

Mémoire présenté le : 16 octobre 2018

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

Par : M. Arnold Caleb MEKONTSO FOTSING

Titre : *L'open DAMIR : apport à la maîtrise des dépenses de santé*

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

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

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

signature

Entreprise :

Nom : CNP Assurances

Signature : *[Signature]*
4 Place Raoul Dautry
75716 PARIS Cedex 15

Mme Catherine PIGEON

Directeur de mémoire en entreprise :

Nom : Mme Khadidiatou-DIENG

Signature : *[Signature]*

M. Olivier BOUGAREL

Membres présents du jury de l'ISFA

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

[Signature]

Signature du candidat

[Signature]

Résumé

Mots clés : *Marché français de l'assurance maladie, tarification risque santé, provision pour sinistres à payer, provision pour risques croissants, open data, imputation des données manquantes, arbre de régressions, forêts aléatoires, modèle linéaire généralisé, série chronologique.*

«Après plus d'un quart de siècle de déficits continus - une génération entière -, le retour à l'équilibre de l'assurance maladie constitue une priorité majeure». Cette phrase est extraite du rapport 2017 de la cour des comptes sur l'application des lois de financement de la Sécurité Sociale. Elle illustre l'accélération du rythme des mesures prises ces dernières années et visant à la maîtrise des coûts. Parmi ces mesures, l'Ondam fixe des objectifs de dépenses, la législation intervient sur les niveaux de remboursement des contrats des organismes complémentaires à la Sécurité Sociale, Mutuelles, Institutions de Prévoyance et Compagnies d'Assurances, via des contrats dits responsables. Leur objectif est de responsabiliser les assurés et certaines professions de santé. Afin que ces mesures de maîtrise des coûts n'entraînent pas de renoncement aux soins, un nouveau dispositif visant un reste à charge égal à zéro pour les assurés va prochainement être mis en place.

Parallèlement à ces éléments, s'ouvre l'ère de la digitalisation et ses promesses de bases de données conséquentes permettant d'appréhender une nouvelle manière d'aborder l'assurance santé. *L'Open DAMIR* (base complète sur les Dépenses d'Assurance Maladie Inter Régimes) en est un exemple. Cette base de données brute est une extraction du Système National d'Information Inter Régime de l'Assurance Maladie (SNIIRAM) qui rend compte des dépenses nationale de santé.

Ce mémoire s'intéresse particulièrement à ce jeu de données publiques. Son caractère relativement récent nécessite une approche prudente. Des méthodes de complétion de données ont dû notamment être mises en œuvre. À partir des données traitées, il propose des statistiques descriptives des dépenses réelles de santé, des prestations de la sécurité sociale et des restes à charges sécurité sociale par année, région, tranche d'âge et genre. Les informations obtenues ont ensuite été confrontées à celles issues d'une base d'assurés du portefeuille CNP Assurances. Ces éléments ont permis l'étude d'une approche du "reste à charge zéro", promesse de campagne du Président Macron, tant en terme de niveau de garanties qu'en terme de tarification.

Abstract

Key words : *French health insurance market, pricing health risk, provisions for claims to be paid, provisions for increasing risks, open data, imputations of missing data, regression tree, random forest classifier, generalized linear model, time series.*

“After more than a quarter century of continuing deficits - a full generation -, the return to a balance of health insurance is a major priority.” This sentence is quoted from the 2017 Court of Audit report on the application of Social Security financing laws. It illustrates the acceleration of the pace of the measures taken in the recent years, aiming to control the costs. Amongst these measures, the National Health Insurance Spending Objective (ONDAM) fixes spending targets. The legislation intervenes on the reimbursement levels of Social Security additional organizations, health mutuals, pension institutions and insurance companies, via so-called responsible contracts. Their goal is to empower contract holders and certain health professions. In order for these cost control measures to not lead to people declining health care, a new measure aiming for a zero-residual cost for health insurance contract holders will be implemented soon.

In addition to these elements, an era of digitalization opens up, along with its promises of substantial databases that help to apprehend a new way to address health insurance. The Open Health Insurance Interplan Expenses (Open DAMIR) is an example. This gross database is retrieved from the Health Insurance Interplan National Information System (SNIRAM), which reports national health expenses.

This thesis is particularly interested in this public dataset. Its relative recentness requires a cautious approach. The completion of data methods had to be implemented. From the processed data, it offers descriptive statistics of actual expenditures of health, social security benefits and the costs that remain the responsibility of social security per year by region, age and genre. The information obtained was then confronted to that of a CNP Assurances portfolio database. These elements made possible the research of a zero-residual cost approach, one of the campaign promises of French President Emmanuel Macron, equally in terms of levels of guarantees as in terms of pricing.

Remerciements

Mes remerciements s'adressent en premier lieu à mes directeurs de mémoire Mme Khadidia-tou DIENG et M. Frédéric PLANCHET pour leur aide précieuse, leur disponibilité et pour les orientations d'étude qu'ils m'ont fournies.

Je remercie également M. Éric DEMOLLI et M. Pierre LAFARGUE, mes responsables, pour les précisions et informations techniques qu'ils m'ont apportées . Un grand merci à Mme Marie BOUIS pour son appui logistique.

Je remercie particulièrement Mme Nadia EL MASRIOUI, M. Sylvain CARO et M. Thomas PETEUL pour la relecture de mon mémoire et pour les suggestions proposées.

J'exprime ma reconnaissance à chacun de mes collègues pour la charge de travail supplémentaire supportée ces derniers mois et pour ces échanges qui m'ont aidés dans mes analyses.

Une pensée pour Éthan, Sérah et Asaël.

Sommaire

I	<i>Le système d'assurance maladie</i>	10
1	La Sécurité Sociale et l'Assurance Maladie Obligatoire	12
1.1	Généralité sur la sécurité sociale	12
1.1.1	Les régimes de la sécurité sociale	12
1.1.2	Les branches de la sécurité sociale	12
1.2	L'Assurance Maladie Obligatoire (AMO)	15
1.2.1	Un financement multiple	16
1.2.2	Des prestations diversifiées	16
2	L'Assurance Maladie Complémentaire	19
2.1	Principe de l'assurance maladie complémentaire	19
2.1.1	Les garanties du contrat responsable	20
2.1.2	Les remboursements supplémentaires	20
2.1.3	Le tiers payant	21
2.2	Les types de complémentaire santé	21
2.2.1	Les complémentaires santé collectives à adhésion obligatoire	21
2.2.2	Les complémentaires santé collectives à adhésion facultative	22
2.2.3	Les complémentaires santé individuelles	22
2.3	Le marché de l'assurance complémentaire	23
3	Tarification des contrats d'assurance maladie complémentaire	25
3.1	Quelques définitions	25
3.1.1	L'hospitalisation	25
3.1.2	Consultation et actes médicaux courants	26
3.1.3	L'optique	26
3.1.4	Le dentaire	27
3.1.5	Les autres actes médicaux	27
3.1.6	Principes de remboursement	28
3.2	Principe général de la tarification	29
3.2.1	La prime pure	29
3.2.2	Méthodologie du BCAC	30
3.2.3	La prime commerciale	31
4	Provisionnement et rentabilité d'un contrat assurance maladie complémen- taire	34
4.1	Les Provisions en assurance maladie	34
4.1.1	Provisionnement des sinistres non connus	34
4.1.2	Mise en œuvre de la méthode	36
4.1.3	La provision pour risque croissant	38
4.2	Rentabilité et pilotage d'un contrat d'assurance maladie complémentaire	42

II	<i>L'open DAMIR, traitements des données et statistiques descriptives</i>	46
5	Présentation générale des données DAMIR	48
5.1	Présentation générale	48
5.1.1	Source des données	48
5.1.2	Données et outils	52
5.2	Forme et type des données	52
5.2.1	Définitions (<i>Variable aléatoire et type d'une variable</i>)	52
5.2.2	Types des variables de base DAMIR	53
5.2.3	Les valeurs manquantes	53
6	Traitement des données manquantes et remarques sur les données atypiques	56
6.1	Les méthodes de complétions des données manquantes	56
6.1.1	Méthode <i>LOCF</i> (<i>Last Observation Carried Forward</i>)	56
6.1.2	Méthode d'imputation par la moyenne	56
6.1.3	Méthode d'imputation par la médiane	57
6.1.4	Méthode <i>LOESS</i> (<i>LOcal regrESSion</i>)	57
6.1.5	Méthode <i>kNN</i> (<i>k-Nearest Neighbors</i>)	59
6.1.6	Méthode <i>MissForest</i>	63
6.2	Comparaison et choix des modèles	64
6.2.1	Échantillonnage des données	64
6.2.2	Méthode de sélection du modèle	67
6.3	Les données atypiques	71
7	Statistiques descriptives de la base de données	73
7.1	Description qualitative	73
7.1.1	Nature d'assurance	74
7.1.2	Qualité du bénéficiaire	74
7.1.3	Âge du bénéficiaire	75
7.1.4	Couverture du bénéficiaire (CMU-C ou non)	76
7.1.5	Sexe du bénéficiaire	77
7.1.6	Région de résidence du bénéficiaire	78
7.2	Description quantitative	80
7.2.1	Les dépenses contre la maladie de 2009 à 2016	80
7.2.2	Les prestations de l'assurance maladie obligatoire	83
III	<i>Le reste à charge après le remboursement de la sécurité sociale</i>	88
8	Le reste charge des dépenses de santé	90
8.1	Évolution et répartition du reste à charge	90
8.1.1	Progression annuelle du reste à charge	90
8.1.2	Le reste à charge par genre	90
8.1.3	Le reste à charge par tranche d'âge, par nature d'acte et par région	92
8.2	Quelques mesures d'influences	94
8.2.1	Impact de la zone géographique, de la tranche d'âge et du genre sur le reste à charge	94
8.2.2	Impact du reste à charge sur la fréquence d'actes	102
8.3	Le reste à charge zéro	111
8.3.1	De quoi est-il question ?	111
8.3.2	Les prothèses dentaires, les verres optiques et les audioprothèses	111

9	Analyse temporelle du reste à charge sécurité sociale : cas des verres optiques	113
9.1	Chronique des données	113
9.2	Type de décomposition : additive ou multiplicative	115
9.2.1	Choix du type de décomposition par la méthode de la bande	115
9.2.2	Choix du type de décomposition par la méthode des profils	116
9.2.3	Méthode analytique du tableau de Buys-Ballot	117
9.3	Décomposition de la série	118
9.3.1	Méthode de Buys-Ballot généralisée (BB)	118
9.3.2	Méthode <i>seasonal decomposition of times series by loess(STL)</i>	120
9.3.3	Méthode moyenne mobile (MM)	120
9.4	Comparaison des modèles de décomposition et prévision	122
9.4.1	Les autocorrélogrammes	122
9.4.2	Tests sur les résidus	123
9.4.3	Prévision par la méthode STL	125
9.5	Autres méthodes de prévision	126
9.5.1	Lissage exponentiel de Hold-Winters	126
9.5.2	Prévision par l'ajustement d'un modèle SARIMA	128
9.6	Choix du meilleur modèle de prévision	133

INTRODUCTION GENERALE

Les organismes d'assurance sont aujourd'hui confrontés à des défis de plus en plus prégnants du fait d'un environnement socio-économique et réglementaire en perpétuel mouvement. Pour y faire face, la gestion et la maîtrise des risques sont des enjeux majeurs dans cet environnement de plus en plus complexe. Les assureurs doivent mettre en place des méthodes agiles de gestion des risques qui leur permettent d'identifier, de modéliser et de quantifier les risques auxquels ils sont confrontés. Outre les évolutions socio-économiques et réglementaires, le contexte est marqué par une forte transformation vers le digital qui se caractérise notamment par un mouvement dynamique en matière d'ouverture des données publiques.

Depuis le lancement du mouvement *open data*, un nombre grandissant de jeux de données numériques sont disponibles. L'accès et l'usage à ces données sont libres. A ce jour, 21420 séries de données sont disponibles sur la plate-forme ouverte des données publiques françaises. Ces données sont de natures diverses et constituent une énorme ressource encore trop peu exploitée. Le mouvement *open data* a pour objectif notamment de créer de la valeur et de favoriser l'innovation.

En matière de santé, l'Assurance Maladie met à disposition du public des données de cadrage et de bases brutes qui sont des extractions du système national d'information inter-régimes de l'Assurance Maladie. Ces données concernent principalement la biologie médicale, les médicaments et les dépenses d'assurance maladie.

