en particulier dans la partie supérieure de la distribution avec certains résidus bien plus élevés que ceux attendus pour une loi normale. Cette observation montre que certaines inflations sont sous-estimées par le modèle.

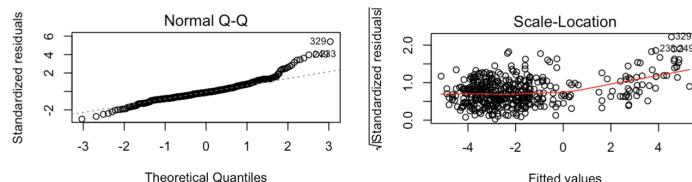


FIGURE 4.14 – Observation des hypothèses du modèle linéaire - avant transformation de l'inflation

D'un point de vue statistique, il est possible d'utiliser la méthode des moindres carrés pour l'estimation des paramètres du modèle même si les résidus ne sont pas gaussiens. Les tests d'hypothèses associés au modèle linéaire gaussien ne sont cependant plus valides et les observations présentant un résidu important risquent de perturber l'estimateur des moindres carrés. L'hypothèse d'homoscédasticité des résidus ne semble pas vérifiée non plus. Afin d'éviter l'utilisation de modèles erronés, une transformation de la variable à expliquer est effectuée.

La transformation logarithmique est très classique dans le domaine des statistiques. Elle permet de régulariser la distribution en écrétant les fortes valeurs et semble alors pertinente dans la situation de l'inflation. L'inflation comporte cependant des valeurs négatives. La transformation est alors réalisée en deux étapes. La première consiste à ramener l'inflation en valeurs positives à l'aide d'une translation. La seconde étape consiste à appliquer la fonction logarithme aux données. Après cette transformation, les hypothèses du modèle linéaire gaussien sont à nouveau observées et apparaissent cette fois-ci acceptables.

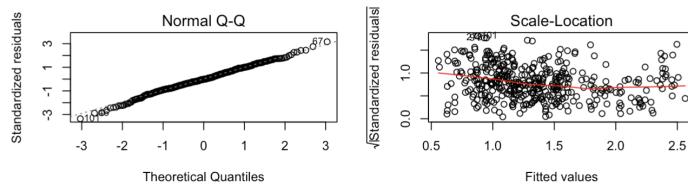


FIGURE 4.15 – Validation des hypothèses du modèle linéaire - après transformation de l'inflation

L'allure de la distribution de l'inflation peut être observée avant et après transformation. Après transformation, la variable à expliquer semble suivre une loi normale. En appliquant le test statistique de Kolmogorov-Smirnov, utilisé pour vérifier l'adéquation à une loi normale, une p-value supérieure à 5% est obtenue ce qui amène à conclure que l'hypothèse selon laquelle la variable à expliquer suit une loi normale ne peut pas être rejetée. Il convient toutefois de rappeler que même si la variable à expliquer suit un modèle linéaire gaussien, une loi normale n'est pas forcément retrouvée en représentant l'histogramme des observations puisqu'il peut s'agir d'un mélange de gaussiennes.

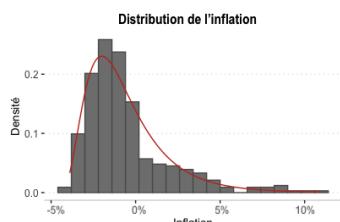


FIGURE 4.16 – Distribution de l'inflation - Médicaments PH65

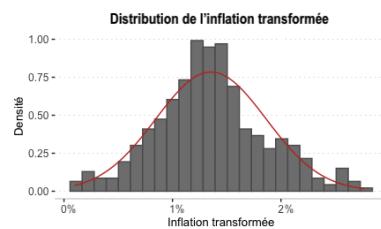


FIGURE 4.17 – Distribution de l'inflation après transformation - Médicaments PH65

Après cette transformation, il est possible de réaliser des analyses de la variance avec les variables de la granularité. Elles permettent de confirmer les précédentes observations et ne sont donc pas présentées ici.

Le modèle linéaire est implémenté avec les variables retenues par la méthode RFE comme variables explicatives, puisqu'elles ont été identifiées comme influentes sur l'inflation.

Performance du modèle

Afin de vérifier la pertinence des modèles, chacun des modèles implémentés par la suite est ajusté sur une base d'apprentissage comme pour la forêt aléatoire et les indicateurs sont calculés sur une base de test. Ces indicateurs sont calculés à partir des inflations non transformées afin de pouvoir être comparés avec les résultats de la forêt aléatoire. La transformation inverse de celle réalisée est alors appliquée sur la prédiction du modèle pour le calcul des indicateurs.

Modèle avec les variables numériques

Un premier modèle est implémenté avec les variables sélectionnées par RFE au format numérique. L'absence de multicolinéarité parmi les variables explicatives est vérifiée à l'aide des VIF (Facteurs d'Inflation de la Variance ou *Variance Inflation Factors*). Parmi les variables qui ne concernent pas les professionnels de santé prescripteurs, celles non significatives selon le test de Student sont supprimées. Les indicateurs de performance prennent alors les valeurs suivantes.

$RMSE_{test}$	$R^2_{aj,test}$
1,23	68,9%

Les coefficients peuvent être observés.

	Coefficients estimés avec variables centrées réduites	Coefficients estimés avec variables non centrées réduites	t value	Pr(> t)
(Intercept)	1,355	1,374	15,116	< 2E-16 ***
Prop. de prescripteurs DERMATOLOGIE - VENEROLOGIE	0,116	0,171	4,821	2,02E-06 ***
Prop. de prescripteurs GYNECOLOGIE OBSTETRIQUE - SAGE-FEMME	0,085	0,023	4,637	4,76E-06 ***
Prop. de prescripteurs OPHTALMOLOGIE	-0,027	-0,010	-0,924	0,356
Prop. de prescripteurs PEDIATRIE	0,214	0,045	9,064	< 2E-16 ***
Prop. de prescripteurs PSYCHIATRIE - NEUROLOGIE	-0,010	-0,010	-0,408	0,684
Prop. de prescripteurs RHUMATOLOGIE	-0,114	-0,267	-5,204	3,10E-07 ***
Variation Chômage (par région)	0,067	0,274	4,166	3,79E-05 ***

FIGURE 4.18 – Coefficients du modèle linéaire avec les variables numériques

Il apparaît alors que l'inflation augmente avec les proportions de prescripteurs pédiatres, dermatologues/vénérologues et gynécologues/sage-femmes, ce qui est cohérent avec le résultat de la forêt aléatoire. Les proportions de prescripteurs ophtalmologues et psychiatres/neurologues sont quant à elles non significatives. Étant donné que les relations entre ces variables et l'inflation sont fortement non-linéaires, une discréétisation est envisagée afin de mieux appréhender ces relations.

Discrétisation des variables numériques

Lors de la mise en place d'un modèle linéaire, les variables numériques ayant une liaison non-linéaire avec la variable à expliquer sont généralement discrétisées. Ce découpage en catégories a pour but d'appréhender au mieux les liaisons non linéaires et non monotones entre les variables numériques.

L'influence de chacune des variables numériques retenues par la méthode RFE sur la variable à expliquer est observée. Lorsque la relation entre deux variables est non-linéaire, des catégories sont créées. Ces regroupements sont effectués en tenant compte de l'inflation mais aussi en tentant de rechercher un nombre convenable de classes pour chaque variable. Un nombre élevé de modalités implique en effet de petits effectifs et une moindre lisibilité tandis qu'un nombre faible de modalités peut faire perdre de l'information.

Pour illustrer cette démarche, le découpage en classes de la variable "Proportion de prescripteurs d'ophtalmologie" est présenté. L'inflation moyenne est observée en fonction de la proportion de ce type de prescripteur. Les classes sont choisies en fonction de seuils identifiés par les droites rouges sur le graphique ci-dessous (seuils strictement inférieurs). La répartition des observations dans les différentes classes est ensuite représentée.

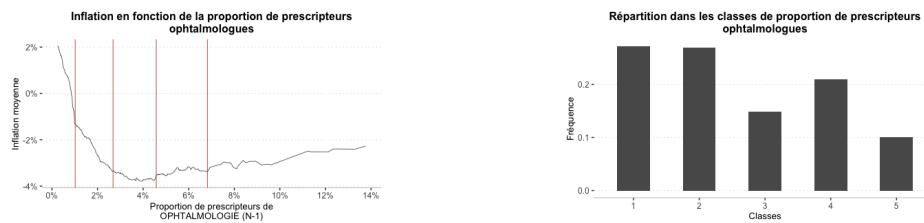


FIGURE 4.19 – Discrétisation de la proportion de prescripteurs ophtalmologues

Le meilleur modèle est obtenu lorsque toutes les variables numériques sont catégorisées. Il convient de mentionner que le danger en utilisant une discrétisation est d'occasionner une perte d'information ou d'introduire une information supplémentaire qui n'existe pas dans les données.

Corrélation entre les variables discrétisées

La corrélation entre les variables explicatives pose des problèmes dans un modèle linéaire puisqu'elle entraîne une augmentation de la variance des estimateurs et l'interprétation devient hasardeuse.

Le V de Cramer est utilisé pour mesurer la corrélation entre les variables. La corrélation avec les tranches d'âges est également observée afin de vérifier que les proportions discrétisées caractérisent toujours l'âge.

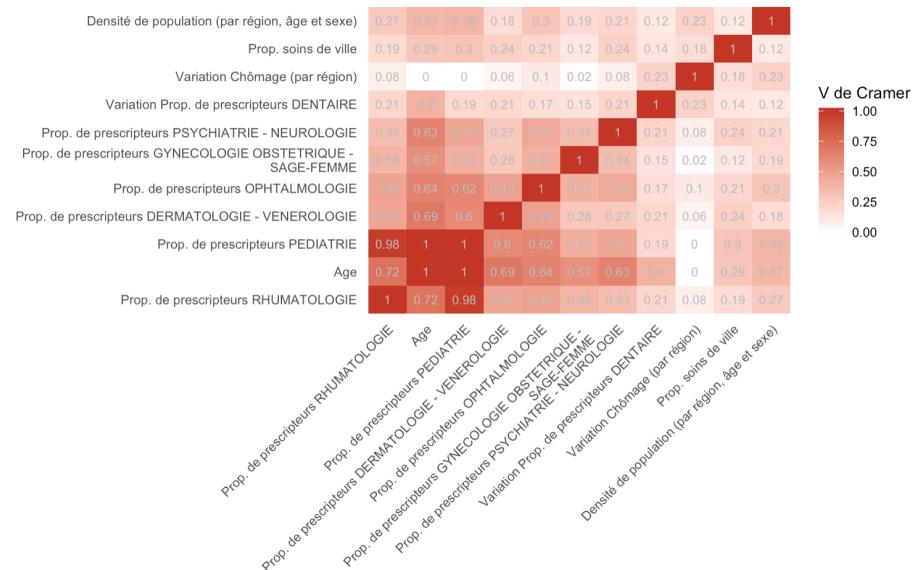


FIGURE 4.20 – Matrice des corrélations

Il apparaît que la proportion de prescripteurs pédiatres est très corrélée à la proportion de rhumatologues. Après avoir été catégorisée, la variable associée aux prescripteurs pédiatres permet seulement de distinguer la classe d'âge 0-19 ans des autres classes d'âge, tandis que la proportion de rhumatologues permet d'écartier cette tranche d'âges, ce qui explique la forte corrélation entre ces variables. La proportion de prescripteurs pédiatres n'est par conséquent pas conservée dans le modèle afin de se rapprocher de l'absence d'information redondante dans les prédicteurs supposée dans un modèle linéaire.