Ce mémoire traite du *risque maladie*, ce risque auquel tout le monde est confronté et qui engendre des dépenses qui ne cessent d'augmenter du fait de l'élévation générale du niveau de vie, de l'accroissement démographique et du vieillissement de la population.

Ce mémoire s'intéresse à la maîtrise des dépenses de santé. Il propose une analyse en trois étapes :

- La première étape consiste à présenter le système d'assurance maladie français. Il s'agit d'une présentation qui s'est avérée nécessaire de rajouter lors de la rédaction de ce document afin de rendre son contenu accessible à tout type de lecteur.
- La deuxième partie porte sur le traitement et la description statistique des dépenses d'assurance maladie inter-régime.
- La troisième partie s'intéresse à la maîtrise des dépenses de santé par l'analyse des restes à charge sécurité sociale.

Depuis 2012 nous assistons à une légère hausse de la part des cotisations publiques dans le financement de l'assurance maladie obligatoire. Indépendamment des revenus des assurés sociaux, ce financement favorise un mécanisme de solidarité face aux dépenses contre la maladie. L'encadrement de ces dépenses est aujourd'hui indispensable. C'est dans ce contexte que s'inscrit le projet "reste à charge zéro" lancé par le Ministère de la Santé depuis le 23 janvier 2018. Ce

mémoire s'inscrit dans ce cadre. Il propose :

- Une mesure de l'importance de la région de résidence, de la tranche d'âge et du genre des bénéficiaires de soins sur le reste à charge sécurité sociale.
- Une mesure de l'impact du reste à charge après remboursement sécurité sociale et remboursement d'un organisme d'assurance maladie complémentaire sur la demande de soins des assurés du portefeuille considéré.
- Un argumentaire sur le projet "reste à charge zéro" en optique, dentaire et audioprothèse.
- Une analyse chronologique du reste à charge sécurité sociale par paire de verre optique.

Première partie

Le système d'assurance maladie

Introduction

Cette partie vise principalement le lecteur non spécialiste de la santé et de l'assurance maladie. Elle a pour but de lever un pan de voile sur le système d'assurance maladie. Le premier chapitre présente l'organisation en branche de la sécurité sociale et s'intéresse plus longuement à la branche maladie, aux risques qu'elle gère, à son financement et aux prestations qu'elle offre. Le second chapitre s'intéresse à l'assurance maladie complémentaire. Il donne au lecteur de connaître son principe, les types de complémentaire santé et les modes d'adhésion. Il fait également un état des lieux du marché de l'assurance maladie complémentaire. Le chapitre trois rend compte de la méthode courante de tarification des contrats d'assurance maladie complémentaire. Par quelques définitions, il présente les principaux risques couverts par les organismes d'assurance maladie complémentaire et il donne la méthode de détermination du prix d'un contrat. Cette partie se termine par le quatrième chapitre qui présente les principales provisions en assurance maladie et la méthode d'analyse de la rentabilité d'une police d'assurance maladie complémentaire.

Chapitre 1

La Sécurité Sociale et l'Assurance Maladie Obligatoire

Créée par les ordonnances du 4 et du 19 octobre 1945 du gouvernement du Général de Gaulle, la sécurité sociale a pour objectif d'établir une solidarité entre les différentes classes sociales afin de garantir à chacun une protection contre les incertitudes de la vie.

1.1 Généralité sur la sécurité sociale

Sous sa forme actuelle, la sécurité sociale s'organise en plusieurs régimes de base.

1.1.1 Les régimes de la sécurité sociale

L'organisation de la sécurité sociale est fortement marquée par la répartition socioprofessionnelle. Les régimes qui la structurent sont les suivants :

- **le régime général** qui concerne tous les salariés et les inactifs et qui est géré par la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS) ;
- **les régimes des exploitants agricoles**, gérés par la Caisse Centrale de la Mutualité Sociale Agricole (CCMSA) ;
- **le régime social des indépendants (RSI)** qui protège les artisans, les commerçants et les professions libérales. Depuis sa création en 2006, ce régime présente de nombreux dysfonctionnements. Un plan de redressement consistant à l'adosser au régime général a été voté par l'assemblée nationale : sur une période transitoire de deux ans à compter du 1^{er} janvier 2018, les indépendants seront progressivement intégrés au régime général.

A cette liste, il faut ajouter d'autres régimes singuliers, propres à certaines familles de métier comme ceux des mineurs, des militaires, de la SNCF, de la RATP et des industries électriques et gazières, de la banque de France, des parlementaires, pour ne citer que ceux-ci.

Le régime général couvre à lui seul plus de 80% des français. Avec les autres régimes, ils assurent la gestion des branches de la sécurité sociale.

1.1.2 Les branches de la sécurité sociale

Les régimes de la sécurité sociale sont organisés en branches séparées et autonomes gérées par des caisses nationales. La sécurité sociale est ainsi organisée en 5 branches.

La branche maladie

La branche maladie de la Sécurité sociale a pour mission de prendre en charge les dépenses de santé de ses affiliés malades et de garantir l'accès aux soins. Grâce à son action sociale, elle mène de nombreux programmes de prévention, permet aux plus démunis l'accès aux soins par la solidarité qu'elle instaure et participe à la gestion des établissements médico-sociaux. A ces missions de la branche maladie, vient s'ajouter le rôle qu'elle joue dans la gestion des risques suivants :

- **la maternité** en prenant en charge les frais d'examens pré et postnataux et en assurant les indemnités journalières durant le congé de maternité,
- **l'invalidité** en fournissant une pension dans des proportions déterminées à l'assuré présentant une capacité de travail ou de gain réduite (du fait de son invalidité),
- **le décès** par le paiement d'un capital correspondant à un multiple du revenu journalier de base l'assuré décédé.

La branche maladie de la sécurité sociale a également au cœur de ses missions l'amélioration du système de santé, de l'état de santé de la population et la maîtrise de l'évolution du coût des soins. Cette branche est représentée par les trois principaux régimes d'assurance maladie : le régime général (CNAMTS), le régime agricole (MSA) et le régime social des indépendants (RSI). Ils appartiennent tous les trois à l'Union nationale des caisses d'assurance maladie (Uncam) qui a pour but de fixer les taux de prise en charge des soins, de déterminer les actes de santé admis au remboursement et de préciser les liens entre l'assurance maladie et les professionnels de santé libéraux.

Les remboursements effectués par la branche maladie correspondent à deux types de prestations : prestations en nature et prestations en espèce.

Les frais de santé pour lesquels sont versées des prestations en nature correspondent aux :

- frais de médecine générale et spécialisée,
- frais de soins et de prothèses dentaires,
- frais pharmaceutiques et d'appareillage,
- frais d'analyses et d'examens de laboratoire,
- frais d'hospitalisation et de traitements lourds dans les établissements de soins, de réadaptation fonctionnelle et de rééducation ou d'éducation professionnelle,
- frais d'examen prénuptial,
- frais afférents aux vaccinations dont la liste est fixée par arrêté,
- frais relatifs aux examens de dépistage effectués dans le cadre de programmes de santé publique,
- frais d'hébergement et de traitement des enfants ou adolescents handicapés dans les établissements d'éducation spéciale et professionnelle,
- frais de transport des malades dans des conditions et limites tenant compte de l'état du malade et du coût du transport.

Les soins et les produits de santé doivent remplir deux critères pour être pris en charge par l'assurance maladie obligatoire. Ils doivent être réalisés par un professionnel de santé habilité à exercer et doivent obligatoirement être dispensés par un établissement public ou privé autorisé. S'agissant d'un médicament, il est recommandé que celui-ci figure dans la liste des médicaments et produits remboursables. Les actes de soins doivent quant à eux faire partie de la nomenclature

des actes de santé.

Une participation visant à responsabiliser l'assuré est très souvent laissée à sa charge, il s'agit du *ticket modérateur*. Cette participation est proportionnelle ou forfaitaire et dépend de la nature de la prestation. Il existe de nombreuses situations qui limitent ou suppriment le ticket modérateur. Les hospitalisations d'une durée supérieure à trente jours ou qui nécessitent des techniques lourdes et les médicaments qui sont reconnus irremplaçables et très coûteux sont quelques-unes des situations pour lesquels le ticket modérateur est inexistant. Pour certaines catégories d'assurés sociaux, une exonération du ticket modérateur est également prévue : les victimes de maladies professionnelles ou d'accidents de travail, les femmes enceintes à partir du sixième mois de grossesse, les bénéficiaires d'une pension d'invalidité, etc. L'assuré est tenu d'avancer les frais de santé et la Sécurité Sociale rembourse par la suite l'assuré. Toutefois, pour de nombreux actes et produits de santé, il existe des conventions ("tiers-payant") qui prévoient le paiement direct par la caisse au fournisseur de service.

Les prestations en espèces sont quant à elles fournies afin de compenser une perte de revenu pour un assuré qui doit cesser son activité professionnelle pour une raison de santé. Cette rémunération de remplacement correspond à une pension d'invalidité pour les personnes présentant une invalidité qui réduit leurs capacités de travail et de gain dans des proportions déterminées. Elle correspond également à des indemnités journalières en cas d'arrêt maladie, de congés de maternité ou paternité et à un capital en cas de décès.

Fortement marqué par des facteurs historiques et socio-économiques, le régime auquel est rattaché un assuré dépend de son activité professionnelle présente ou passée. Pour les personnes qui ne remplissent pas les conditions permettant leur rattachement à un régime sur la base de leur profession mais résidant en France de façon stable et en situation régulière, il existe la Protection Universelle Maladie (PUMa). Certains bénéficiaires de la PUMa sont redevables d'une cotisation annuelle dite « spécifique ». L'affiliation sur la base du critère de résidence est subsidiaire à l'affiliation au titre de l'activité professionnelle.

L'affiliation d'une personne donne également des droits aux prestations en nature de l'assurance maladie et maternité à son conjoint, concubin ou pacsé lorsque celui-ci ne bénéficie d'aucun régime de protection sociale. Sont aussi couverts les enfants à sa charge ou à la charge de son conjoint, concubin, pacsé (dans la limite d'un maximum) et toute personne étant à la charge effective et permanente de l'assuré et qui ne bénéficie pas d'un régime de protection sociale.

La branche accidents du travail et maladies professionnelles

La branche accidents du travail et maladies professionnelles (AT/MP) est en charge de la gestion des accidents liés aux activités des salariés. Elle est également en charge de la gestion des accidents de trajet et des maladies professionnelles. Elle assure la diffusion d'information sur les risques professionnels et met en œuvre une prévention dont le but est d'améliorer la sécurité et la santé des salariés en entreprise. Elle a également un rôle de formation, de conseil et de contrôle. Elle indemnise les victimes et détermine la participation de l'entreprise au financement du système d'assurance contre les dommages corporels. Composée à part égale des représentants des partenaires sociaux, employeurs et salariés, la Commission des accidents du travail et des maladies professionnelles (CAT/MP) de la Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS) est chargée de définir la politique de prévention et d'assurance des risques professionnels. Cette politique de prévention s'organise régionalement au sein de chaque Caisse régionale d'assurance maladie (CRAM), et, pour les DOM, au sein de chaque Caisse générale de sécurité sociale (CGSS). Des comités techniques nationaux et régionaux (CTN et CTR), composés eux aussi à parts égales de représentants des employeurs et des salariés, assistent les par-

tenaires sociaux pour la définition des actions de prévention dans les différents secteurs d'activité.

La branche retraite

Les prestations offertes par les régimes de retraite en France correspondent à la redistribution sous forme de pensions retraite au cours d'une année des cotisations perçues la même année auprès des actifs. Ce mode de fonctionnement est dit par **répartition**. L'ensemble des régimes de retraite obligatoires de base ou complémentaire coopèrent de façon coordonnée et sont solidaires les uns des autres. Cette solidarité permet de garantir une pension minimum (le minimum vieillesse) à toutes les personnes âgées à faibles ressources. La particularité des régimes de retraite obligatoire est d'intégrer un principe de solidarité entre les générations par le mécanisme de la redistribution. Ils favorisent aussi une solidarité au sein d'une même génération car ils redistribuent les cotisations collectées entre les différentes catégories socio-professionnelles et sans distinction de sexe. Le système de retraite repose sur quatre grands mécanismes :

- le taux de cotisation n'est pas fonction des écarts d'espérance de vie,
- les aléas de carrière sont pondérés dans les régimes de base surtout, avec l'attribution d'un minimum de pension et la prise en compte de périodes peu ou pas travaillées,
- le calcul des pensions intègre les avantages liés à la famille,
- tous les régimes attribuent, avec ou sans condition de ressources, des pensions de réversion au conjoint survivant.

Ce système s'organise en régimes de base obligatoire, en régime complémentaire souvent obligatoire et en régime facultatif. Le calcul des pensions diffère selon les régimes, on distingue le calcul en annuités et le calcul en points. Les régimes de base sont majoritairement des régimes en annuités, les annuités étant de 50% pour les affiliés au régime général, de 75% pour les affiliés au régime des fonctionnaires et forfaitairement fixés pour les assurés des professions libérales. La méthode de calcul en points consiste à attribuer une valeur d'achat unitaire à un point, le nombre de points étant déterminé sur la base du montant des cotisations versées en fonction du salaire de référence. Par cette méthode chaque assuré acquiert un certain nombre de points au cours de sa carrière. Les prestations de retraite sont alors fonction de ce nombre de points. La méthode en points est celle des régimes complémentaires de retraite.

La branche famille

La branche famille de la sécurité sociale est chargée de la gestion des prestations familiales. Par cette prestation, son action contribue à réduire les différences de niveau de vie entre les ménages selon le nombre d'enfants. Les actions menées par cette branche portent sur l'accompagnement des familles dans leur vie quotidienne, l'accueil du jeune enfant, l'accès au logement et la lutte contre la précarité et le handicap.

Ce mémoire étant consacré à l'assurance maladie complémentaire et collective, nous nous intéresserons uniquement à la branche maladie de la sécurité sociale.

1.2 L'Assurance Maladie Obligatoire (AMO)

L'AMO est gérée par la branche maladie de la sécurité sociale, elle garantit le minimum obligatoire d'assurance santé à toute personne exerçant une activité professionnelle.

1.2.1 Un financement multiple

La branche maladie de la sécurité sociale se finance essentiellement par les cotisations sociales qui sont constituées de la proportion de CSG (Contribution Sociale Généralisée) dédiée au risque maladie et des cotisations sociales patronales diminuées de la part dédiée au financement des indemnités journalières. Le reste du financement est assuré par la perception de taxes sur le tabac, l'alcool et des recettes annexes.

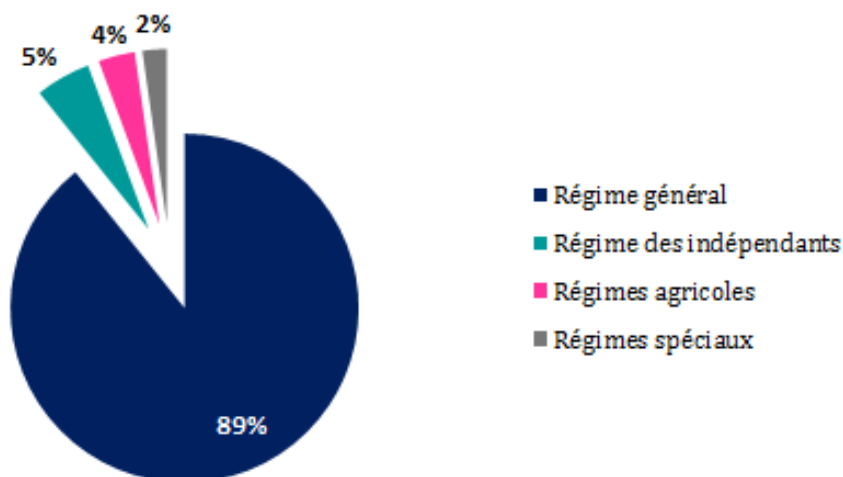


FIGURE 1.1 – Financement des régimes de base en 2015 (en pourcentage de l'assiette globale des recettes de l'AMO).

Le financement de l'assurance maladie obligatoire concerne tous les régimes de base de la sécurité sociale. Comme le présente la figure 1.1, les cotisants du régime général sont la principale source des recettes de l'AMO. Sans exclure les autres régimes de nos analyses, dans la suite de ce mémoire nous nous intéresserons principalement au régime général.

1.2.2 Des prestations diversifiées

L'assurance maladie obligatoire fournit des prestations diversifiées qui peuvent être observées suivant les différents axes de consommation des biens et services médicaux :

- Soins hospitaliers
- Soins de ville
- Médicaments
- Autres biens médicaux
- Transports de malades

En 2015, les soins hospitaliers représentaient 47% des dépenses de l'AMO (cf. figure 1.2), les soins de ville 26% et les autres axes de consommation des biens et services médicaux 27%. Par rapport à 2014, le volume de prestations est en hausse de 1,8%, après une augmentation de 2,7% de 2013 à 2014. De façon générale, les montants annuels des prestations de l'AMO sont (d'une année à l'autre) en hausse depuis 2008 malgré la perte de vitesse illustrée par la figure 1.3.

Cette diminution globale des taux d'accroissement annuel des dépenses de santé peut s'expliquer par une volonté de pallier au déficit chronique de la sécurité sociale et de responsabiliser les assurés. L'assurance maladie a instauré des forfaits journaliers et hospitaliers qui se rajoutent à la charge des ménages et dont le montant dépend de la prestation médicale dont bénéficie l'assuré.

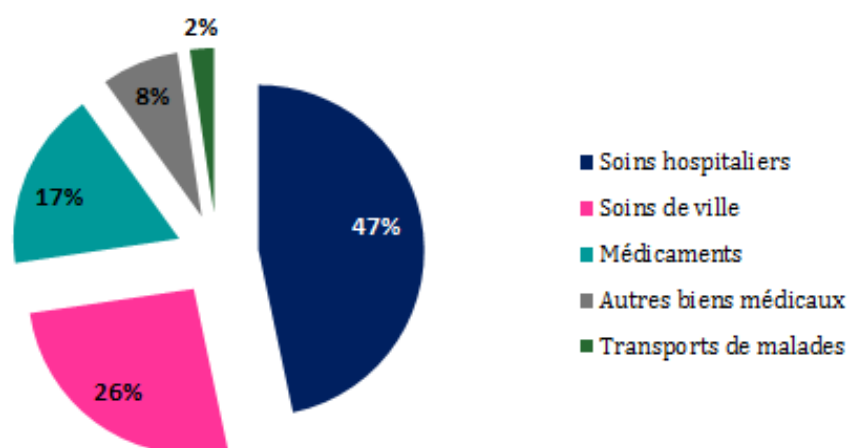


FIGURE 1.2 – Structure des dépenses de l'AMO (en pourcentage du volume globale de dépense de 2015).

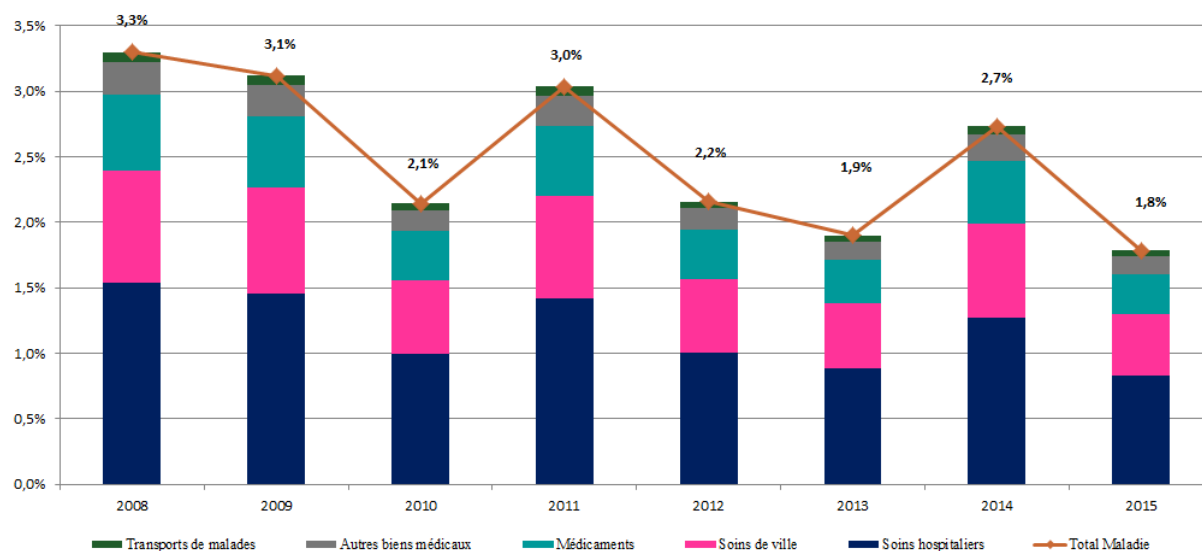


FIGURE 1.3 – Évolution du taux d'accroissement annuel des prestations de l'AMO de 2008 à 2015.

Les règles de remboursement

Les tarifs pratiqués par les médecins et le montant qui sert de base pour le calcul des remboursements de l'AMO sont fonction de la discipline du médecin et de son secteur d'activité :

- **Le Secteur 1** : Les médecins du secteur 1 sont ceux qui appliquent des tarifs fixés par convention avec l'AMO. Ces médecins ne pratiquent des dépassements d'honoraires qu'en cas de demande du patient, par exemple une demande de consultation en dehors des heures d'ouverture habituelles du cabinet médical. Ces dépassements ne sont pas remboursés par l'AMO.
- **Le Secteur 2** : Les médecins du secteur 2 sont ceux qui appliquent des tarifs libres, non fixés par l'AMO. Ces médecins sont ainsi autorisés à pratiquer des dépassements d'honoraires. Les dépassements ne sont pas remboursés par l'AMO et doivent être pratiqués avec tact et mesure.

Remarque : Les médecins conventionnés des secteurs 1 et 2 qui adhèrent à l'OPTAM (Option de Pratique Tarifaire Maîtrisée) s'engagent à modérer et stabiliser leurs honoraires afin de faciliter aux patients l'accès aux soins. Les actes de santé réalisés par ces médecins sont mieux remboursés que ceux faits par des médecins non conventionnés.

Chapitre 2

L'Assurance Maladie Complémentaire

En France comme dans beaucoup d'autres pays où il existe un système d'assurance maladie obligatoire, il existe également des mécanismes d'assurances privées. Ces assurances privées peuvent être :

- *duplicative*. L'assurance privée est dite duplicative lorsqu'elle permet aux assurés de l'AMO de bénéficier des prestations individuelles non prises en charge ou des prestations individuelles au sein du secteur public. Ce système existe en Grande-Bretagne, Espagne et Italie.
- *substitutive*. L'assurance privée est dite substitutive lorsqu'elle couvre des catégories particulières de population pour lesquelles elle remplace l'assurance publique. En Allemagne par exemple, au-delà d'un certain niveau de revenu, l'affiliation à l'assurance sociale n'est pas obligatoire.
- *supplémentaire*. L'assurance privée est dite supplémentaire lorsqu'elle prend en charge des prestations non couvertes par le système d'assurance sociale. Ce système existe par exemple au Canada et 67% des Canadiens y ont recours.
- *complémentaire*. L'assurance privée est dite complémentaire lorsqu'elle prend en charge une part des dépenses de santé laissées à la charge des assurés de l'AMO. C'est le principal système d'assurance privée en France. Il permet de compléter les remboursements de la sécurité sociale. En France, l'assurance complémentaire prend également en charge certaines dépenses de santé non couvertes par la sécurité sociale (les frais de chambre particulière par exemple).

A l'origine, les assurances complémentaires en France étaient essentiellement mutualistes et assuraient essentiellement la prise en charge du ticket modérateur. Suite aux évolutions de la réglementation et des comportements des consommateurs, l'intervention des organismes complémentaires s'est diversifiée et les mutuelles ont dû partager le marché avec les sociétés d'assurance et les institutions de prévoyance.

2.1 Principe de l'assurance maladie complémentaire

L'assurance maladie complémentaire a pour but de fournir des prestations en complément de celles de l'assurance maladie obligatoire, dans la limite des frais réels d'actes de soins. Ces prestations permettent de réduire voire de supprimer le reste à charge laissé aux ménages. De plus en plus, l'assurance maladie complémentaire intervient dans la prise en charge de prestations non

couvertes par la sécurité sociale. Avant le 1^{er} janvier 2018, les complémentaires santé pouvaient être "responsables" ou non. Le 31 décembre 2017 était le délai fixé pour mettre tous les contrats d'assurance maladie complémentaire en accord avec la réforme contrats responsables.