Modèle avec les variables discrètes

Un modèle linéaire est construit à partir des variables discrétisées et sa performance est observée.

$RMSE_{test}$	$R^2_{aj,test}$
1,06	76,3%

Chacune des variables présentes dans le modèle possède au moins une modalité significative. En testant la nullité de tous les coefficients associés à chacune des variables, toutes les variables s'avèrent significatives. Les observations semblent donc cohérentes avec celles de la forêt aléatoire. Les variables identifiées comme influentes sont également significatives dans ce modèle linéaire.

Dans ce modèle, les paramètres peuvent être interprétés en les considérant conjointement aux catégories créées. L'ensemble des coefficients obtenus est disponible en annexe B.2.4.

Ce modèle est cependant de moins bonne qualité que la forêt aléatoire au vu de sa RMSE. De nombreuses améliorations pourraient alors être apportées. En particulier, la discréétisation des variables pourrait être améliorée et des interactions pourraient être ajoutées. Ce n'est cependant pas le but recherché dans ce mémoire.

Pour conclure, un modèle linéaire permet de confirmer les résultats obtenus à l'aide de la forêt aléatoire. Une forêt aléatoire semble cependant plus rapide à mettre en oeuvre puisque les liaisons non linéaires sont automatiquement prises en considération. De plus, l'utilisation d'un modèle paramétrique tel que le modèle linéaire gaussien nécessite la vérification d'hypothèses qui peuvent être difficiles à remplir. Malgré le fait que les forêts aléatoires appartiennent aux modèles de type "boîte noire", de nombreux outils d'interprétation existent, rendant ces modèles interprétables pour l'étude réalisée. La méthode des forêts aléatoire est alors retenue pour l'étude des autres familles de remboursements.

4.4.4 L'apport des modèles pour la compréhension de l'inflation

Les modèles ont mis en évidence que l'inflation mesurée sur la famille des médicaments remboursés à 65% est principalement impactée par l'âge du bénéficiaire des soins. Ils permettent également d'observer que l'inflation diffère en fonction des proportions de professionnels de santé prescripteurs, en particulier des proportions de prescripteurs gynécologues/sage-femmes, ophtalmologues et dermatologues. Ces proportions de professionnels de santé prescripteurs permettent de caractériser les tranches d'âges. Cette observation permet de mettre en évidence que l'inflation est liée à l'âge en raison du type de médicaments consommés qui diffère en fonction de l'âge. Les proportions de professionnels de santé prescripteurs apportent également des explications supplémentaires. Pour un âge donné, l'inflation peut différer en fonction de la proportion de certains prescripteurs et donc de la proportion consommée de certains types de médicaments.

Le modèle mis en place permet alors de donner une indication sur les profils d'assurés associés à des inflations plus ou moins élevées. Les variables importantes permettent de caractériser ces profils afin d'expliquer l'inflation.

L'âge est une variable disponible sur le portefeuille de Generali, qui peut donc être suivie pour expliquer l'évolution de l'inflation. Son influence est observée sur le portefeuille de Generali afin de voir si l'observation de la base Open DAMIR se confirme. Les âges sont regroupés en tranches d'âges et l'inflation est calculée. Il apparaît effectivement que l'inflation sur les médicaments remboursés à 65% par la Sécurité Sociale est significativement différente en fonction de l'âge. L'influence du sexe et de la région sont également appréciées et se révèlent non significatives. Étant donné que l'âge exact des assurés est disponible dans les données de Generali, une étude plus fine pourrait être réalisée afin d'affiner l'identification des profils concernés par une inflation élevée.

L'information relative au professionnel de santé prescripteur n'est en revanche pas disponible sur le portefeuille de Generali. Il pourrait être intéressant de récupérer cette information afin d'étudier si les observations se confirment sur les données de Generali. Le cas échéant, le suivi de ces variables permettrait de caractériser des profils d'assurés concernés par des inflations élevées et d'améliorer la compréhension de l'inflation observée sur le portefeuille.

L'étude réalisée permet d'observer que l'inflation est liée au type de médicaments consommés. Les variables disponibles n'apportent cependant pas d'autre information sur les médicaments pour permettre de comprendre l'origine de l'inflation. L'inflation peut alors résulter de l'évolution du prix de ces médicaments ou peut par exemple provenir d'une variation de la consommation de médicaments génériques. En effet, si plus de génériques ont été consommés dans certaines familles de médicaments, le coût moyen de ces médicaments a par conséquent diminué. Afin de mieux comprendre l'origine de l'inflation, il pourrait être intéressant de disposer de la répartition entre les médicaments génériques et princeps. La possibilité de disposer cette information semble toutefois limitée.

4.5 Étude de l'inflation associée aux consultations et visites des généralistes

La seconde famille de remboursements étudiée est la famille "Généralistes". Comme pour la famille "Pharmacie PH65", l'objectif est d'établir une relation entre l'inflation et les différentes variables explicatives dans le but d'identifier les variables qui peuvent influencer l'inflation observée. Seule la méthode des forêts aléatoire est présentée ici, puisqu'elle permet de prendre en compte efficacement les relations non-linéaires observées pour cette seconde famille. Une étude des associations entre les variables est préalablement réalisée afin de comprendre les données.

4.5.1 Les associations entre les variables

Dépendance entre les variables explicatives et l'inflation

L'existence d'une relation entre l'inflation et chacune des variables explicatives est testée à l'aide des méthodes retenues pour l'étude relative aux médicaments.

- Lien entre les variables qualitatives et l'inflation :

Le test de Kruskal-Wallis est utilisé pour déterminer si les variables explicatives qualitatives sont liées à l'inflation. Ce test donne les résultats suivants.

Variable	Statistique du test	P-value du test
Région	137,78	< 0,01%
Âge	113,14	< 0,01%
Sexe	32,63	< 0,01%

TABLE 4.2 – Résultats du test de Kruskal-Wallis

Les hypothèses nulles des tests associés à ces trois variables sont rejetées au seuil de significativité $\alpha = 1\%$ ($p\text{-value} << 1\%$). Pour chacune de ces variables, la médiane d'au moins un échantillon est différente de celle d'un autre échantillon. Des tests de Dunn sont réalisés pour l'âge et la région afin d'observer si seulement l'une des classes se démarque des autres ou si plusieurs différences significatives existent. Ces tests sont disponibles en Annexe B.3.1. Ils montrent de nombreuses différences significatives entre les tranches d'âges et entre les régions. L'inflation est alors significativement différente en fonction de chacune des variables de la granularité de calcul de l'inflation.

- Lien entre les variables quantitatives et l'inflation :

La base de données relative à la famille "Généralistes" contient 26 variables numériques. Le test basé sur les coefficients de corrélation de Spearman est utilisé pour examiner l'existence d'une relation monotone entre ces dernières et l'inflation. Les résultats associés aux variables pour lesquelles le test est rejeté au seuil de significativité $\alpha = 1\%$ sont affichés en Annexe B.3.2.

Ils ne sont pas présentés dans le corps de ce mémoire puisque la forêt aléatoire implémentée par la suite permettra de mettre en évidence ces relations.

Bien que les tests indiquent l'existence de liaisons monotones pour 10 variables, les niveaux de corrélation observés sont plutôt faibles. Le coefficient de corrélation de Spearman maximal s'élève à 0,49. Ces niveaux faibles s'expliquent par la nature des variables externes, souvent non disponibles à une granularité fine. Les variables externes ne possèdent généralement pas une valeur différente pour chaque combinaison des caractéristiques des bénéficiaires des soins utilisées comme granularité. Une multitude d'inflations est alors associée à une unique valeur de ces variables. Par exemple, la densité d'établissements FINESS par région ne diffère pas en fonction de l'âge et du sexe. Une valeur de cette variable est alors associée à 16 inflations différentes, puisqu'il existe 8 tranches d'âges et deux sexes.

Corrélation entre les variables explicatives

La matrice des corrélations entre les variables quantitatives est observée. Elle est présentée ici pour les variables numériques présentant une relation monotone significative avec l'inflation selon le test basé sur le coefficient de corrélation de Spearman ainsi que pour les variables d'intérêt identifiées dans la suite de l'étude. Compte tenu de la manière dont ont été construites ces variables, l'observation de nuages de points semble toutefois plus pertinente afin d'appréhender les relations.

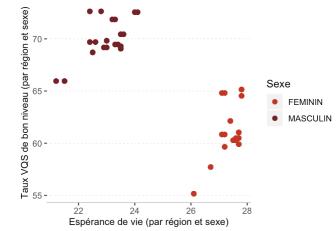


FIGURE 4.21 – Relation entre deux variables numériques caractérisant le sexe

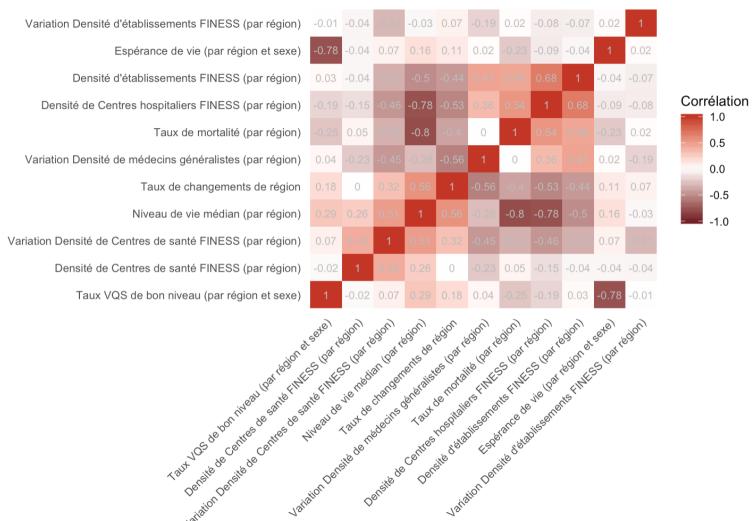


FIGURE 4.22 – Matrice des corrélations entre des variables numériques

Ces corrélations sont à rapprocher des variables qualitatives en fonction desquelles les variables numériques diffèrent. Les rapports de corrélation sont alors observés. Ils mettent notamment en évidence que la densité d'établissements FINESS et le niveau de vie caractérisent la région, tandis que le taux de bon niveau VQS et l'espérance de vie différencient principalement les sexes.

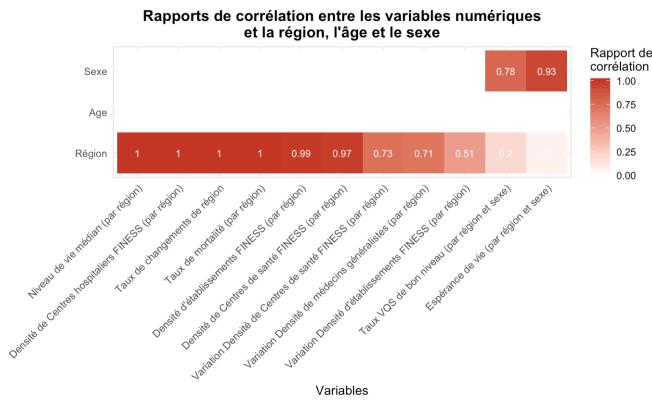


FIGURE 4.23 – Rapports de corrélation entre les variables qualitatives et numériques

4.5.2 Les forêts aléatoires

Après avoir observé les caractéristiques principales des données, la méthode des forêts aléatoires est utilisée pour établir une relation entre l'inflation et les variables explicatives dans le but d'en extraire les variables les plus influentes.