2.1.1 Les garanties du contrat responsable

Une complémentaire santé est dite responsable lorsqu'elle respecte un certain nombre de critères de prise en charge. Le contrat responsable doit couvrir la totalité du ticket modérateur : différence entre le tarif de référence (la base de remboursement) et le montant réellement remboursé par l'assurance maladie obligatoire (généralement exprimé en pourcentage de la base de remboursement) avant d'appliquer sur celui-ci une participation forfaitaire d'un euro ou une franchise. Cette participation forfaitaire ayant pour but de responsabiliser les assurés ne s'applique pas aux frais de cures thermales, de médicaments qui sont pris en charge par l'assurance maladie à 15% ou à 30% de la base de remboursement et aux médicaments et préparations homéopathiques.

Un contrat responsable prévoit également le remboursement du forfait journalier facturé par les établissements hospitaliers de santé, sans limitation de durée. Concernant la couverture des équipements optiques au-delà du ticket modérateur, la prise en charge des verres et de la monture est fonction du niveau de correction. En effet pour les verres simples, le minimum de remboursement est fixé à 50 euros et le maximum à 470 euros. Pour les verres très complexes, le minimum de remboursement est fixé à 200 euros et le maximum à 850 euros. La prise en charge de la monture est quant à elle plafonnée à 150 euros. L'adulte a droit à une prestation d'optique tous les deux ans sauf en cas d'une modification de la vue qui contraint à un renouvellement de l'équipement. Pour les mineurs la période est réduite à un an. Lorsque l'assuré ne peut ou ne souhaite pas un équipement progressif ou multifocal, la complémentaire santé peut prévoir la prise en charge d'un équipement de correction autre sur une période de 2 ans.

Les contrats responsables ne prennent pas en charge la participation forfaitaire (1 euro par consultation) et les franchises médicales laissées à la charge de l'assuré. Ils ne couvrent également pas la majoration de la participation de l'assuré due à la non désignation d'un médecin traitant ou consultation d'un autre médecin sans prescription du médecin traitant : l'assuré est alors dit « hors parcours de soins ». Ils ne prennent également pas en charge les dépassements d'honoraires lorsque l'assuré consulte un spécialiste auquel la réglementation ne permet pas d'accéder directement sans passer par son médecin traitant, il s'agit d'un spécialiste autre qu'un gynécologue, ophtalmologue, psychiatre, neuropsychiatre et stomatologue.

Remarque : La prise en charge des dépassements d'honoraires, lorsqu'elle est proposée par une complémentaire santé responsable, est limitée à 100% des frais réels pour les médecins (généralistes et spécialistes) ayant adhéré à l'OPTAM. Elle est limitée à 200% de la base de remboursement de l'assurance maladie obligatoire pour les médecins qui n'y sont pas adhérents. De plus, chaque année, les organismes qui proposent des complémentaires santé responsables doivent communiquer aux assurés le montant et la composition des frais de gestion et d'acquisition des contrats.

2.1.2 Les remboursements supplémentaires

Les complémentaires santé peuvent proposer des garanties autres que celles définies par le contrat responsable en prenant en charge : le forfait journalier hospitalier appliqué par les établissements médico-sociaux, des coûts supplémentaires pour la chambre particulière, des lentilles et la chirurgie réfractive, les coûts des prothèses dentaires et d'orthopédie dentofaciale qui sont en sus du ticket modérateur, des implants, les coûts d'orthopédie et de prothèses qui dépassent le ticket modérateur, des actes de prévention (les vaccins par exemple), des actes de médecine

douce non pris en charge par le régime obligatoire et un forfait pour les cures thermales. A ces garanties, certains assureurs rajoutent des prestations en cas de maternité, de naissance et un forfait obsèques.

De nombreux assureurs proposent à leurs assurés des réseaux de professionnels de santé. Principalement mis en place dans le domaine de l'optique, du dentaire et des audio-prothèses, ces réseaux permettent aux assurés de bénéficier de tarifs négociés et d'un engagement en terme de qualité des biens et services de santé. En se rendant chez un professionnel qui appartient à un réseau, un assuré peut voir le reste à charge disparaître pour certains actes de soins, selon les caractéristiques de son contrat, voire même disparaître totalement son « reste à charge » (c'est-à-dire la part des dépenses qui reste à sa charge après remboursement de l'assurance maladie obligatoire).

Il est de plus en plus courant que les complémentaires santé offrent des services singuliers adaptés à la demande des consommateurs. Ces services peuvent consister en des informations permettant d'orienter l'assuré dans le système de santé, des services d'analyse de devis, de coaching, de soutien psychologique et des services d'assistance comme le rapatriement, le soutien scolaire, la garde d'enfants, la garde d'animaux, l'aide ménagère et l'aide à la garde des malades.

2.1.3 Le tiers payant

Le tiers payant est le mécanisme consistant à dispenser les assurés de l'avance des frais de santé à leur charge (en totalité ou en partie). Ce mécanisme est mis en place par une convention signée entre l'assureur et des professionnels de santé. En pratique, il suffit que le patient présente une attestation de tiers payant délivrée par l'assureur pour être dispensé de l'avance des frais par le professionnel de santé. Pour les garanties minimales définies par le contrat responsable, tous les assurés adhérant à un contrat responsable peuvent également se voir proposer le tiers payant par leur médecin à hauteur du ticket modérateur. Aussi, les bénéficiaires de l'ACS (aide à l'acquisition d'une complémentaire santé) bénéficient pour la plupart du tiers payant intégral, c'est-à-dire qu'il n'avancent aucun frais de la part obligatoire ou complémentaire pour l'ensemble des soins de ville.

2.2 Les types de complémentaire santé

Couramment dénommé "complémentaire santé", les contrats d'assurance maladie complémentaire peuvent être distingués par l'obligation d'affiliation. On distingue des contrats à adhésion obligatoire et des contrats à adhésion facultative.

2.2.1 Les complémentaires santé collectives à adhésion obligatoire

Les complémentaires santé collectives sont des contrats couvrant un groupe de salariés. Dans une entreprise, cette couverture peut être faite par catégories objectives de personnes. Depuis le 1^{er} janvier 2016, les entreprises ont toutes l'obligation de couvrir leurs salariés par une complémentaire santé. A partir du 1^{er} janvier 2018, ces complémentaires devront obligatoirement être "responsables". Tous les salariés doivent être couverts et pour ceux appartenant au même contrat le tarif et la participation de l'employeur doivent être identiques c'est-à-dire sans distinction de sexe et d'âge. Le caractère obligatoire de ces contrats est dû au fait que les salariés doivent inmanquablement y adhérer, bien que, dans certains cas, le salarié peut, pour des raisons particulières renoncer à cette couverture. Lorsque le contrat le prévoit, les ayants droits du salariés peuvent eux aussi bénéficier de cette couverture. Toutes les spécificités des contrats collectifs obligatoires sont précisées par la loi du 14 juin 2013 ; cette loi a imposé la généralisation des complémentaires santé à tous les salariés du privé. Les contrats sont annuels. Les tarifs et

les garanties sont révisables. L'entreprise et l'assureur disposent d'un droit de résiliation sous conditions de respect des délais et préavis fixés.

Que se passe-t-il en cas de départ de l'entreprise ?

Le législateur a prévu que l'ex-salarié puisse continuer à bénéficier des garanties en vigueur dans les cas suivants :

- En cas de **licenciement** ou de **rupture conventionnelle**, le salarié qui se retrouve au chômage (sauf pour faute lourde) est bénéficiaire de la portabilité de ses droits. La portabilité est le maintien des garanties complémentaires, dont bénéficiait l'ex-salarié, de manière totalement gratuite et pour une période correspondante à la durée du dernier contrat de travail et plafonnée à 12 mois. À l'issue de cette période, par un contrat dit de sortie, le chômeur peut continuer de bénéficier de ces garanties en contrepartie d'une cotisation.
- En cas de fin du contrat du fait d'un **départ en retraite** ou du passage en **incapacité ou invalidité**, le salarié retraité ou en arrêt de travail peut être bénéficiaire d'un contrat de sortie du contrat d'assurance complémentaire collectif (loi du 31 décembre 1989). Ce contrat de sortie présente des garanties identiques à celles du précédent contrat. La première année, le coût de cette couverture est identique à celui payé par les salariés actifs, à la différence qu'il n'y a pas de contribution de la part de l'employeur. La totalité de la prime d'assurance est à la charge de l'ex-actif. Le coût de cette couverture ne peut excéder 25% la deuxième année et 50% la troisième année. Au-delà de cette troisième année, la prime à payer est laissée à la libre appréciation de l'assureur.

L'assureur a l'obligation de faire une proposition de contrat de sortie dans un délai de 2 mois à compter de la date de départ en retraite ou, le cas échéant à partir de l'expiration de la période de portabilité. En cas de sortie d'une complémentaire santé du fait du décès, les ayants droit du salarié décédé peuvent bénéficier (dans les mêmes conditions) d'un contrat de sortie. L'ex-salarié est libre de souscrire ou non à un autre contrat et même de changer d'assureur. L'adhésion au contrat de sortie est facultative.

2.2.2 Les complémentaires santé collectives à adhésion facultative

Encore appelés "contrat optionnel" ou "surcomplémentaire", les complémentaires santé collective à adhésion facultative ont pour but de donner au salarié la possibilité de bénéficier d'une couverture complémentaire plus complète que celle proposée par la complémentaire santé obligatoire. Ce contrat est souscrit par l'entreprise mais les salariés y adhèrent volontairement et individuellement. Lorsque le contrat le prévoit, l'adhérent à une surcomplémentaire peut en faire bénéficier ses ayants droit. Le financement du complément de couverture donné par le contrat optionnel est très souvent à la charge exclusive du salarié qui y souscrit. En cas de départ de l'entreprise, le souscripteur peut conserver cette couverture dans les mêmes conditions qui prévalent pour la complémentaire santé obligatoire.

2.2.3 Les complémentaires santé individuelles

Les contrats d'assurance maladie complémentaire peuvent être souscrits de façon individuelle par des personnes qui peuvent, si le contrat le permet, en faire bénéficier leurs ayants droit. Ces contrats à souscription libre sont principalement destinés aux étudiants : lorsqu'ils ne sont pas couverts par les complémentaires de leurs parents, aux indépendants bien qu'ils adhèrent généralement aux contrats Madelin (loi n° 94-126 du 11 février 1994), aux fonctionnaires, aux

chômeurs, aux retraités qui ne souhaitent pas souscrire au contrat de sortie qui leur est proposé par l'ancien assureur et aux assurés en collectif qui souhaitent compléter leur couverture par ailleurs par un contrat individuel.

Depuis le 1^{er} juillet 2015, les personnes qui bénéficient de l'aide au paiement de leur complémentaire santé ont accès à des contrats sélectionnés sur des critères de qualité et de prix. Ces personnes bénéficient d'une dispense d'avance de frais chez le médecin, leurs médicaments, lunettes, prothèses dentaires et auditives sont mieux remboursés que ceux des autres assurés sociaux et elles sont exonérées des franchises médicales. En cas de non paiement des cotisations ou de déclaration fausse, la complémentaire santé peut être résiliée, dans le cas contraire il est tacitement reconduit jusqu'au décès de l'assuré. Ces assurés ne peuvent être exclus par l'assureur qui ne peut non plus réduire les garanties ou hausser les tarifs du fait de l'état de santé. Néanmoins, l'assureur a le droit de réviser les tarifs de l'ensemble des affiliés à un contrat.

2.3 Le marché de l'assurance complémentaire

La part de l'assurance complémentaire dans le financement des dépenses d'assurance maladie est en moyenne de 13,5% du montant total de la couverture maladie depuis 2010. Ce taux relativement constant permet à plus 94% de la population de bénéficier d'une protection complémentaire contre la maladie. Ces 94% sont constitués des 88% des personnes couvertes par des organismes complémentaires et des 6% de bénéficiaires de la CMU-C. Le manque de moyens financiers est l'une des principales raisons pour lesquels 6% de la population reste sans couverture maladie complémentaire.

Le marché de l'assurance maladie complémentaire reste fortement dominé par les mutuelles même si elles perdent des parts de marché au profit des sociétés d'assurances (cf. figure 2.1).

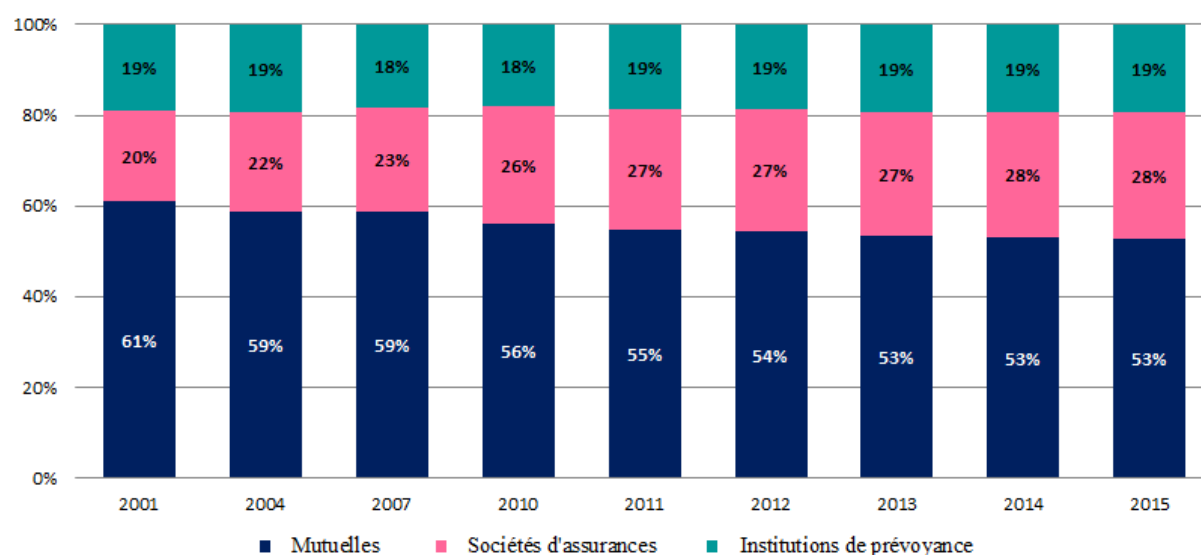


FIGURE 2.1 – Répartition des volumes annuels de cotisations par type d'organisme d'assurance maladie complémentaire de 2001 à 2015.

Suite au renforcement des exigences réglementaires en matière de solvabilité des organismes assureurs, le marché de la complémentaire santé s'est fortement concentré ces dernières années. De 2001 à 2014, le nombre d'organismes complémentaires est passé de 1702 à 573, une baisse 66% au global. Les mutuelles ont diminué de 70%, les institutions de prévoyance de 54% et les sociétés d'assurance de 20% (cf. figure 2.2).

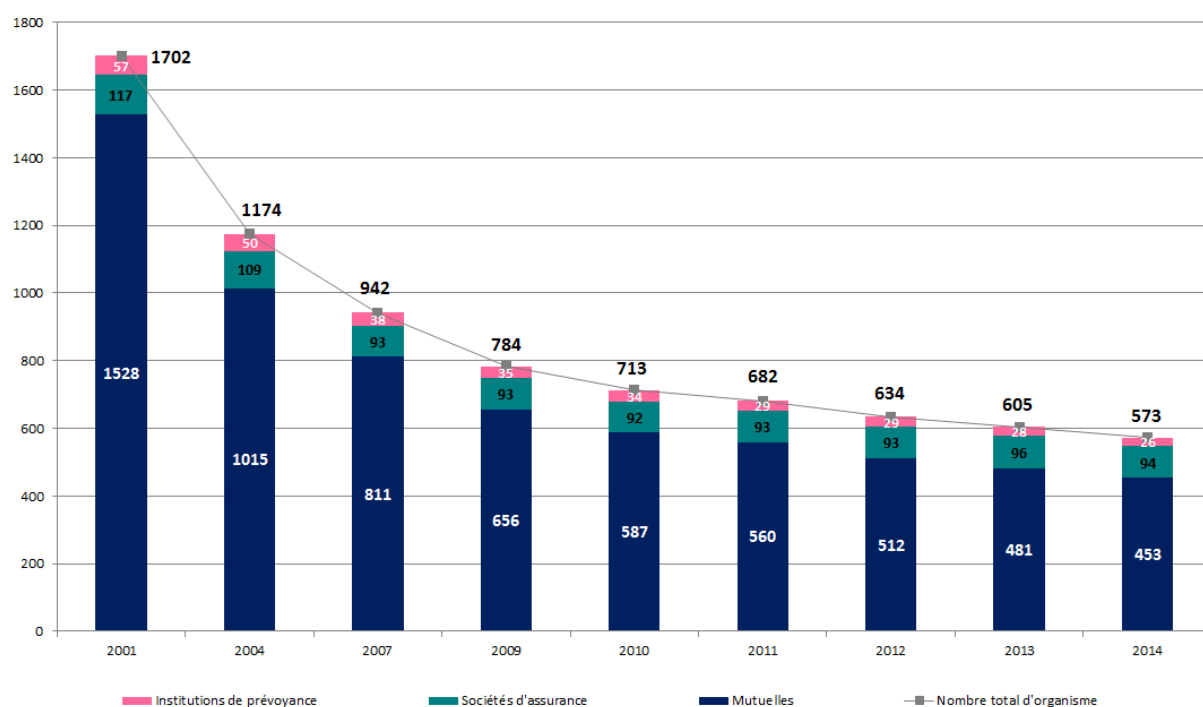


FIGURE 2.2 – Nombre d'organismes exerçant une activité de complémentaire santé de 2001 à 2014.

Ce recul du nombre d'assureurs complémentaires s'explique également par la concurrence qui existe sur le marché. Suite à des fusions et absorptions, on compte aujourd'hui 3 fois moins de mutuelles qu'en 2001 et deux fois moins d'institutions de prévoyance. Le nombre de sociétés d'assurance a, quant à lui, diminué dans une moindre mesure.

Chapitre 3

Tarification des contrats d'assurance maladie complémentaire

En France, nous l'avons vu, un peu plus de 94% de la population bénéficie d'une complémentaire. Qu'ils soient couverts par une société d'assurance, une mutuelle ou une institution de prévoyance, les souscripteurs aux complémentaires santé payent une prime qui est la contrepartie de la protection dont ils bénéficient. Cette prime est le résultat d'une analyse statistique des facteurs déterminants de la consommation des biens et services médicaux et du niveau de garantie souhaité. Dans ce chapitre nous proposons une présentation de quelques méthodes de tarification et leur mise en œuvre.

3.1 Quelques définitions

Classiquement, les actes et services médicaux garantis par un contrat d'assurance maladie complémentaire peuvent être regroupés par catégories : *hospitalisation, consultation et actes médicaux courants, optique, dentaire et pharmacie*.

3.1.1 L'hospitalisation

Le risque **hospitalisation** est principalement représenté par les dépenses de santé relatives aux frais de séjour, aux honoraires médicaux et chirurgicaux (en secteur conventionné ou non), aux frais d'hébergement et d'entretien du malade générés par l'hospitalisation (le forfait hospitalier), au surcoût pratiqué par les établissements de santé lorsque l'assuré souhaite disposer d'une chambre individuelle durant le séjour à l'hôpital (la chambre particulière) et au coût supplémentaire dû à la mise à disposition d'un lit pour une personne souhaitant accompagner le malade (le lit d'accompagnement).

Pour un assureur complémentaire, l'hospitalisation n'est pas le risque le plus important en matière de prestations. La sécurité sociale couvre l'essentiel des frais. L'assurance maladie complémentaire prend principalement en charge le forfait journalier et la chambre particulière (cf. table 3.1).

Remarque sur les tableaux 3.1, 3.2, 3.3, 3.4 et 3.5 : Le régime Alsace-Moselle est un régime d'Assurance maladie venant en complément du régime général français. Il est destiné aux habitants des départements du Haut-Rhin (68), du Bas-Rhin (67) et de la Moselle (57). Il est réglementé par le décret du 31 mars 1995, issu de la loi du 31 décembre 1991, le définissant comme un régime dérogatoire, complémentaire et obligatoire du régime général.

Liste des garanties	Régime Général	Régime Alsace-Moselle
Forfait Journalier	0%	0%
Chambre Particulière	0%	0%
Honoraires Chirurgicaux	80%	100%
Honoraires Médicaux	80%	100%
Frais de Séjour	80%	100%
Lit d'Accompagnant	0%	0%

TABLE 3.1 – Prestations de la sécurité sociale (en % de la base de remboursement) pour la catégorie hospitalisation.

3.1.2 Consultation et actes médicaux courants

Le risque **consultation et actes médicaux courants** est principalement représenté par les dépenses de santé relatives aux tarifs des consultations pratiquées par les médecins généralistes et spécialistes, signataires ou non de l'OPTAM. A ces dépenses, il faut rajouter les frais générés par des actes de laboratoire, de radiologie, de "petite chirurgie" et les services rendus par les auxiliaires médicaux. Le tableau 3.2 donne les taux de remboursement pratiqués par la sécurité sociale pour cette catégorie.

Liste des garanties	Régime Général	Régime Alsace-Moselle
Consultations / Visites Généralistes	70%	90%
Consultations / Visites Spécialistes	70%	90%
Acte de radiologie / imagerie médicale	70%	90%
Actes de laboratoire	60%	90%
Auxiliaires Médicaux	60%	90%
Petite Chirurgie / Acte Technique	70%	90%

TABLE 3.2 – Prestations de la sécurité sociale (en % de la base de remboursement) pour la catégorie consultation et actes médicaux courants.

3.1.3 L'optique

Le risque **optique** est relatif aux dépenses générées par l'acquisition des équipements optiques : les verres, les montures et les lentilles. Pour cette catégorie, de nombreux contrats prévoient également le couvreur de la chirurgie réfractive, qui n'est pas remboursée par la sécurité sociale (cf. table 3.3).

Remarque : Les dépenses liées aux verres dépendent de la correction visuelle dont a besoin l'assuré. On distingue ainsi des verres simples, complexes et hypercomplexes. Les verres simples sont des équipements optique unifocaux à sphère ou cylindre positif ou négatif. La sphère permet de déterminer le degré d'astigmatisme et le cylindre permet de déterminer le degré de myopie (cylindre négatif) et d'hypermétropie (cylindre positif). Les verres complexes sont des équipements unifocaux donc la sphère est en absolue supérieure à 6 dioptries et/ou le cylindre en absolu

Liste des garanties	Régime Général	Régime Alsace-Moselle
Verre Adulte	60%	90%
Verre Enfant	60%	90%
Monture Adulte	60%	90%
Monture Enfant	60%	90%
Lentilles	60%	90%
Chirurgie réfractive (myopie laser)	0%	0%

TABLE 3.3 – Prestations de la sécurité sociale (en % de la base de remboursement) pour la catégorie optique.

supérieur à 4 dioptries. Il peut aussi s’agir de verres multifocaux à sphère en absolue inférieure à 4 dioptrie et sans cylindre. Les verres hypercomplexes sont quant à eux des verres multifocaux ou progressifs dont la sphère est en absolue supérieure à 4 dioptries et sans cylindre, ou en absolue inférieure à 8 dioptries et avec cylindre, ou en absolue supérieure à 8 dioptries et avec cylindre.

3.1.4 Le dentaire

Le risque **dentaire** est relatif aux les dépenses de santé générées à l’occasion de soins dentaires, d’actes prothétiques, orthodontiques, implantaires et parodontologiques. Pour cette catégorie, il recommande à l’assureur de ne prendre en charge que les soins ayant fait l’objet d’un remboursement de la sécurité sociale. De nombreux contrats complémentaires prévoient toutefois la couverture de certains actes dentaires non couverts par la sécurité sociale (cf. table 3.4).

Liste des garanties	Régime Général	Régime Alsace-Moselle
Consultations et Soins Dentaires	70%	90%
Orthodontie	100%	100%
Prothèse Dentaire	70%	90%
Inlay-Core	70%	90%
Inlay / Onlay	70%	90%
Implants Dentaires	0%	0%
Parodontologie	0%	0%

TABLE 3.4 – Prestations de la sécurité sociale (en % de la base de remboursement) pour la catégorie dentaire.