Importance des variables de la granularité retenue

Le premier modèle ajusté ne contient que les variables utilisées comme granularité de calcul de l'inflation. Il permet de mettre en évidence que la région semble être la variable la plus discriminante pour l'explication de l'inflation associée aux consultations et visites des médecins généralistes. Son importance est cependant modérément plus élevée que celle des autres variables, ce qui est cohérent avec les résultats des tests de Kruskal-Wallis indiquant que l'inflation diffère également en fonction de l'âge et du sexe.

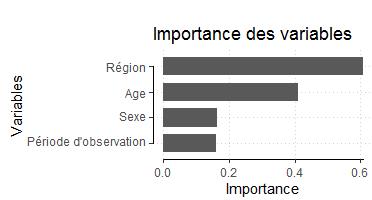


FIGURE 4.24 – Importance des variables de la granularité

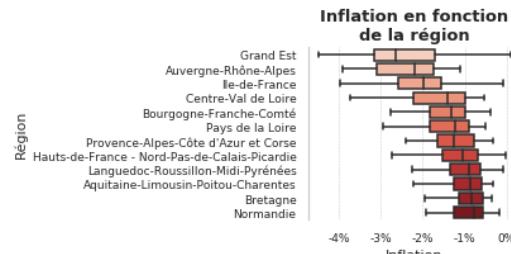


FIGURE 4.25 – Inflation en fonction de la région

Le fait que la région soit importante confirme l'intérêt d'introduire des données externes pour caractériser cette variable et comprendre son importance.

Ajout des variables de la base d'étude

Toutes les variables de la base d'étude sont prises en considération dans le second modèle implanté. Les étapes réalisées pour la familles "Pharmacie PH65" sont répliquées sur cet ensemble de données. La performance du modèle sur la base de test est ensuite mesurée. L'adéquation du modèle aux données semble suffisante pour permettre l'étude de l'importance des variables.

$RMSE_{oob}$	$RMSE_{test}$	$R^2_{aj,test}$
0,31	0,30	78,1%

Sélection des variables liées à l'inflation

L'étude préliminaire des données a montré que la région est bien caractérisée par les variables externes. En la retirant du modèle, ce dernier ne semble pas dégradé au regard de la RMSE qui reste identique. Il est alors décidé de ne pas la conserver dans le modèle afin de concentrer l'attention sur les variables caractérisant les régions.

Les méthodes de sélection de variables VSURF et Boruta sont appliquées à la base d'étude afin d'identifier les variables liées à l'inflation. Ces variables sont ensuite comparées à celles retenues par la méthode RFE permettant de tenir compte des groupes de variables corrélées, en accordant une importance non diluée à au moins l'une des variables de chaque groupe de variables interdépendantes.

La méthode de Boruta n'exclut qu'une unique variable et sélectionne les 27 autres. La méthode VSURF retient 22 variables, tandis que la méthode RFE en conserve 18.

Les variables sélectionnées sont alors comparées. Tout d'abord, toutes les variables présentant une relation monotone significative avec l'inflation selon le test basé sur les coefficients de corrélation de Spearman sont sélectionnées avec VSURF et Boruta. Certaines d'entre elles ne sont pas retenues dans la sélection de RFE. Il s'agit notamment du taux de mortalité et du taux de changements de région. Ces différences semblent s'expliquer par leur corrélation avec d'autres variables explicatives. Ensuite, la méthode RFE n'a sélectionné que des variables comprises dans l'ensemble sélectionné par VSURF. Il s'avère notamment que les 7 variables les plus importantes selon RFE, qui pour rappel recalcule l'ordre d'importance après la suppression de chaque variable, sont identiques aux 7 variables les plus importantes retenues par VSURF.

La performance du modèle réduit à l'aide de la méthode RFE est donnée dans le tableau ci-dessous.

$RMSE_{oob}$	$RMSE_{test}$	$R^2_{aj,test}$
0,31	0,30	80,7%

Une fois les variables liées à l'inflation identifiées, leur importance et leur effet dans le modèle sont étudiés. Seule l'étude des variables les plus importantes est réalisée dans ce mémoire. Étant donné que les 7 variables les plus importantes sont identiques selon VSURF et RFE, il est décidé de les observer. Afin de vérifier que ce nombre de variables étudiées soit pertinent, la RMSE du modèle est considérée en supprimant successivement les variables les moins importantes. Il s'agit de la courbe d'erreur obtenue en sortie de l'algorithme RFE. L'observation des 7 premières variables peut alors sembler pertinente.

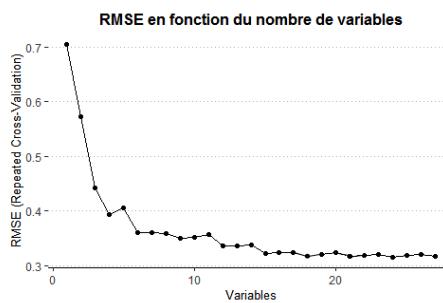


FIGURE 4.26 – Évolution de l'erreur de prédiction en supprimant successivement les variables (algorithme RFE)

Le classement par ordre d'importance obtenu avec l'indice par permutation propre aux forêts aléatoires est alors présenté. L'importance de la région apparaît au travers des variables externes la caractérisant.

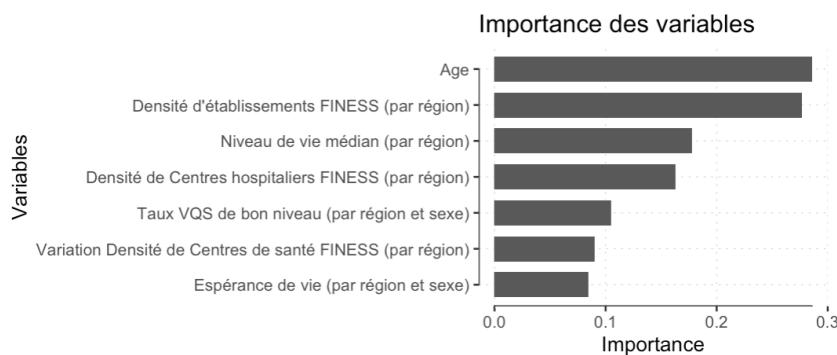


FIGURE 4.27 – Importance des variables

L'âge apparaît à nouveau comme une variable importante. Les variables relatives aux densités d'établissements FINESS par région ressortent ensuite comme des variables importantes dans le modèle. Ces données permettent de décrire l'offre de soins par région par rapport à la population. Le niveau de vie médian par région, indicateur de la capacité financière des assurés, contribue également au modèle.

Deux variables fortement liées au sexe ressortent enfin parmi les variables les plus importantes du modèle. Il s'agit du taux de bon niveau VQS et de l'espérance de vie par région. Ces variables apparaissent plus importantes que le sexe lui-même. Elles caractérisent en effet le sexe mais diffèrent également en fonction de la région. Elles apportent en ce sens une information supplémentaire au sexe dans le modèle.

Effet des variables sur l'inflation

Un diagramme de synthèse de SHAP est ensuite utilisé pour observer l'impact des variables en fonction de leur valeur. Il convient de rappeler que la valeur de Shapley traduit, pour chacune des observations de l'ensemble de données, la contribution de chaque variable à la différence entre la prédiction obtenue et la prédiction moyenne.



FIGURE 4.28 – Diagramme de synthèse de SHAP

D'après les valeurs de Shapley, les tranches d'âges associées aux bénéficiaires de soins les plus jeunes et les plus âgés ont un impact positif (inflation en sortie du modèle en hausse). Une densité d'établissements FINESS par région pour 100 000 habitants faible a un impact négatif. Il en est de même pour la densité de centres hospitaliers, ce qui souligne le fait qu'une offre de soins (ramenée au niveau de population) faible a un impact à la baisse sur l'inflation dans le modèle. Une variation négative de la densité de centres de santé par région semble également avoir un impact négatif. Un niveau de vie médian élevé a quant à lui un impact négatif. Concernant les variables liées au sexe, un taux de bon niveau VQS (i.e. de bon niveau général d'état de santé) élevé a un impact positif. Ce taux est plus élevé pour les hommes que pour les femmes. Le sexe féminin a donc un impact négatif. L'espérance de vie confirme cette observation puisqu'elle est plus élevée pour les femmes et a un impact négatif.

L'âge apparaît comme une variable importante dans le modèle. Son effet peut être observé à l'aide d'un graphique ALE. Les tranches d'âges inférieures à 29 ans ou supérieures à 80 ans sont associées à une inflation plus élevée. Les variables disponibles n'apportent cependant pas d'informations supplémentaires sur ce phénomène.

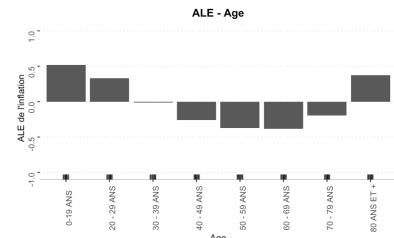


FIGURE 4.29 – Graphique ALE associé à l'âge

Les effets de la densité d'établissements FINESS pour 100 000 habitants et du niveau de vie médian par région sont également observés à l'aide de graphiques ALE. Ces derniers montrent que l'inflation est plus faible avant un certain seuil de densité d'établissements par région. Ils montrent aussi qu'elle est plus faible au delà d'un certain niveau de vie médian.

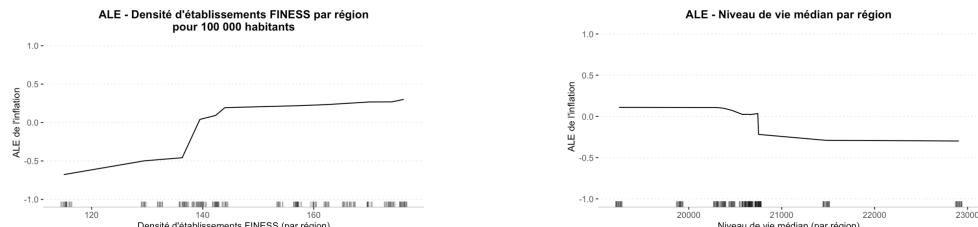


FIGURE 4.30 – Graphiques ALE associés à la densité d'établissements FINESS et au niveau de vie par région

En représentant l'effet conjoint de la densité d'établissements FINESS et du niveau de vie par région dans le modèle, les observations issues des graphiques ALE apparaissent également.

Il convient de noter que ces variables sont corrélées. De nombreuses observations possèdent soit une densité élevée d'établissements FINESS et un niveau de vie médian faible, ou soit un niveau de vie médian élevé et un nombre faible d'établissements FINESS pour 100 000 habitants. Les différents points de données (associés à une région et une période d'observation) sont alors affichés sur ce graphique afin d'identifier les zones dépourvues d'observations.

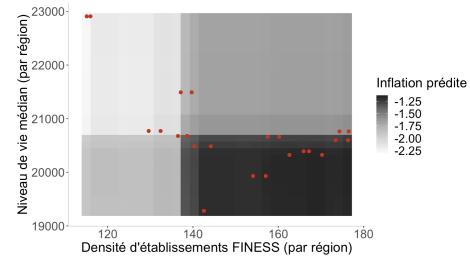


FIGURE 4.31 – Graphique de l'effet conjoint de la densité d'établissements FINESS et du niveau de vie médian par région

4.5.3 L’apport du modèle pour la compréhension de l’inflation

Le modèle met en évidence que l’inflation associée à la famille des consultations et visites des généralistes est liée à l’âge. Il montre également que l’inflation mesurée dépend fortement de la région de résidence des bénéficiaires des soins. Des variables externes sont alors introduites pour enrichir le modèle. Il apparaît que le niveau de vie médian et le nombre d’établissements sanitaires et sociaux (répertoriés dans le FINESS) ramené au niveau de population par région contribuent de manière importante au modèle. Ces variables caractérisent différemment les régions et distinguent d’une autre manière leurs inflations. Contrairement à la variable région, il s’agit de variables numériques. L’observation de leur effet laisse penser qu’elles pourraient de plus permettre d’expliquer plus précisément l’inflation. L’étude est cependant limitée par le fait que la base Open DAMIR ne donne pas accès à une variable de localisation plus précise que la région.