3.1.5 Les autres actes médicaux

Cette catégorie est constituée de l’ensemble des actes et services de santé réalisés par des professionnels qualifiés autres que sus-cités. Il s’agit notamment de la prise en charge : des actes dits de médecine douce (ostéopathie, chiropractie, acupuncture, étio-pathie, pédicurie-podologie, diététique, psychologie, psychomotricité et tabacologie), de l’orthopédie, des audio-prothèses, de l’appareillage, de la cure thermique et du forfait naissance et d’un nombre grandissant d’actes

de prévention. Pour ces quelques garanties, le tableau 3.5 donne le niveau des prestations de la sécurité sociale.

Liste des garanties	Régime Général	Régime Alsace-Moselle
Médecines douces	0%	0%
Appareillage	60%	90%
Orthopédie	60%	90%
Prothèse Auditive	60%	90%
Cure Thermale	70%	90%
Forfait Naissance	0%	0%

TABLE 3.5 – Prestations de la sécurité sociale (en % de la base de remboursement) pour la catégorie autres actes médicaux.

3.1.6 Principes de remboursement

Les prestations de l'assurance maladie complémentaire ont vocation à couvrir les dépenses de soins et services médicaux non pris en charge par la sécurité sociale. Le montant du **remboursement de la sécurité sociale**(RSS) est déterminé (pour un acte) en appliquant le taux de remboursement correspondant à un tarif de convention. Ce tarif de convention est dénommé **base de remboursement**(BR) ou **Tarif de responsabilité**(TR). La différence entre la base de remboursement et le remboursement réalisé par la sécurité sociale est le **ticket modérateur** (TM). Lorsque les **frais réels**(FR) ou dépenses totales d'un acte ou service de santé vont au-delà de la base de remboursement, le montant en sus correspond au **dépassement d'honoraire**(Dep). la figure 4.1 représente les relations qui existent entre ces différentes quantités.

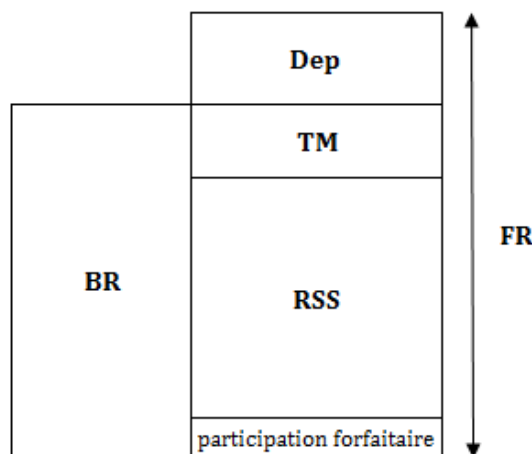


FIGURE 3.1 – Illustration du mécanisme de remboursement de la sécurité sociale

En résumé :

$$RSS = x \times BR; \quad TM = BR - RSS; \quad Dep = FR - BR$$

où x est le taux de remboursement.

Remarque : Les remboursements des régimes de base ne prennent généralement pas en charge le montant total de la dépense de santé. L'article 52 de la loi de financement de la sécurité sociale pour 2008 a instauré dans le mécanisme de remboursement des franchises médicales. Ces franchises ont pour objectif de réduire les remboursements effectués par les caisses d'assurance maladie et de responsabiliser les patients. Cette franchise est de 0,50 euro pour le remboursement des médicaments. **La franchise médicale est différente de la participation forfaitaire de 1 euro** demandée aux assurés sociaux âgés de plus de 18 ans lorsqu'ils bénéficient de consultations ou d'actes réalisés par des professionnels de santé et d'examen de radiologie et d'analyses de biologie médicale. Elle se distingue aussi du **forfait hospitalier**, facturé à 20 euros par jour en hôpital ou en clinique et à 15 euros par jour dans le service psychiatrique d'un établissement de santé, et du forfait de 18 euros prévu pour tout acte médical d'un coût supérieur à 120 euros.

3.2 Principe général de la tarification

Comme dans tous les autres domaines de l'assurance, en assurance maladie complémentaire, les actuaires sont confrontés à l'estimation des coûts qui pourraient être générés par les risques couverts. Pour quantifier, analyser et comprendre de façon adéquate l'engagement qui sera pris par l'assureur au moment de la souscription, l'actuaire doit construire un modèle simple, facilement utilisable mais assez élaboré afin de prendre en compte la complexité du risque.

3.2.1 La prime pure

Considérons une variable aléatoire X représentant le coût (ou la perte) pour l'assureur associée à un contrat d'assurance maladie complémentaire d'un quelconque portefeuille d'assuré. La représentation théorique de cette variable est donnée par :

$$X = \sum_{k=1}^N Z_k 1_{N>0}$$

où N est une variable représentant le nombre d'actes et services médicaux associés à la garantie considérée et dont ont bénéficié les affiliés et Z_k est la variable de montant correspondant au k -ième acte ou service médical. Les Z_k sont supposées indépendantes et identiquement distribuées suivant une loi L . Le coût espéré du risque est

$$pp = E(L) \cdot E(N)$$

En effet, Le coût espéré du risque est donné par $pp = E(X)$. En conditionnant X par N on a :

$$E(X) = E(E(X|N))$$

or

$$E(X|N = n) = E\left(\sum_{k=1}^n Z_k\right) = n \times E(L), \quad n \in N^+$$

d'où

$$E(X) = E(N \cdot E(L)) = E(N) \cdot E(L)$$

donc

$$pp = E(L) \cdot E(N)$$

D'après cette équation, la prime pure est le produit du montant et du nombre espéré des actes et services médicaux associés à la garanties. Il s'agit d'une tarification par la méthode "fréquence-coût moyen".

Ainsi évalué, si on note i la garantie considérée, et pp_i la prime pure associée à la garantie i , alors la prime pure associée à l'ensemble des garanties contractuelles est :

$$PP = \sum_i pp_i$$

Si T représente la taille du portefeuille d'assurés ayant souscrit au contrat considéré, la prime par tête à payer est :

$$\frac{PP}{T}$$

3.2.2 Méthodologie du BCAC

La tarification par l'approche fréquence-coût présentée dans la sous-section précédente est celle utilisée par le BCAC (Bureau Commun des Assurances Collective). Le montant de la **prime** résulte de l'application des **fréquences** observées par type d'acte au **coût moyen** étant à la charge de l'assureur. Le tableau 3.6 donne la codification des actes médicaux retenus par le BCAC.

Code	Acte
C	Consultations (omnipraticien)
CS	Consultations de spécialiste
V	Visites (omnipraticien)
VS	Visites de Spécialiste
K	Actes de spécialité
KC	Actes de chirurgie
Z	Actes d'électroradiologie
B	Frais d'analyses et d'examens de laboratoire
AM	Auxiliaires médicaux
AMI	Infirmiers
AMM	Kinésithérapeutes
AMO	Orthophoniste
AMY	Orthoptiste
D	Soins dentaires

TABLE 3.6 – Codification des actes médicaux retenus par le BCAC

Dans son approche, le BCAC définit un **assuré-type** à partir des statistiques du régime général. Il s'agit d'une personne âgée de 40 ans et d'un ayant droit pouvant être dans 75% des cas un enfant et dans 25% des cas le conjoint ou concubin. Afin d'adapter le montant de la prime

aux différents cas de composition familiale, le BCAC a également défini la **famille type**. Elle est composée de 37% d'hommes, 38% de femmes et 25% d'enfants.

Remarque : La majorité des organismes d'assurance maladie complémentaire collective propose des tarifs issus de méthodes dérivées de celle du BCAC. Elles sont basées sur la définition du **salarié moyen** du groupe à assurer : il s'agit d'une personne étant dans H% des cas un homme et dans F% une femme. Cette personne est soit :

- Célibataire, Veuf ou Divorcé et sans enfant à charge (CVD0)
- Célibataire, Veuf ou Divorcé et avec Enfant(s) à charge (CVDE)
- Marié (ou pacsé ou en concubinage) et sans enfant à charge (M0)
- Marié (ou pacsé ou en concubinage) et avec Enfant(s) à charge (ME)
- et elle a Enf% d'enfants.

Ces caractéristiques du salarié moyen à assurer sont telles que

$$H\% + F\% = 100\%$$

et

$$CVD0\% + CVDE\% + M0\% + ME\% = 100\%$$

Le BCAC détermine le coût associé à un acte i par la formule

$$Cout_i = BR_i \times TM_i \times Ptm_i \times Dep_i$$

où BR est la base de remboursement de l'AMO, TM le ticket modérateur (en % de la BR), Ptm la proportion d'actes pour lesquels le ticket modérateur est maintenu et Dep la proportion des éventuels dépassements.

La prime pure de l'acte i est donc

$$pp_i^{BCAC} = freq_i \times cout_i$$

e.i

$$pp_i^{BCAC} = freq_i \times BR_i \times TM_i \times Ptm_i \times Dep_i$$

où $freq_i$ est la fréquence de l'acte i

La prime pure est donc

$$PP^{BCAC} = \sum_i pp_i^{BCAC}$$

3.2.3 La prime commerciale

La prime pure d'un contrat d'assurance maladie complémentaire est la valeur probable des engagements de l'assureur. Elle correspond au coût du risque à couvrir. Les contrats collectifs en particulier les complémentaires santé créent une mutualisation des risques entre les assurés, mais cela ne suffit pas pour garantir la solvabilité de l'organisme d'assurance maladie complémentaire.

Le principe de la mutualisation se fonde sur la loi des grands nombres : étant donnée une suite de variables aléatoire $(X_n)_n$ indépendantes, identiquement distribuées et intégrable,

$$\frac{1}{n} \sum_{i=1}^n X_i \longrightarrow E(X)$$

En assurance maladie complémentaire les salariés appartenant à la même catégorie objective ont une consommation de biens et services médicaux supposée identique et indépendamment distribuée. La loi des grands nombres, et donc le principe de mutualisation, garantit que lorsque le nombre de salariés couverts par le contrat augmente, la probabilité que le coût moyen par tête des actes et services médicaux garantis par la complémentaire s'écarte marginalement de la prime pure tend vers zéro.

En effet, d'après l'inégalité de Bienaymé-Tchebychev on a :

$$P(|\bar{X}_n - E(\bar{X}_n)| \geq \alpha) \leq \frac{E((\bar{X}_n - E(\bar{X}_n))^2)}{\alpha^2} = \frac{\sigma_n^2}{\alpha^2}$$

avec $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$; $\alpha > 0$ et $\sigma_n^2 = E((\bar{X}_n - E(\bar{X}_n))^2)$

e.i $\sigma_n^2 = \text{var}(\bar{X}_n) = \text{var}(\frac{1}{n} \sum_{i=1}^n X_i) = \frac{1}{n^2} \text{var}(\sum_{i=1}^n X_i) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{1}{n^2} \times n \times \text{var}(X)$
En notant $\sigma^2 = \text{var}(X)$ on a

$$P(|\bar{X}_n - E(X_n)| \geq \alpha) \leq \frac{\sigma^2}{n\alpha^2}$$

d'où

$$\lim P(|\bar{X}_n - E(X_n)| \geq \alpha) = 0$$

Ainsi la mutualisation créée par le contrat collectif nous garantit que le coût moyen par assuré converge en probabilité vers la prime pure par tête. Toutefois $\lim P(|\bar{X}_n - E(X_n)| \geq \alpha) = 0$ n'implique pas $\lim P(|\sum_{i=1}^n X_i - nE(X)| \geq \alpha) = 0$ car $\text{var}(\bar{X}_n) < \text{var}(X)$.

La mutualisation créée par la complémentaire santé collective ne garantit donc pas la convergence du coût total vers la prime pure globale. Pour des raisons de solvabilité, il est donc nécessaire que la prime demandée par l'assureur comprenne une marge de sécurité. Afin d'apprécier l'importance de la marge de sécurité, intéressons nous à la probabilité de ruine de l'assureur qui demande une prime $\pi = E(X) + \epsilon$ à chaque salarié, ϵ étant le chargement de sécurité. Pour un nombre n de salariés cette probabilité est donnée par

$$\varphi_n(u) = P(\sum_{i=1}^n X_i > n\pi + u)$$

où $u > 0$ est la réserve initiale de l'assureur.