L’influence de la variable région est observée sur le portefeuille de Generali. L’inflation sur les consultations et visites des généralistes est calculée selon la même granularité que sur la base d’étude. Il apparaît alors que l’inflation est significativement différente en fonction de la région.

Sur le portefeuille de santé individuelle, la localisation est plus détaillée. Il pourrait être intéressant d’observer l’inflation à une granularité géographique plus fine afin d’être en mesure de conclure sur une éventuelle influence de la densité d’établissements FINESS et du niveau de vie. Un zonier pourrait être construit à partir des variables identifiées comme importantes récupérées par exemple au niveau des communes. La densité d’établissements FINESS serait notamment évaluée en comptabilisant le nombre d’établissements dans un rayon autour de chaque commune rapporté au niveau de population. L’inflation pourrait alors être mesurée pour chaque zone, en veillant à avoir des zones avec suffisamment d’observations. Cette étude permettrait ainsi de confirmer ou non les observations réalisées sur la base Open DAMIR. Le cas échéant, il serait possible d’identifier des zones géographiques et des profils d’assurés présentant une inflation plus ou moins élevée. Le suivi de ces variables permettrait d’améliorer la compréhension de l’inflation observée sur le portefeuille de Generali et d’envisager la mise en place d’actions de prévention afin de piloter l’inflation.

Conclusion

Une étude de l'inflation est réalisée dans le but d'améliorer le pilotage du portefeuille de santé individuelle de Generali. L'objectif de ce mémoire est d'exploiter les données de marché disponibles afin d'en extraire des enseignements permettant d'améliorer la compréhension de l'inflation de la sinistralité du portefeuille. La base Open DAMIR enrichie à l'aide de données externes est alors utilisée pour étudier l'inflation.

La compréhension de la base Open DAMIR et de sa structure particulière est une étape primordiale afin de manipuler correctement les données. L'analyse de la qualité des données est ensuite d'autant plus importante que des données de mauvaise qualité risquent d'altérer les études réalisées. L'exploitation de ces données l'assurance maladie nécessite plusieurs étapes de préparation afin de se placer du point de vue d'un organisme d'assurance complémentaire santé. Une étude spécifique du portefeuille de Generali est préalablement réalisée afin d'appliquer ses caractéristiques à la base Open DAMIR. Des retraitements des évolutions réglementaires sont effectués afin de mesurer l'inflation réelle. La réalisation de ces étapes demande du temps mais permet d'obtenir une base de données pertinente pour l'étude de l'inflation. L'ensemble de ces travaux permet également de mieux appréhender la base Open DAMIR qui pourra être réutilisée pour de futures études.

L'inflation est mesurée au global sur la base de données construite, ainsi que poste par poste. Une granularité de calcul plus fine, composée de la famille de remboursements à laquelle appartient chaque prestation et des caractéristiques des bénéficiaires des soins (âge, sexe et région de résidence), est ensuite utilisée. Ces différentes inflations mesurées permettent à Generali de positionner son portefeuille par rapport à l'inflation du marché. Sur la granularité fine, des modèles sont appliqués dans le but d'identifier les variables influençant l'inflation. Un modèle global, toutes familles de remboursements confondues, biaise les observations puisque les variables identifiées comme importantes conduisent à une discrimination des différentes familles. Un modèle est alors réalisé pour chaque famille de remboursements dans le but d'expliquer l'inflation. Ce mémoire présente les résultats obtenus sur deux familles de remboursements importantes du portefeuille : les médicaments remboursés à 65% par la Sécurité Sociale et les consultations et visites des médecins généralistes.

La méthode des forêts aléatoires est utilisée pour établir une relation entre l'inflation et les différentes variables explicatives afin d'identifier les variables influentes. Il s'agit d'un modèle non paramétrique qui ne nécessite pas la vérification d'hypothèses, contrairement au modèle linéaire utilisé en comparaison. Cette méthode dispose d'une mesure de l'importance des variables, ce qui a contribué au choix de son utilisation. Elle permet également de gagner en temps de paramétrage par rapport à un modèle linéaire puisque les liaisons non-linéaires sont automatiquement considérées.

L'inflation de la sinistralité peut dépendre d'une multitude de facteurs qui ne sont pas forcément observables. Elle est en ce sens difficile à expliquer. Les données utilisées dans cette étude apportent des informations supplémentaires mais ne permettent pas de comprendre réellement les phénomènes sous-jacents à l'inflation observée. L'étude permet toutefois d'identifier certaines variables explicatives qui pourraient s'avérer pertinentes à suivre sur le portefeuille Generali afin d'expliciter l'inflation.

Concernant le modèle associé à la famille des médicaments remboursés à 65% par la Sécurité Sociale, deux types de variables explicatives ressortent. Il apparaît que l'inflation observée est principalement impactée par l'âge du bénéficiaire des soins. Elle diffère également en fonction des proportions de certains professionnels de santé prescripteurs, notamment des proportions de gynécologues et de sage-femmes, d'ophtalmologues et de dermatologues. Ces observations indiquent que l'inflation diffère en fonction du type de médicaments consommés, caractérisé par l'âge et les proportions de professionnels de santé prescripteurs. Pour les consultations et visites des généralistes, la région contribue principalement au modèle, mais l'âge est également important. L'introduction de variables externes permet d'observer que le nombre d'établissements répertoriés dans le FINESS ramené au niveau de population et le niveau de vie médian caractérisent différemment les régions et permettent de différencier leurs inflations. L'étude doit cependant être réalisée à un niveau géographique plus fin afin de pouvoir conclure sur l'existence d'un éventuel lien entre ces variables et l'inflation.

Dans les deux modèles implémentés, les variables identifiées comme importantes permettent de caractériser des profils d'assurés associés à une inflation plus ou moins élevée. Les observations pourront être étudiées et affinées sur le portefeuille de Generali. Pour les médicaments, l'information relative au professionnel de santé prescripteur est à récupérer afin d'étudier si les observations issues des données de marché se confirment. L'inflation par âge pourra de plus être mesurée en utilisant des tranches moins étendues. Pour les généralistes, la construction d'un zonier permettant l'étude de l'inflation selon une granularité géographique plus fine permettra de conclure à une éventuelle influence de la densité d'établissements sanitaires et sociaux et du niveau de vie. Si les observations issues de la base Open DAMIR se confirment sur le portefeuille de Generali, le suivi de ces variables permettra d'améliorer la compréhension de l'inflation mesurée. En observant les différents indicateurs évoqués dans le temps, certaines actions de prévention pourraient être envisagées afin de réduire l'inflation.

Dans le cadre du pilotage du portefeuille de Generali, les actions de prévention sont actuellement identifiées à partir d'études sur le coût moyen des dépenses de santé. Disposer d'observations relatives à l'inflation permettrait alors d'affiner les actions mises en place.

L'une des limites principales de cette étude est la nature agrégée de la base Open DAMIR. Ces données ont en effet été agrégées préalablement à leur ouverture au public afin que l'identification d'un bénéficiaire ou d'un professionnel de santé soit impossible. Bien que la contrainte de l'anonymisation soit une limite à l'utilisation des données, il s'agit d'une composante incontournable de l'Open Data afin de protéger la vie privée des personnes concernées. Les prestations ne sont donc pas disponibles ligne à ligne, ce qui engendre un biais dans certains retraitements réalisés, en particulier pour les retraitements des évolutions réglementaires. Les études réalisées sont également limitées par le fait que la base Open DAMIR ne donne pas accès à une variable de localisation plus fine que la région et à un âge plus détaillé que les tranches d'âges disponibles. Ces données ne permettent notamment pas de conclure de façon précise sur l'impact de la zone géographique.

Il est également important de préciser que l'étude réalisée ne permet pas d'étudier toutes les garanties couvertes par une complémentaire santé. En particulier, les dépenses d'hospitalisation ne sont pas complètes dans la base Open DAMIR. De plus, l'assurance complémentaire santé peut prendre en charge des soins non remboursés par la Sécurité Sociale, comme par exemple les médecines douces (ostéopathie, acupuncture, etc.).

L'étude réalisée apporte une explication de l'inflation mesurée sur des données de la Sécurité Sociale. Les observations sont à confirmer sur le portefeuille de Generali. Le cas échéant, elles permettront de mieux comprendre l'inflation de la sinistralité et d'améliorer le pilotage du portefeuille. Ces observations sont toutefois valides dans un contexte économique et sanitaire classique. En 2020, la pandémie de Covid-19 et le confinement ont bouleversé le contexte sanitaire. Pendant le confinement, les dépenses de santé ont fortement chuté et un effet rattrapage est observé sur les mois suivants. En particulier, les dépenses de santé ont progressé de 6,9% en juillet 2020 par rapport à juillet 2019 selon le bilan de la Sécurité Sociale. Les dépenses de santé associées aux laboratoires de biologie ont notamment évolué de + 31,1% en raison de la campagne massive de tests de dépistage du coronavirus. Un impact est alors attendu sur l'inflation 2019 - 2020.

Ces observations sont également valides dans un contexte réglementaire stable ou présentant des évolutions facilement identifiables sur les données. A partir du 1er janvier 2020, la réforme 100% santé donne progressivement accès à des paniers de soins dentaires, optiques et auditifs intégralement pris en charge par l'assurance maladie et les organismes d'assurance complémentaire. L'impact de cette évolution réglementaire ne sera cependant pas facile à isoler et l'inflation hors modifications réglementaires sera alors difficilement mesurable.

Table des figures

1.1	Décomposition des remboursements	8
1.2	Les 3 acteurs du remboursement d'un acte médical	20
2.1	Exemple illustratif de la structure de la base de données	25
2.2	Suite de l'exemple illustratif de la structure de la base de données	26
2.3	Résultats des contrôles de qualité des données	29
2.4	Postes et familles de remboursements	32
2.5	Histogrammes annuels des bases de remboursement	35
2.6	Comparaison des montants moyens mensuels réels et <i>as-if</i>	35
2.7	Évolution des proportions d'actes C/CS/V/VS vs G/GS/VG/VGS	37
2.8	Évolution des montants moyens annuels	37
2.9	Comparaison des montants moyens mensuels réels et <i>as-if</i>	38
2.10	Taux de changements de région observés	40
2.11	Agrégation des bases	41
2.12	Coefficients appliqués aux régions, tranches d'âges et sexes des bénéficiaires des soins la base d'étude	45
2.13	Évolution des proportions d'actes par sexe et par région	46
2.14	Évolution de l'âge moyen	46
2.15	Évolution du nombre d'actes	47
2.16	Cadence de règlement	47
2.17	Évolution du coût moyen	48
2.18	Répartition des actes par poste	48
3.1	Décomposition des effets inflation par poste	51
3.2	Inflation 2017-2018 en fonction du recul	52
3.3	Inflation par poste	55
3.4	Inflation en fonction de l'âge et du sexe	55
3.5	Inflation 2016-2017 par région	55
3.6	Inflation 2017-2018 par région	55
3.7	Inflation, calculée avec la granularité retenue, en fonction du nombre d'actes	56
3.8	Dispersion de l'inflation, calculée avec la granularité retenue, par famille de remboursements	56
3.9	Relation entre les espérances de vie à la naissance et à 60 ans	59