Si $\epsilon > 0$

$$\varphi_n(u) = P(\sum_{i=1}^n X_i > n\pi + u) = P(\sum_{i=1}^n X_i - nE(X) > n\epsilon + u) \leq P(|\sum_{i=1}^n X_i - nE(X)| > n\epsilon + u)$$

e.i

$$\varphi_n(u) \leq P(|\sum_{i=1}^n X_i - nE(X)| > n\epsilon + u) \leq \frac{\text{var}(\sum_{i=1}^n X_i)}{(n\epsilon + u)^2} = \frac{n \text{var}(X)}{(n\epsilon + u)^2} = \frac{n\sigma^2}{(n\epsilon + u)^2}$$

d'où

$$\lim \varphi_n(u) = 0$$

Si $\epsilon < 0$

Considérons un nombre n_0 de salariés tel que $\forall n > n_0, n\epsilon + u < 0$

$$\varphi_n(u) = P\left(\sum_{i=1}^n X_i > n\pi + u\right) = P\left(\sum_{i=1}^n X_i - nE(X) > n\epsilon + u\right) = 1 - P\left(\sum_{i=1}^n X_i - nE(X) \leq n\epsilon + u\right)$$

or

$$P\left(\sum_{i=1}^n X_i - nE(X) \leq n\epsilon + u\right) \leq P\left(\left|\sum_{i=1}^n X_i - nE(X)\right| \geq n\epsilon + u\right) \leq \frac{n\sigma^2}{(n\epsilon + u)^2}$$

d'où

$$\varphi_n(u) \geq 1 - \frac{n\sigma^2}{(n\epsilon + u)^2}$$

donc

$$\lim \varphi_n(u) = 1$$

Si $\epsilon = 0$

$$\varphi_n(u) = P\left(\sum_{i=1}^n X_i > n\pi + u\right) = P\left(\sum_{i=1}^n X_i > nE(X) + u\right) = P\left(\frac{\sum_{i=1}^n X_i - nE(X)}{\sigma\sqrt{n}} > \frac{u}{\sigma\sqrt{n}}\right)$$

d'après le théorème central limite

$$\frac{\sum_{i=1}^n X_i - nE(X)}{\sigma\sqrt{n}} \xrightarrow{L} N(0, 1)$$

or $\lim \frac{u}{\sigma\sqrt{n}} = 0$

donc

$$\varphi_n(u) \xrightarrow{n} P(\zeta > 0) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{2}$$

où $\zeta \stackrel{L}{=} N(0, 1)$

En résumé, nous avons les trois résultats suivants :

- si $\epsilon > 0$ c'est-à-dire $\pi > E(X)$, $\lim \varphi_n(u) = 0$
- si $\epsilon < 0$ c'est-à-dire $\pi < E(X)$, $\lim \varphi_n(u) = 1$
- si $\epsilon = 0$ c'est-à-dire $\pi = E(X)$, $\lim \varphi_n(u) = \frac{1}{2}$

En d'autres termes, la solvabilité de l'assureur n'est certaine que si la prime pure est augmentée d'une marge de sécurité strictement positive.

Afin d'obtenir la prime commerciale qui sera contractuellement proposée, l'assureur augmente la prime sécurisée ($\pi = E(X) + \epsilon$) de quelques chargements techniques. Il s'agit des chargements de gestion et d'acquisition, auxquelles il faut rajouter la TSCA (taxe spéciale sur les conventions d'assurances qui est de 7% pour les contrats d'assurance maladie responsables et de 14% pour les contrats d'assurance maladie non responsables) et une taxe de solidarité additionnelle destinée au fond CMU (6,27%). La prime commerciale est donnée par :

$$PC = (E(X) + \epsilon) \times \frac{1 + TSCA + CMU}{1 - \text{frais gestion} - \text{frais acquisition}}; \quad \epsilon > 0$$

Chapitre 4

Provisionnement et rentabilité d'un contrat assurance maladie complémentaire

Le chapitre précédent a mis en évidence le fait que l'espérance mathématique est la principale mesure de risque en matière d'assurance santé. Elle permet de quantifier le risque à couvrir et de déterminer la prime d'assurance correspondante. Elle est donc le premier "outil" de gestion en assurance santé car une prime bien estimée garantit, toutes choses égales par ailleurs, la solvabilité de l'assureur qui a l'obligation de payer les prestations lorsqu'un assuré bénéficie d'un acte ou service de santé contractuellement défini. A la date d'inventaire, l'assureur comptabilise les sinistres et les rapporte à la prime pure qu'il a perçu afin d'apprécier le rentabilité du contrat. Le ratio ainsi déterminé est couramment noté S/P (S pour *sinistres* et P pour *prime pure*) ou P/C (P pour *prestations* et C pour *cotisation pure*). Afin de déterminer la juste valeur des sinistres, l'assureur ajoute aux prestations fournies le montant estimé des prestations futures (il s'agit des provisions). Ainsi, on a :

$$\text{Sinistres} = \text{Prestations} + \text{Provisions}$$

4.1 Les Provisions en assurance maladie

Les risques couverts par les complémentaires santé étant par nature à développement court, les problèmes liés au provisionnement sont moins importants en matière d'assurance maladie complémentaire que dans d'autres domaines d'assurance. L'évaluation des provisions est toutefois importante en assurance maladie car les soins réalisés proches de la date d'inventaire ne sont généralement pas connus par l'assureur. De nombreuses autres raisons peuvent également expliquer le fait que les services de gestion de l'assureur n'ont pas connaissance des soins dont ont bénéficié les assurés, notamment les retards de traitement de la sécurité sociale et le dysfonctionnement du dispositif de télétransmission. Les deux principales provisions effectuées en assurance maladie complémentaire sont : *les provisions pour sinistres non connus et les provisions pour risque croissant*.

4.1.1 Provisionnement des sinistres non connus

Le provisionnement des sinistres non connus consiste en l'évaluation de la PSAP (provision pour sinistres à payer). En assurance santé, les sinistres survenus mais non connus en fin d'exercice sont pour la plupart enregistrés au cours des deux années suivantes. Du fait de cette particularité, le calcul des cadences de règlements se fait par mois de survenance et non par année

de survénance : il s'agit de la méthode de *CHAIN-LADDER*.

Soit $X_{i,j}$ le montant des prestations liées au mois i et observés (ou réglés) le mois j . En date de provisionnement, l'assureur santé a donc connaissance des sinistres $X_{i,j}$ donnés par le tableau 4.1.

	$j = 1$	$j = 2$	\cdots	$j = n$
$i = 1$	$X_{1,1}$	$X_{1,2}$	\cdots	$X_{1,n}$
\vdots	\vdots	\vdots	\ddots	
$i = n - 1$	$X_{n-1,1}$	$X_{n-1,2}$		
$i = n$	$X_{n,1}$			

TABLE 4.1 – Triangle de liquidation

Cette méthode de provisionnement est la plus simple à réaliser. Les hypothèses de la méthode sont les suivantes :

- L'absence de facteurs exogènes tels que l'inflation.
- Indépendance des mois de survénance, $i \cdot e \forall i_1 \neq i_2 (X_{i_1,j})_{j \in [1:n]} \perp (X_{i_2,j})_{j \in [1:n]}$.
- Invariance de la politique de règlement des sinistres par l'assureur.
- Invariance de la cadence de règlement dans le temps, $i \cdot e \forall i, j$

$$E(X_{i,j+1} | X_{i,1}, \dots, X_{i,j}) = X_{i,j} \times \text{constante}$$

Partant du triangle de prestations 4.1, la première étape de la méthode consiste à calculer les quantités $C_{i,j}$ qui sont des sommes cumulées des règlements antérieurs ou égaux à j pour chaque survénance i :

$$C_{i,j} = \sum_{k=1}^j X_{i,k} = C_{i,j-1} + X_{i,j}$$

	$j = 1$	$j = 2$	\cdots	$j = n$
$i = 1$	$C_{1,1}$	$C_{1,2}$	\cdots	$C_{1,n}$
\vdots	\vdots	\vdots	\ddots	
$i = n - 1$	$C_{n-1,1}$	$C_{n-1,2}$		
$i = p$	$C_{n,1}$			

TABLE 4.2 – Triangle des prestations cumulées

Partant du triangle 4.2 la deuxième étape de cette méthode est l'estimation des coefficients f_j de passage d'un mois de développement à l'autre. Ces estimateurs doivent présenter une faible volatilité pour que l'hypothèse d'invariance de la cadence de règlement dans le temps soit vérifiée.

Pour tout $j \in [1 : n]$, la méthode de Chain-Ladder propose l'estimateur

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$$

La troisième étape de cette méthode consiste à déterminer la charge ultime par exercice de survenance. Il s'agit d'estimer les quantités $\hat{C}_{i,n}$, $\forall i \in [1 : n]$ données par :

$$\hat{C}_{i,n} = C_{i,n-i+1} \prod_{j=n-i+1}^{n-1} \hat{f}_j, \forall i \in [1 : n]$$

Pour chaque exercice de survenance, la provision est donnée par :

$$\hat{R}_i = \hat{C}_{i,n} - C_{i,n-i+1} = \left(\prod_{j=n-i+1}^{n-1} \hat{f}_j - 1 \right) \times C_{i,n-i+1}, \forall i \in [1 : n]$$

Le montant total des provisions est donc :

$$\hat{R} = \sum_{i=1}^n \hat{R}_i$$

4.1.2 Mise en œuvre de la méthode

Considérons un contrat d'assurance santé pour lequel le triangle de liquidation donné par le tableau 4.3. Nous supposons que les garanties proposées par ce contrat restent les mêmes dans le temps.

	Règlement							
Survenance	1	2	3	4	5	6	7	8
Mai	44 032	53 278	26 119	16 232	16 234	905	1 547	290
Juin	45 471	66 155	15 504	11 741	11 363	3 862	1 750	
Juillet	47 150	79 501	58 774	20 235	23 062	4 524		
Août	32 500	48 898	18 402	43 510	2 947			
Septembre	48 524	81 980	15 177	14 179				
Octobre	42 017	33 490	9 745					
Novembre	40 125	63 230						
Décembre	48 000							

TABLE 4.3 – Triangle de liquidation

Partant du tableau 4.3, on obtient le triangle des prestations cumulées donné par le tableau 4.4.

A partir du triangle des prestations cumulées (tableau 4.4), nous obtenons les estimateurs de Chain-Ladder données par le tableau 4.5.

Règlement								
Survenance	1	2	3	4	5	6	7	8
Mai	44 032	97 310	123 429	139 661	155 895	156 800	158 347	158 637
Juin	45 471	111 626	127 130	138 871	150 234	154 096	155 846	
Juillet	47 150	126 651	185 425	205 660	228 722	233 246		
Août	32 500	81 398	99 800	143 310	146 257			
Septembre	48 524	130 504	145 681	159 860				
Octobre	42 017	75 507	85 252					
Novembre	40 125	103 355						
Décembre	48 000							

TABLE 4.4 – Triangle des prestations cumulées

\hat{f}_1	\hat{f}_2	\hat{f}_3	\hat{f}_4	\hat{f}_5	\hat{f}_6	\hat{f}_7
2,42	1,23	1,16	1,09	1,02	1,01	1,00

TABLE 4.5 – Estimateurs de Chain-Ladder.

Grâce au tableau des estimateurs de Chain-Ladder et au triangle des règlements cumulés, le montant total de la provision au 31/12 est estimé à 249 155,03.

i	\hat{R}_i
Mai	0,00
Juin	285,42
Juillet	2 905,24
Août	4 394,03
Septembre	18 869,46
Octobre	24 874,49
Novembre	60 956,72
Décembre	136 869,67
Total	249 155,03

TABLE 4.6 – Provision

Remarque : Dans cette sous-section nous avons présenté la méthode de Chain-Ladder traditionnellement utilisée pour le provisionnement des sinistres non connus, toutefois il existe de

nombreuses autres méthodes qui proposent une évaluation stochastique de cette provision. Ces méthodes permettent de valider (au moins en partie) les hypothèses utilisées, d'évaluer la variabilité des provisions à constituer, de construire des intervalles de confiance pour les estimateurs d'intérêt et d'estimer par la méthode de Monte-Carlo les sinistres futurs. Parmi ces méthodes stochastiques nous pouvons citer :

- *La méthode de Mack [1982]* : la provision obtenue par cette méthode est exactement la même que celle donnée par la méthode de Chain-Ladder.
- *Les modèles linéaires généralisés (J. Nelder & R. Wedderburn [1972])* : ces modèles sont appliqués aux triangles non cumulés et les provisions qu'elles donnent ont pour *Benchmark* celles estimées par Chain-Ladder.