3.10 Densité d'établissements FINESS par région en 2017	60
4.1 Découpe de la racine de l'arbre de décision (à gauche) et exemple de partition associée dans l'espace des variables explicatives (à droite)	68
4.2 Principe des forêts aléatoires - Source [3]	69
4.3 Matrice des corrélations entre les proportions de professionnels de santé prescripteurs	78
4.4 Corrélation entre les variables qualitatives et les variables numériques d'intérêt	79
4.5 Proportions de professionnels de santé prescripteurs en fonction de l'âge .	80
4.6 Importance des variables de la granularité - Médicaments PH65	81
4.7 Inflation en fonction de l'âge	81
4.8 Ordre d'importance des variables - méthodes VSURF et RFE	83
4.9 Importance des variables du modèle après sélection de variables	84
4.10 Inflation par rapport aux proportions et aux variations des proportions de certains prescripteurs	85
4.11 Diagramme de synthèse de SHAP	86
4.12 Valeurs de Shapley associées aux proportions de prescripteurs	87
4.13 Graphiques ALE associés aux proportions de prescripteurs	88
4.14 Observation des hypothèses du modèle linéaire - avant transformation de l'inflation	89
4.15 Validation des hypothèses du modèle linéaire - après transformation de l'inflation	90
4.16 Distribution de l'inflation - Médicaments PH65	90
4.17 Distribution de l'inflation après transformation - Médicaments PH65 . .	90
4.18 Coefficients du modèle linéaire avec les variables numériques	91
4.19 Discrétisation de la proportion de prescripteurs ophtalmologues	92
4.20 Matrice des corrélations	93
4.21 Relation entre deux variables numériques caractérisant le sexe	97
4.22 Matrice des corrélations entre des variables numériques	97
4.23 Rapports de corrélation entre les variables qualitatives et numériques . .	98
4.24 Importance des variables de la granularité	98
4.25 Inflation en fonction de la région	98
4.26 Évolution de l'erreur de prédiction en supprimant successivement les variables (algorithme RFE)	100
4.27 Importance des variables	100
4.28 Diagramme de synthèse de SHAP	101
4.29 Graphique ALE associé à l'âge	102
4.30 Graphiques ALE associés à la densité d'établissements FINESS et au niveau de vie par région	102
4.31 Graphique de l'effet conjoint de la densité d'établissements FINESS et du niveau de vie médian par région	102
A.1 Liste des variables de la base Open DAMIR	111

A.2	Tableau récapitulatif des contrôles de qualité des données réalisés	112
A.3	Distribution du nombre de compléments d'acte à supprimer estimé	116
B.1	Illustration de la boucle principale de l'algorithme de Boruta - Source [14]	119
B.2	Test de Dunn : Inflation ~ Âge	121
B.3	Importance des variables de la granularité - Sans les 0-19 ans	123
B.4	Importance des variables - Sans les 0-19 ans	123
B.5	Valeurs de Shapley associées aux proportions de prescripteurs - Sans la tranche d'âges 0-19 ans	123
B.6	Résultat du modèle linéaire	124
B.7	Test de Dunn : Inflation ~ Âge	125
B.8	Test de Dunn : Inflation ~ Région	126

Annexe A

Le traitement de la base Open DAMIR

A.1 Variables de la base Open DAMIR

Variable	Libellé
PERIODE DE TRAITEMENT	
FLX_ANN_MOI	Année et Mois de Traitement
PRESTATION	
PRS_NAT	Nature de Prestation
ASU_NAT	Nature d'Assurance
ATT_NAT	Nature de l'Accident du Travail
CPT_ENV_TYP	Type d'Enveloppe
CPL_COD	Complément d'Acte
EXO_MTF	Motif d'Exonération du Ticket Modérateur
PRS_Rem_tau	Taux de Remboursement
PRS_PPU_SEC	Code Secteur Privé/Public
PRS_FJH_TYP	Type de Prise en Charge Forfait Journalier
ETE_IND_TAA	Indicateur TAA Privé/Public
PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)
DRG_AFF_NAT	Nature du Destinataire de Règlement affiné
PRS_Rem_TYP	Type de Remboursement
ORGANISME	
ORG_CLE_REG	Région de l'Organisme de Liquidation à partir de 2015
PERIODE	
SOI_ANN	Année de Soins
SOI_MOI	Mois de Soins
BENEFICIAIRE	
BEN_SEX_COD	Sexe du Bénéficiaire
AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins
BEN_QLT_COD	Qualité du Bénéficiaire
BEN_RES_REG	Région de Résidence du Bénéficiaire à partir de 2015
MTM_NAT	Modulation du Ticket Modérateur
BEN_CMU_TOP	Top Bénéficiaire CMU-C

EXECUTANT	
PSE_ACT_CAT	Catégorie de l'Exécutant
PSE_SPE_SNDS	Spécialité Médicale PS Exécutant
PSE_ACT_SNDS	Nature d'Activité PS Exécutant
EXE_INS_REG	Région du PS Exécutant à partir de 2015
PSE_STI_SNDS	Statut Juridique PS Exécutant
MFT_COD	Mode de Fixation des Tarifs Etb Exécutant
ETE_REG_COD	Région d'Implantation Etb Exécutant à partir de 2015
ETE_TYP_SNDS	Type Etb Exécutant
ETE_CAT_SNDS	Catégorie Etb Exécutant
DOP_SPE_COD	Discipline de Prestation Etb Exécutant
MDT_TYP_COD	Mode de Traitement Etb Exécutant

PRESCRIPTEUR	
PSP_ACT_CAT	Catégorie du Prescripteur
PSP_SPE_SNDS	Spécialité Médicale PS Prescripteur
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur
PRE_INS_REG	Région du PS Prescripteur à partir de 2015
PSP_STI_SNDS	Statut Juridique PS Prescripteur
ETP_REG_COD	Région d'Implantation Etb Prescripteur à partir de 2015
ETP_CAT_SNDS	Catégorie Etb Prescripteur

TOP PS5	
TOP_PS5_TRG	Top Périmètre hors CMU C et prestations pour information

INDICATEURS	
FLT_ACT_COG	Coefficient Global de la Prestation Préfiltré
FLT_ACT_NBR	Dénombrement de la Prestation Préfiltré
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltré
FLT_PA1_MNT	Montant de la Dépense de la Prestation Préfiltrée
FLT_REM_MNT	Montant Versé/Remboursé Préfiltré
PRS_ACT_COG	Coefficient Global
PRS_ACT_NBR	Dénombrement
PRS_ACT_QTE	Quantité
PRS_DEP_MNT	Montant du Dépassement
PRS_PA1_MNT	Montant de la Dépense
PRS_REM_MNT	Montant Versé/Remboursé
PRS_REM_BSE	Base de Remboursement

FIGURE A.1 – Liste des variables de la base Open DAMIR

A.2 Qualité des données

Nom	Libellé	Contrôle d'exhaustivité	Contrôle de vraisemblance	Contrôle de cohérence	Contrôle distribution
FLX_ANN_MOI	Année et Mois de Traitement	Non vide			
SOI_ANN	Année de Soins	Non vide	2015 < SOI_ANN < 2020	SOI_ANN ≤ FLX_ANN_MOI	Test du poids des années
SOI_MOI	Mois de Soins	Non vide	1 ≤ SOI_MOI ≤ 12		Test du poids des mois
BEN_SEX_COD	Sexe du Bénéficiaire	Non vide	1 ou 2 (MASCULIN ou FEMININ)		Test du poids des sexes
AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins	Non vide	≠ 99 (AGE INCONNU)		Test du poids des tranches d'âge
BEN_RES_REG	Région de Résidence du Bénéficiaire	Non vide	≠ 99 (INCONNU)		Test du poids des régions
PRS_NAT	Nature de Prestation	Non vide	≠ 9999 (INCONNU) et 0 (SANS OBJET)	Signification du code inscrite dans le lexique Open DAMIR	
PRS_ACT_QTE	Quantité	Non vide	PRS_ACT_QTE > 0		Test de distribution
PRS_PA1_MNT	Montant de la Dépense	Non vide	PRS_PA1_MNT ≥ 0	PRS_PA1_MNT ≠ 0 si PRS_Rem_MNT = 0 PRS_PA1_MNT > 0 si prestation de référence ou complément d'acte	
PRS_DEP_MNT	Montant du Dépassement	Non vide	PRS_DEP_MNT ≥ 0	PRS_DEP_MNT ≤ PRS_PA1_MNT	
PRS_Rem_TAU	Taux de Remboursement	Non vide	0 ≤ PRS_Rem_TAU ≤ 1	PRS_Rem_TAU = 0 si PRS_Rem_MNT = 0 et PRS_Rem_BSE = 0	
PRS_Rem_BSE	Base de Remboursement	Non vide	PRS_Rem_BSE ≥ 0	PRS_Rem_BSE ≤ PRS_PA1_MNT si prestation de référence ou complément d'acte PRS_Rem_BSE > 0 si PRS_Rem_MNT > 0 et si prestation de référence ou complément d'acte	
PRS_Rem_MNT	Montant Versé/Remboursé	Non vide	PRS_Rem_MNT ≥ 0	PRS_Rem_MNT = PRS_Rem_BSE x PRS_Rem_TAU si prestation de référence ou complément d'acte (Avec un delta de 10 cts x PRS_ACT_QTE) PRS_Rem_MNT ≤ PRS_PA1_MNT si prestation de référence ou complément d'acte	Test de distribution
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée	Non vide	FLT_ACT_QTE ≥ 0	FLT_ACT_QTE ≤ PRS_ACT_QTE FLT_ACT_QTE = 0 si FLT_PA1_MNT = 0 FLT_ACT_QTE > 0 si FLT_PA1_MNT > 0	Test de distribution
FLT_PA1_MNT	Montant de la Dépense de la Prestation Préfiltrée	Non vide	FLT_PA1_MNT ≥ 0	FLT_PA1_MNT ≤ PRS_PA1_MNT FLT_PA1_MNT ≥ PRS_Rem_MNT si prestation de référence FLT_PA1_MNT ≥ PRS_Rem_BSE si prestation de référence FLT_PA1_MNT > 0 si prestation de référence ; FLT_PA1_MNT = 0 sinon	Test de distribution
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltrée	Non vide	FLT_DEP_MNT ≥ 0	FLT_DEP_MNT ≤ FLT_PA1_MNT FLT_DEP_MNT ≤ PRS_DEP_MNT	
PRS_Rem_TYP	Type de Remboursement	Non vide	≠ 99 (INCONNU)		
CPL_COD	Complément d'Acte	Non vide	€ [0,1,2,3]	CPL_COD €[1,2,3] si complément d'acte	
PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)	Non vide	≠ 99 (INCONNU)		
PRS_PPU_SEC	Code Secteur Privé/Public	Non vide	€ [1,2]		
BEN_QLT_COD	Qualité du Bénéficiaire	Non vide	≠ 99 (INCONNU)		
ETE_TYP_SNDS	Type Etb Exécutant	Non vide			
PSE_ACT_CAT	Catégorie de l' Exécutant	Non vide	≠ 99 (INCONNU)		
PSP_ACT_CAT	Catégorie du Prescripteur	Non vide	≠ 99 (INCONNU)		
PSE_SPE_SNDS	Spécialité Médicale PS Exécutant	Non vide	≠ 99 (INCONNU) et 0 (NON RENSEIGNE)		
PSP_SPE_SNDS	Spécialité Médicale PS Prescripteur	Non vide	≠ 99 (INCONNU) et 0 (NON RENSEIGNE)		
PSE_ACT_SNDS	Nature d'Activité PS Exécutant	Non vide	≠ 99 (INCONNU) et 0 (NON RENSEIGNE)		
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur	Non vide	≠ 99 (INCONNU) et 0 (NON RENSEIGNE)		

FIGURE A.2 – Tableau récapitulatif des contrôles de qualité des données réalisés

A.3 Évolutions réglementaires

Certaines prestations ont subi une évolution de leur base de remboursement. Des montants "as-if" sont alors calculés. Ils correspondent aux montants qui auraient été observés si les tarifs actuels étaient en vigueur. Ainsi, les montants moyens sont stables et représentatifs du futur. Cette liste détaille les prestations retraitées.

Avis de consultant :

- Pour les médecins généralistes et spécialistes :

Au 1er octobre 2017, les codes APC (Avis Ponctuel de Consultant) et APV (Avis Ponctuel de consultant en Visite) sont créés. Ils remplacent les codes C2 et V2 et sont revalorisés de 46 € à 48 € (et de 55,20 € à 57,60 € dans les DROM¹). Ils subissent une nouvelle évolution au 1er juin 2018, de 48 € à 50 € (et de 57,60 € à 60 € dans les DROM¹).

- Pour les psychiatres, neurologues et neuropsychiatres :

Les codes APY (Avis Ponctuel de Consultant) et AVY (Avis Ponctuel de consultant en Visite) sont créés au 1er octobre 2017. Anciennement côtés C2,5 et V2,5, ils sont revalorisés de 57,50 € à 60 € (et de 69 € à 72 € dans les DROM¹). Ils subissent une nouvelle évolution au 1er juin 2018, de 60 € à 62,50 € (et de 72 € à 72 € dans les DROM¹).

Consultations et visites :

- Pour les médecins généralistes et spécialistes :

La consultation de base a été revalorisée de 23 € à 25 € au 1er mai 2017 pour les médecins généralistes de secteur 1 et de secteur 2 adhérents à l'OPTAM. Une nouvelle lettre clé G (resp. GS) avec une base de remboursement de 25 € est créée pour remplacer le C (resp. CS). De même pour les visites, la lettre VG (resp. VGS) est créée pour remplacer le V (resp. VS).

Les généralistes de secteur 2 hors OPTAM continuent à facturer le C/CS ou V/VS avec une base de remboursement de 23 €. Ils ne peuvent coter le G/GS ou VG/VGS (25 €) que pour les patients bénéficiant de la CMU-C ou de l'ACS et pour les consultations facturées au tarif opposable. De manière similaire, les consultations et visites des généralistes ont été revalorisées de 2 € dans les DROM.

- Pour les psychiatres, neurologues et neuropsychiatres :

La consultation CNP et la visite VNP sont revalorisées de 37 € à 39 € au 1er juillet 2017 (et de 44,40 € à 46,80 € dans les DROM).

- Pour les cardiologues :

Au 1er juillet 2017, la consultation CSC est revalorisée de 45,73 € à 47,73 € (et de 52,44 € à 54,73 € dans les DROM).

1. Ces tarifs concernent les DROM hors Mayotte.

Majorations :

- Majoration de coordination :
Elle est revalorisée de 3 € à 5 € au 1er juillet 2017 pour les médecins généralistes et spécialistes et de 4 € à 5 € pour les psychiatres, neuropsychiatres et neurologues.
- Majoration Généraliste Enfant :
Une majoration MEG (Majoration Enfant Généraliste) de 5 € est créée à partir du 1er mai 2017. Elle remplace la majoration pour les enfants de moins de 2 ans (MNO, 5 €) et celle pour les enfants de 2 à 6 ans (MGE, 3 €).
- Majoration Enfant Pédiatre :
Au 1er mai 2017, la nomenclature évolue avec la création d'une nouvelle majoration MEP (Majoration Enfant Pédiatre) à 4 €. Elle remplace les codes MPE (Majoration Pédiatre Enfant) et MNP (Majoration Nourrisson Pédiatre), tous deux à 3 €.
- Autres majorations :
Au 1er novembre 2017, la Majoration Première consultation Famille (MPF) et la Majoration consultation Annuelle Familiale (MAF) sont revalorisées de 10 à 20 €. La Majoration Consultation Endocrinologue (MCE) est revalorisée de 10 à 16 €. La Majoration consultation Appareillage (MTA) est revalorisée de 20 à 23 €.

Autres :

- Participation forfaitaire de 18/24 € :
La participation forfaitaire, s'appliquant sur les actes médicaux dont le tarif est supérieur ou égal à 120 € ou ayant un coefficient supérieur ou égal à 60, a été revalorisée de 18 € à 24 € au 1er janvier 2019.
- Forfait journalier :
Au 1er janvier 2018, le forfait journalier hospitalier est passé de 18 € à 20 € par jour en hôpital ou en clinique et de 13,50 € à 15 € par jour dans le service psychiatrique d'un établissement de santé.
- 100 % santé pour le poste Dentaire :
Les premières mesures de la réforme 100 % santé en dentaire sont entrées en vigueur au 1er avril 2019. Cette réforme engendre une revalorisation progressive de soins bucco-dentaires fréquents et l'instauration progressive de plafonds tarifaires pour certains actes prothétiques.

A.4 Traitement des anomalies après agrégation de la base d'étude

Suite à l'agrégation de la base d'étude, 3 % des lignes possèdent une anomalie. Ces anomalies sont liées au rassemblement des différents types d'actes. Trois sortes d'anomalies sont repérées et corrigées. Voici le détail des traitements réalisés.

- Anomalie 1 : compléments d'acte ou prises en charge complémentaires associés à aucune prestation de référence

Suite à l'agrégation, il existe des lignes qui ne contiennent que des compléments d'acte et des prises en charge complémentaires mais pas de prestation de référence associée. Cette anomalie ne concerne que 0,03 % des compléments d'acte et des prises en charge complémentaires de la base. Les lignes en anomalie sont donc supprimées.

- Anomalie 2 : compléments d'actes associés à des prestations de référence mais avec des volumes incohérents

Certaines lignes de la base de données possèdent une base de remboursement supérieure au montant de frais réels diminué du montant du dépassement. Cette anomalie concerne les lignes possédant au moins un complément d'acte. Elle ne peut pas concerner les remboursements au titre de prises en charge complémentaires puisque leur base de remboursement n'a pas été prise en compte pour éviter les doubles comptages.

81 % des compléments d'acte se trouvent sur une ligne concernée par l'anomalie. Il n'est pas envisageable de tous les supprimer. Il est cependant impossible d'identifier précisément les compléments d'acte qui n'ont pas de prestation de référence associée sur chaque ligne en anomalie, ni les montants remboursés correspondant. Le nombre de compléments d'acte à supprimer est donc estimé.

1. Avant l'agrégation réalisée sans conserver le type de remboursement, le montant remboursé et la base de remboursement de chaque ligne de compléments d'acte sont stockés dans une nouvelle variable afin d'être sommés durant l'agrégation et de pouvoir être utilisés pour calculer le montant moyen remboursé et la base de remboursement moyenne d'un complément d'acte pour chaque ligne de la base agrégée.
2. Pour chaque ligne en anomalie, l'"Écart observé" est défini comme étant la différence entre la base de remboursement et le montant de frais réels diminué du montant du dépassement.
3. Le nombre de compléments d'actes erronés pour chaque ligne en anomalie est alors estimé de la manière suivante :

Nombre de compléments d'acte erronés estimé

$$= \frac{\text{Écart observé}}{\text{Base de remboursement moyenne d'un complément d'acte}}$$

4. La distribution du nombre de compléments d'acte erronés estimé est ensuite observée. Elle est centrée autour des valeurs entières. Ce nombre estimé est alors arrondi pour obtenir le nombre de compléments d'actes à supprimer.



FIGURE A.3 – Distribution du nombre de compléments d'acte à supprimer estimé

5. Afin de corriger l'anomalie, le nombre de compléments d'acte à supprimer est retiré du nombre de compléments d'acte. La base de remboursement associée aux compléments d'acte erronés estimés, calculée en prenant le nombre de compléments d'acte erronés estimé multiplié par la base de remboursement moyenne d'un complément d'acte, est retirée de la base de remboursement totale de la ligne. De la même manière, le montant remboursé associé aux compléments d'acte erronés estimés, correspondant au nombre de compléments d'acte erronés estimé multiplié par le montant remboursé moyen d'un complément d'acte, est soustrait du montant remboursé total.

Avec cette méthode, seulement 4 % des compléments d'actes de la base sont supprimés.

6. Suite à ces étapes, les lignes pour lesquelles le nombre de compléments d'acte à supprimer estimé a été arrondi à l'entier inférieur restent en anomalie. La base de remboursement de ces lignes est toujours légèrement supérieure au montant de frais réel diminué du montant de dépassement. Elles sont retraitées en mettant la base de remboursement égale au montant de frais réel diminué du montant de dépassement.

- Anomalie 3 : prises en charge complémentaires associées à des prestations de référence mais avec des volumes incohérents

Certaines lignes de la base de données possèdent un montant remboursé supérieur à la base de remboursement. Cette anomalie concerne uniquement les lignes possédant au moins un remboursement au titre d'une prise en charge complémentaire. Elle ne concerne pas les prestations de référence ni les compléments d'acte puisque, pour ces deux types de remboursement, les montants remboursés et les bases de remboursement sont prises en compte et leur cohérence a été vérifiée lors des contrôles de qualité des données effectués sur les bases mensuelles. L'anomalie vient donc du fait que la base de remboursement n'est pas prise en compte pour les prises en charge complémentaires.

0,8 % des remboursements complémentaires se trouvent sur une ligne concernée par l'anomalie. La méthode appliquée pour l'anomalie précédente est utilisée. 0,2 % des prises en charge complémentaires sont finalement supprimées.

Annexe B

L'analyse de l'inflation

B.1 Méthodes de sélection de variables basées sur des forêts aléatoires

Trois méthodes ont été utilisées et sont rappelées ici.

- L'algorithme RFE (*Recursive Feature Elimination*) avec les forêts aléatoires :

La méthode RFE est probablement la méthode la plus utilisée pour la sélection de variables. Il s'agit d'une méthode de type *backward* qui élimine itérativement les variables les moins importantes. Les forêts aléatoires s'intègrent bien dans ce type d'algorithme grâce à leur mesure de l'importance des variables explicatives sur la variable à expliquer. L'algorithme fonctionne de la manière suivante. À chaque itération, une forêt aléatoire est construite. L'importance de chaque variable est calculée et la variable la moins importante est retirée du modèle. Ces étapes sont répétées jusqu'à ce qu'il ne reste plus de variables. L'erreur de prédiction du modèle est calculée à chaque itération. Le sous-ensemble de variables minimisant cette erreur est retenu. En pratique, l'algorithme RFE est incorporé dans une boucle de rééchantillonage, comme par exemple une validation croisée.

Cet algorithme permet de plus d'obtenir un classement des variables corrigé des effets de la corrélation. En effet, comme mentionné précédemment, la mesure d'importance par permutation peut être affectée par la corrélation entre les variables [13]. Cet algorithme retire au fur et à mesure les variables corrélées et recalcule l'importance à chaque étape, ce qui permet aux variables corrélées non supprimées de se positionner à leur vraie place dans le classement. D'un point de vue interprétation, il convient de noter que cet algorithme ne retient pas toutes les variables liées à la variable à expliquer lorsque celles-ci sont corrélées.

- La méthode de Boruta :

La méthode de Boruta est une méthode de sélection de variables souvent citée dans la littérature. Elle permet de sélectionner toutes les variables liées à la variable à expliquer en jugeant de la pertinence des variables. Cette méthode est utilisée lorsque l'objectif recherché est la compréhension des mécanismes liés à la variable d'intérêt, plutôt la seule construction d'un modèle prédictif de type "boîte noire" avec une bonne précision de prédiction.

L'algorithme de Boruta fonctionne de la manière suivante. Il commence par ajouter un caractère aléatoire à la base de données en créant des copies randomisées de toutes les variables explicatives. Autrement dit, les variables sont dupliquées puis leurs valeurs sont mélangées. Une forêt aléatoire est ensuite ajustée sur l'ensemble de données obtenu et l'importance de chaque variable est calculée en utilisant le Z-score. Comme l'importance d'une caractéristique peut être non nulle uniquement en raison de fluctuations aléatoires, les importances des variables mélangées sont utilisées pour décider quelles variables sont vraiment importantes. L'importance de chaque variable des données initiales est alors comparée à un seuil qui est défini comme la plus grande importance enregistrée parmi les copies mélangées des variables. Lorsque l'importance d'une caractéristique est supérieure à ce seuil, l'algorithme marque la caractéristique comme étant importante.

L'idée est qu'une variable n'est utile que si elle est capable de faire mieux que la meilleure variable randomisée. La mesure de l'importance dépend notamment de la réalisation particulière des variables mélangées. Par conséquent, l'algorithme effectue plusieurs itérations pour valider l'importance ou non de chaque variable. Le nombre de fois où une caractéristique s'est mieux comportée que les caractéristiques mélangées est compté et une distribution binomiale est utilisée pour valider ou non l'importance.

Il n'y a ainsi pas de seuil strict entre la zone de refus et la zone d'acceptation mais trois zones. La première zone est la zone de refus (marquée par des carrés rouges sur le schéma ci-dessous) dans laquelle les variables sont considérées comme du bruit et sont supprimées de l'ensemble de données. La deuxième est une zone d'irrégularité (la zone blanche), pour laquelle l'algorithme est indécis quant aux caractéristiques qui se trouvent dans cette zone. La dernière zone est la zone d'acceptation (la zone verte). Les caractéristiques qui se trouvent dans cette zone sont considérées comme pertinentes et sont conservées. Les zones sont définies en sélectionnant les queues de la distribution, par défaut chaque queue représente 1% de la distribution. Cette boucle est illustrée sur la figure B.1.

Cette boucle est réitérée, sans les variables éliminées lors des itérations précédentes. L'algorithme s'arrête soit lorsque toutes les caractéristiques conservées sont confirmées, soit lorsqu'il atteint un nombre maximal prédéfini d'itérations (100 par défaut).

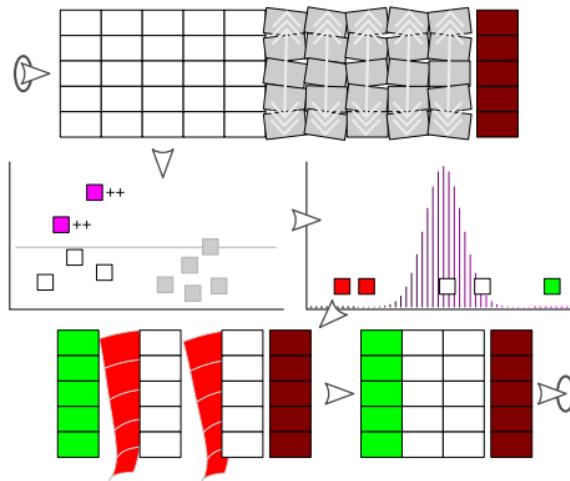


FIGURE B.1 – Illustration de la boucle principale de l'algorithme de Boruta - Source [14]

- La méthode VSURF :

La méthode VSURF (*Variable Selection Using Random Forests*) est retenue pour son fonctionnement en plusieurs étapes permettant notamment d'obtenir un sous-ensemble comprenant les variables fortement corrélées avec la variable à expliquer. Cette méthode a été introduite par Genuer et al. [17] et est implémentée dans le package *VSURF* de R. Elle est composée de deux étapes.

La première étape est descendante et consiste à éliminer un grand nombre de variables inutiles, qui sont définies par une faible importance. Pour cela, un seuil est estimé en utilisant les écart-types des importances et toutes les variables ayant une importance inférieure à ce seuil sont supprimées. Les écart-types des importances sont utilisés puisque la variabilité des importances est plus faible pour les variables inutiles.

La seconde étape est ascendante et consiste à introduire des variables dans des forêts aléatoires. Elle permet de créer deux sous-ensembles de variables, répondant à deux objectifs différents.

Le premier sous-ensemble est construit dans un but d'interprétation. Les variables sélectionnées sont celles qui sont fortement reliées à la variable à expliquer, même si elles sont corrélées entre elles. En notant m le nombre de variables sélectionnées suite à la première étape, la séquence emboîtée de forêts impliquant les k premières variables est construite, pour k allant de 1 à m . Les moyennes des erreurs OOB des modèles emboîtés sont calculées, en général sur 25 forêts. L'idée est de sélectionner les variables du modèle ayant la plus petite erreur OOB. Afin de faire face à l'instabilité, les variables du plus petit modèle avec une erreur OOB inférieure à l'erreur OOB minimale augmentée de son écart-type (basée sur les 25 mêmes répétitions de forêts) sont retenues.

Le second sous-ensemble retenu par la méthode répond à un objectif de prédiction. Le plus petit sous-ensemble de variables suffisant pour bien prédire la variable à expliquer est recherché. Cet ensemble contiendra peu de variables avec très peu de corrélation entre elles. Une suite ascendante de forêts aléatoires est construite en ajoutant les variables de manière progressive dans l'ordre d'importance. Une variable n'est ajoutée que si la diminution de l'erreur OOB est supérieure à un seuil. L'idée est que la diminution de l'erreur OOB doit être significativement supérieure à la variation moyenne obtenue en ajoutant des variables bruitées. Les sous-ensemble sélectionné contient les variables du dernier modèle construit.

Le principal inconvénient de cette méthode est qu'elle ne permet pas de détecter toutes les variables liées à la variable à expliquer dans le cas de nombreux prédicteurs corrélés. La stratégie proposée par Boruta est plus précise que VSURF pour récupérer les faibles corrélations redondantes entre les prédicteurs et la variable à expliquer. Le risque avec la méthode de Boruta est de sélectionner des "faux positifs". La méthode VSURF est plus parcimonieuse que la méthode de Boruta, dans le sens où elle sélectionne moins de variables.

B.2 Compléments de l'étude de l'inflation sur la famille "Médicaments PH65"

B.2.1 Test de Dunn : Inflation ~ Âge

Le test de Dunn est employé à la suite d'un test de Kruskal-Wallis ayant permis de rejeter l'hypothèse H_0 selon laquelle les échantillons proviennent de la même population ou de populations ayant des caractéristiques identiques. Il s'agit d'une procédure de comparaisons multiples. Il permet de réaliser des comparaisons par paires afin de déterminer quelles sont les médianes spécifiques qui sont significativement différentes par rapport aux autres.

Ce test est utilisé ici dans le cadre de l'étude de l'association entre l'âge et l'inflation, après avoir observé qu'au moins un échantillon associé à une tranche d'âges possède une médiane significativement différente d'un autre. Il permet d'observer que de nombreuses différences entre les catégories d'âge sont significatives.

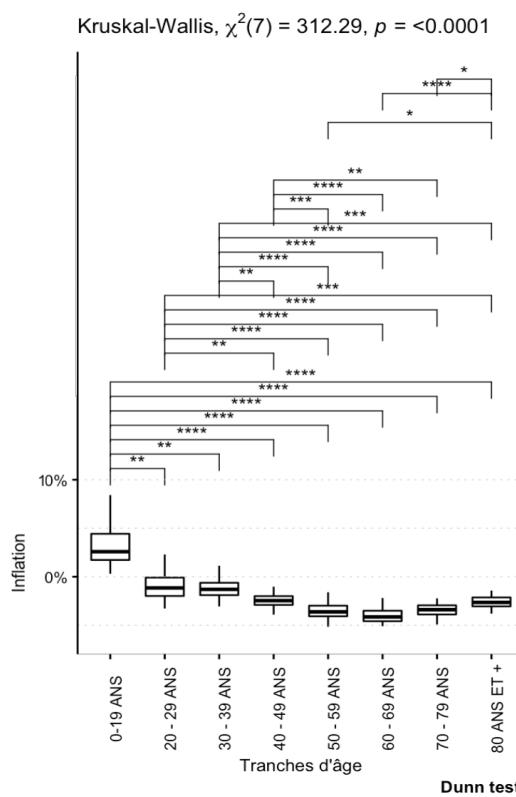


FIGURE B.2 – Test de Dunn : Inflation ~ Âge

B.2.2 Test basé sur les coefficients de corrélation de Spearman

La base de données utilisée pour les "Médicaments PH65" contient 40 variables quantitatives. Afin de simplifier l'affichage, seuls les résultats associés aux variables relatives aux proportions de professionnels de santé prescripteurs sont présentés. Ce choix s'explique notamment par le fait que les variables les plus corrélées avec l'inflation au sens de la corrélation de Spearman sont des proportions de professionnels de santé prescripteurs.

L'hypothèse nulle est rejetée pour 7 variables parmi les 9 au seuil de significativité de 1% (p-value << 1%), ce qui signifie qu'il existe une relation monotone significative entre ces variables et l'inflation.

Variable : Proportion de prescripteurs	Coefficient de corrélation de Spearman	P-value du test
OPHTALMOLOGIE	-0,72	< 0,01%
PÉDIATRIE	0,70	< 0,01%
RHUMATOLOGIE	-0,70	< 0,01%
DERMATOLOGIE - VÉNÉRÉOLOGIE	0,67	< 0,01%
MÉDECINE GÉNÉRALE	-0,51	< 0,01%
GYNÉCOLOGIE - SAGE-FEMME	0,36	< 0,01%
DENTAIRE	0,25	< 0,01%
CHIRURGIE - ANESTHÉSIOLOGIE	-0,07	13%
PSYCHIATRIE - NEUROLOGIE	-0,02	68%

TABLE B.1 – Résultats du test basé sur les coefficients de corrélation de Spearman

B.2.3 Modèle sans les données relatives à la tranche d'âges 0-19 ans

Une forêt aléatoire est implémentée sans les données relatives à la classe d'âge 0-19 ans afin de vérifier que les résultats obtenus sont cohérents avec ceux de la forêt aléatoire contenant toutes les tranches d'âge. Les variables importantes du modèle et leurs effets sont présentés ici.

Il apparaît que les variables les plus importantes sont identiques à celles du premier modèle, à l'exception des proportions de prescripteurs rhumatologues et pédiatres. La proportion de prescripteurs pédiatres n'est pas une variable pertinente lorsque la tranche d'âges 0-19 ans est absente des données. La proportion de prescripteurs rhumatologues servait principalement à différencier la tranche d'âges associée aux jeunes, ce qui explique que cette variable ne soit plus autant importante dans ce modèle. Les tendances observées sur les autres variables sont conformes aux observations issues du modèle contenant la tranche d'âges 0-19 ans.

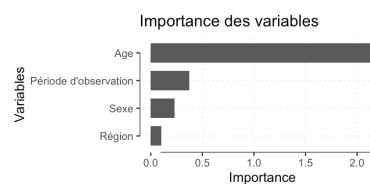


FIGURE B.3 – Importance des variables de la granularité - Sans les 0-19 ans

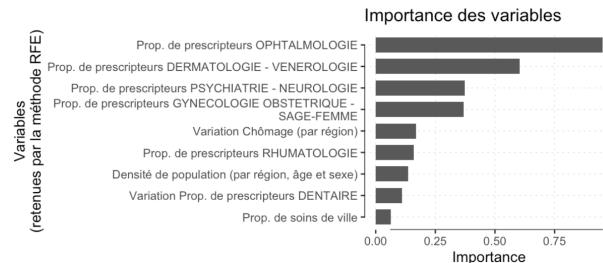


FIGURE B.4 – Importance des variables - Sans les 0-19 ans

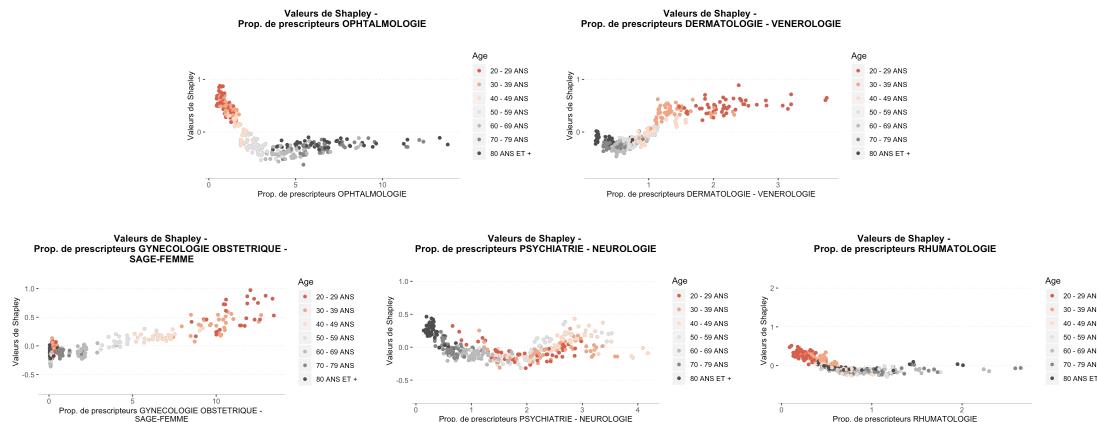


FIGURE B.5 – Valeurs de Shapley associées aux proportions de prescripteurs - Sans la tranche d'âges 0-19 ans

B.2.4 Résultat du modèle linéaire

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.10572	0.10307	20.430	< 2e-16 ***
Prop. de prescripteurs DERMATOLOGIE - VENEROLOGIE2	0.18177	0.05135	3.540	0.000449 ***
Prop. de prescripteurs DERMATOLOGIE - VENEROLOGIE3	0.18562	0.06321	2.937	0.003517 **
Prop. de prescripteurs DERMATOLOGIE - VENEROLOGIE4	0.18970	0.06943	2.732	0.006579 **
Prop. de prescripteurs GYNECOLOGIE OBSTETRIQUE - SAGE-FEMME2	0.12324	0.04638	2.657	0.008208 **
Prop. de prescripteurs GYNECOLOGIE OBSTETRIQUE - SAGE-FEMME3	0.16295	0.05059	3.221	0.001386 **
Prop. de prescripteurs GYNECOLOGIE OBSTETRIQUE - SAGE-FEMME4	0.15558	0.05825	2.671	0.007886 **
Prop. de prescripteurs OPHTALMOLOGIE2	-0.13108	0.05723	-2.290	0.022546 *
Prop. de prescripteurs OPHTALMOLOGIE3	-0.44157	0.07571	-5.833	1.15e-08 ***
Prop. de prescripteurs OPHTALMOLOGIE4	-0.31652	0.08387	-3.774	0.000186 ***
Prop. de prescripteurs OPHTALMOLOGIES	-0.27631	0.09957	-2.775	0.005784 **
Prop. de prescripteurs PSYCHIATRIE - NEUROLOGIE2	-0.22502	0.05175	-4.348	1.76e-05 ***
Prop. de prescripteurs PSYCHIATRIE - NEUROLOGIE3	-0.09316	0.06201	-1.502	0.133806
Prop. de prescripteurs PSYCHIATRIE - NEUROLOGIE4	-0.12141	0.08327	-1.458	0.145638
Prop. de prescripteurs RHUMATOLOGIE2	-0.49947	0.09458	-5.281	2.15e-07 ***
Prop. de prescripteurs RHUMATOLOGIE3	-0.56199	0.10483	-5.361	1.42e-07 ***
Prop. de prescripteurs RHUMATOLOGIE4	-0.61509	0.10873	-5.657	2.99e-08 ***
Prop. soins de ville2	0.02232	0.05935	0.376	0.707095
Prop. soins de ville3	-0.06381	0.05426	-1.176	0.240240
Prop. soins de ville4	-0.15493	0.05102	-3.037	0.002554 **
Variation Prop. de prescripteurs DENTAIRE2	-0.08285	0.03474	-2.385	0.017571 *
Variation Prop. de prescripteurs DENTAIRE3	-0.07401	0.03961	-1.868	0.062457 .
Variation Prop. de prescripteurs DENTAIRE4	0.05151	0.05895	0.874	0.382801
Densité de population (par région, âge et sexe)2	-0.08186	0.03904	-2.097	0.036656 *
Densité de population (par région, âge et sexe)3	-0.14299	0.04659	-3.069	0.002299 **
Densité de population (par région, âge et sexe)4	-0.03410	0.06074	-0.561	0.574836
Variation Chômage (par région)2	0.08613	0.04051	2.126	0.034130 *
Variation Chômage (par région)3	0.18330	0.02989	6.133	2.13e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

FIGURE B.6 – Résultat du modèle linéaire

B.3 Compléments de l'étude de l'inflation sur la famille "Généralistes"

B.3.1 Tests de Dunn

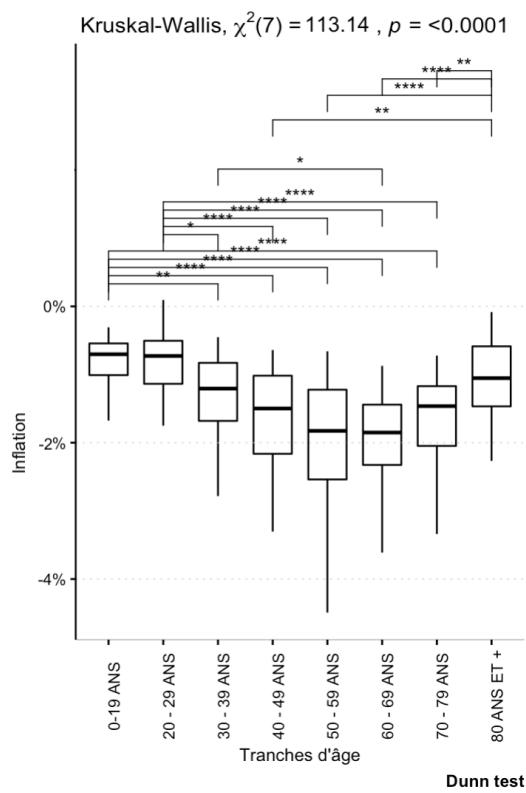


FIGURE B.7 – Test de Dunn : Inflation ~ Âge

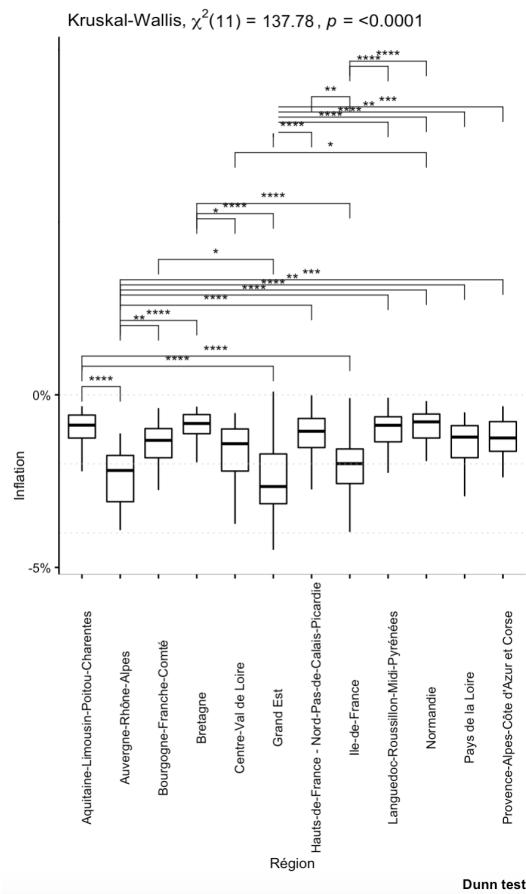


FIGURE B.8 – Test de Dunn : Inflation ~ Région

B.3.2 Tests basé sur les coefficients de Spearman

Variable	Corrélation de Spearman	P-value du test
Densité d'établissements FINESS (par région)	0,49	< 0,01%
Niveau de vie médian (par région)	-0,46	< 0,01%
Densité de Centres hospitaliers FINESS (par région)	0,44	< 0,01%
Espérance de vie (par région et sexe)	-0,33	< 0,01%
Taux de mortalité (par région)	0,29	< 0,01%
Taux VQS de bon niveau (par région et sexe)	0,27	< 0,01%
Densité de Centres de santé FINESS (par région)	-0,24	< 0,01%
Variation Densité de médecins généralistes (par région)	0,19	0,01%
Variation Densité d'établissements FINESS (par région)	-0,18	0,05%
Taux de changements de région	-0,15	0,40%

TABLE B.2 – Résultats des tests basés sur les coefficients de corrélation de Spearman, associés aux variables pour lesquelles le test est rejeté au seuil de significativité $\alpha = 1$

Bibliographie

- [1] *Open Data de l'Assurance Maladie.*
<http://open-data-assurance-maladie.ameli.fr>.
- [2] AILLIOT P. *Modèles linéaires.* Cours de Master 1, 2018.
- [3] BARBASTE M. *Une méthode de provisionnement individuel par apprentissage automatique.* Mémoire d'actuariat, 2017.
- [4] BREIMAN L. *Bagging predictors.* Machine Learning, 1996.
- [5] BREIMAN L. *Random Forests.* Machine Learning, 2001.
- [6] CAILLOL H. *Ouverture des données de santé : l'expérience de l'Assurance maladie.* Informations sociales, n° 191, pages 60 à 67, 2015.
- [7] Commission Open Data en santé. *Rapport de la commission Open Data en santé.* 2014.
- [8] DELCAILLAU D. *Contrôle et Transparence des modèles complexes en actuariat.* Mémoire d'actuariat, 2017.
- [9] DREES. *La complémentaire santé : acteurs, bénéficiaires, garanties.* 2019.
- [10] GAUVILLE R. *Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des SdS.* Mémoire d'actuariat, 2017.
- [11] HANIN J. *La réforme 100% Santé : quelles origines et quels effets ?* Mémoire d'actuariat, 2019.
- [12] MARCIANO L. *Modélisation de la dérive des soins de santé à court terme.* Mémoire d'actuariat, 2018.
- [13] GREGORUTTI B. MICHEL B., SAINT-PIERRE P. *Corrélation et importance des variables dans les forêts aléatoires.*
- [14] MIRON B. KURSA. *Boruta for those in a hurry.* 2020.
- [15] BREIMAN L. FRIEDMAN J. OLSHEN R., STONE C. *Classification and Regression Trees.* Chapman & Hall, New York, 1984.
- [16] GENUER R. POGGI J.M. *Arbres CART et Forêts aléatoires, Importance et sélection de variables.* 2017.
- [17] GENUER R. POGGI J.M., TULEAU-MALOT C. *Variable selection using random forests.* Pattern Recognition Letters, 2010b.
- [18] VERMET F. *Outils avancés pour la data science.* Cours de Master 1, 2018.