Quelle que soit la méthode de provisionnement utilisée, l'estimation des sinistres futurs peut être affinée par l'utilisation de la méthode de Monte-Carlo. L'idée est de prendre comme provision le montant \hat{R} estimé par le moyen empirique des B provisions obtenues en appliquant le modèle de provision considéré aux B déterminés par la méthode de *Bootstrap* : $\hat{R} = \sum_{b=1}^B \hat{R}^b$.

4.1.3 La provision pour risque croissant

Il peut paraître peu naturel de parler de provisions pour risque croissant en assurance maladie car les contrats étant annuels, l'assureur a la possibilité d'ajuster le niveau de la prime à la juste valeur des engagements auxquels il souscrit chaque année. En pratique cet ajustement tarifaire est soumis à des contraintes principalement commerciales (anti-sélection et concurrence) et juridique (plafonnement des taux de cotisations des futurs retraités par la loi Evin). Cette provision a pour but de compenser l'inadéquation des cotisations futures à l'évolution estimée des prestations que l'assureur s'engage à offrir.

La provision pour risque croissant est généralement constituée lorsque le groupe de personnes couvert est essentiellement composé de seniors. Le nombre de seniors étant de plus en plus important, la gestion des risques qui les concernent constitue un enjeu important pour les assureurs. L'évaluation de la provision pour risque croissant n'est encadrée (à ce jour) par aucune exigence réglementaire, les facteurs à prendre en compte pour l'évaluer sont donc appréciés au cas par cas. Toutefois, il est judicieux de tenir compte des paramètres suivants :

- le périmètre couvert (actifs et/ou retraités),
- le départ en retraite, lorsque le périmètre couvert contient des actifs,
- la mortalité,
- un taux d'actualisation économique,
- et le niveau de cotisation à maintenir.

Ces facteurs sont ceux que nous utiliserons dans la suite mais ils ne sont pas exhaustifs. La variation des cotisations, le taux d'ajustement des tarifs lors du passage en retraite, le taux d'embauche, le désengagement éventuel de la sécurité sociale sont quelques paramètres qui pourraient également être utilisés. Les dépenses qui existent entre certains des paramètres sus-cités peuvent eux aussi être des facteurs à considérer.

Dans la pratique, le groupe d'assurés considéré est supposé fermé (pas d'entrée) et pour des faibles effectifs, la méthode consiste en une estimation tête par tête des provisions. Lorsque l'effectif est important, les assurés doivent être regroupés par profil de risque, les critères de segmentation pouvant être l'âge, le sexe, la catégorie socioprofessionnelle et la composition familiale.

Formulation mathématique

Considérons un groupe fermé constitué de n salariés d'âge moyen x pour lequel la sortie se fait par le décès. Les futurs retraités bénéficieront de la même couverture que les actifs en contrepartie d'une cotisation majorée d'un taux $\alpha_{k(k>0)}$. Soient h la proportion d'hommes, $f = 1 - h$ la proportion de femmes, d le taux de dérive annuelle moyenne de la consommation médicale et i le taux d'actualisation économique.

Le principal déterminant de la provision pour risque croissant en assurance maladie est l'âge. Elle était par le passé dénommée *réserve de vieillissement*, ce qui suggère d'utiliser le taux d'actualisation réglementairement préconisé pour des engagement viagers, soit 60% du taux moyen des emprunts d'États (TME) calculé sur une base semestrielle et plafonné à 0,25%. Ainsi, $i = \min(0,25\% ; 60\% \times TME)$.

$$\begin{aligned}
 VAP(\text{assurés}) &= n\pi + \underbrace{\sum_{k=1}^T \frac{n\pi \times {}_k\bar{p}_x \times (1 - {}_k r_x)}{(1+i)^k}}_{\text{actifs}} + \underbrace{\sum_{k=1}^T \frac{(1+\alpha_k) \times n\pi \times {}_k\bar{p}_x \times {}_k r_x}{(1+i)^k}}_{\text{retraites}} \\
 VAP(\text{assurés}) &= n\pi + n\pi \sum_{k=1}^T \frac{{}_k\bar{p}_x \times (1 + \alpha_k {}_k r_x)}{(1+i)^k} \\
 VAP(\text{assureur}) &= n\pi + \underbrace{\sum_{k=1}^T \frac{(1+d)^k \times n\pi \times {}_k\bar{p}_x \times (1 - {}_k r_x)}{(1+i)^k}}_{\text{actifs}} + \underbrace{\sum_{k=1}^T \frac{(1+d)^k \times n\pi \times {}_k\bar{p}_x \times {}_k r_x}{(1+i)^k}}_{\text{retraites}} \\
 VAP(\text{assureur}) &= n\pi + n\pi \sum_{k=1}^T \frac{(1+d)^k \times {}_k\bar{p}_x}{(1+i)^k}
 \end{aligned}$$

où ${}_k r_x$ est le taux de départ en retraite dans k années du groupe de salariés d'âge moyen x et ${}_k\bar{p}_x = h \times {}_k p_x^h + f \times {}_k p_x^f$, avec ${}_k p_x^h$ la probabilité de survie entre x et $x+k$ donnée par la TGH05 et ${}_k p_x^f$ celle donnée par la TGF05 (la TGH05 et la TGF05 sont respectivement des tables de mortalité générationnelles établies en 2005 des hommes et des femmes). π est ici la cotisation nette annuelle par tête que l'assureur souhaite maintenir sur l'horizon de temps T (en année).

La provision pour risques croissants est donnée par :

$$PRC = VAP(\text{assureur}) - VAP(\text{assurés})$$

Mise en œuvre

Considérons un portefeuille d'assurance maladie collective caractérisé par les données suivantes : $n = 1518$, $x = 54$, $\pi = 788,48$ euros, $h = 53,7\%$, et $f = 46,3\%$. Au 31 décembre 2017, le TME semestriel était de 0,75%, soit un taux d'actualisation économique $i = \min(0,25\% ; 60\% \times 0,75\%) = 0,25\%$. Pour cette mise en œuvre, nous retenons une dérive annuelle $d = 1\%$ de la consommation médicale (il s'agit du taux couramment utilisé sur le marché). L'horizon de projection retenu est $T = 40$ ans.

Les assurés de ce portefeuille constituent un groupe fermé composé de deux sous-groupes (les actifs et les retraités). Pour ce groupe, l'évolution des probabilités de survie par sexe et du taux de départ en retraite (ou proportion d'actifs) est illustrée par le figure 4.1.

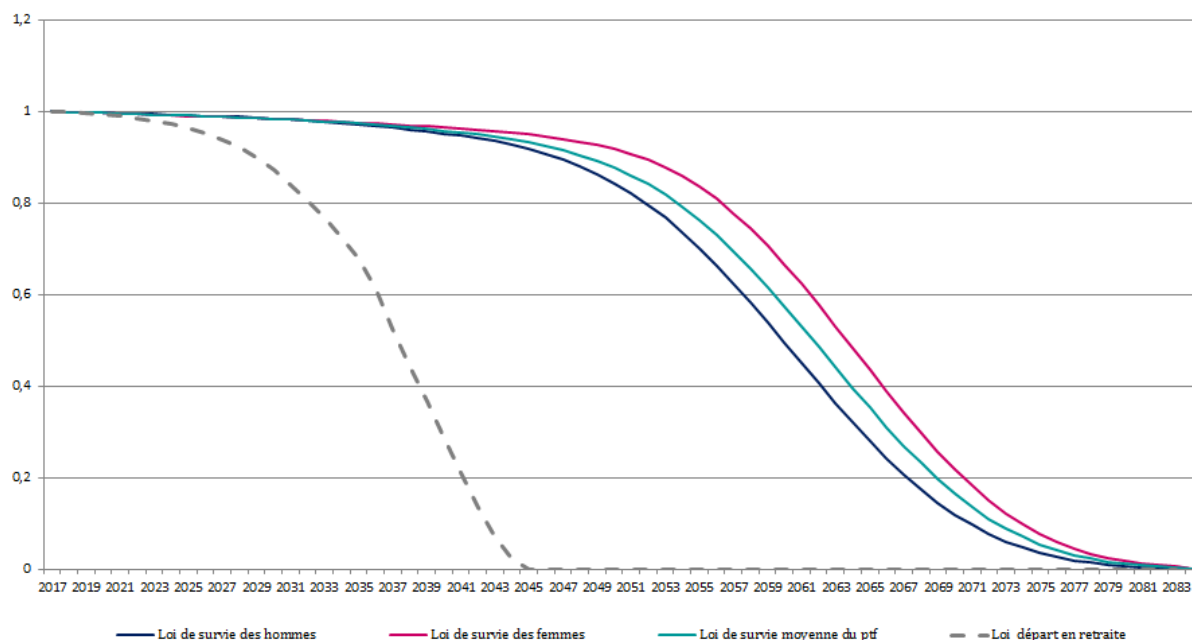


FIGURE 4.1 – Évolution des Probabilités de survie et du taux de départ en retraite.

Lorsque les données démographiques ne donnent pas d'informations sur le sexe, il est recommandé par prudence d'utiliser la TGF05, ce qui a pour effet d'augmenter le niveau de la provision car les femmes vivent plus longtemps que les hommes (cf. figure 4.1). Sur les 40 ans de projection que nous avons choisis, les cotisations et les prestations des actifs et des retraités vont évoluer comme l'illustre la figure 4.2.

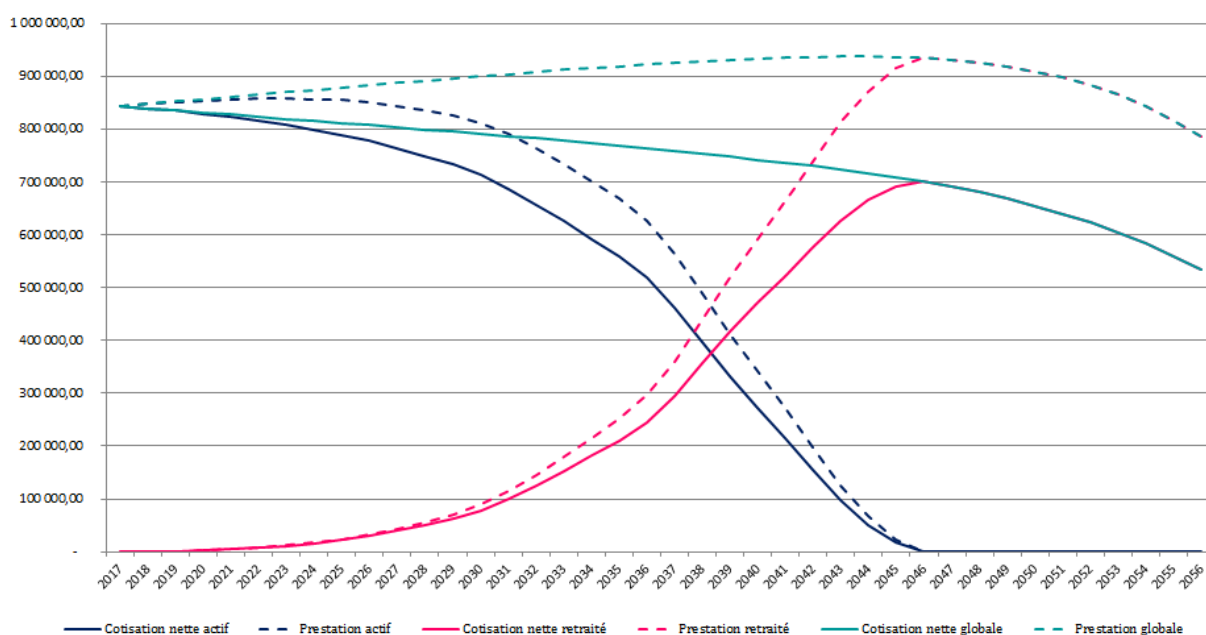


FIGURE 4.2 – Évolutions des cotisations et prestations.

La provision pour risque croissant est déterminée en faisant la différence entre les valeurs

actualisées des prestations et des cotisations représentées par la figure 4.2. Ainsi au 31 décembre 2017, la PRC des actifs est de 1 869 583 euros et celle des retraités est de 4 285 876 euros. La provision globale est donc 6 155 459 euros. Sur l'horizon de temps considérée, ces provisions vont être consommées de la manière représentée par la figure 4.3.

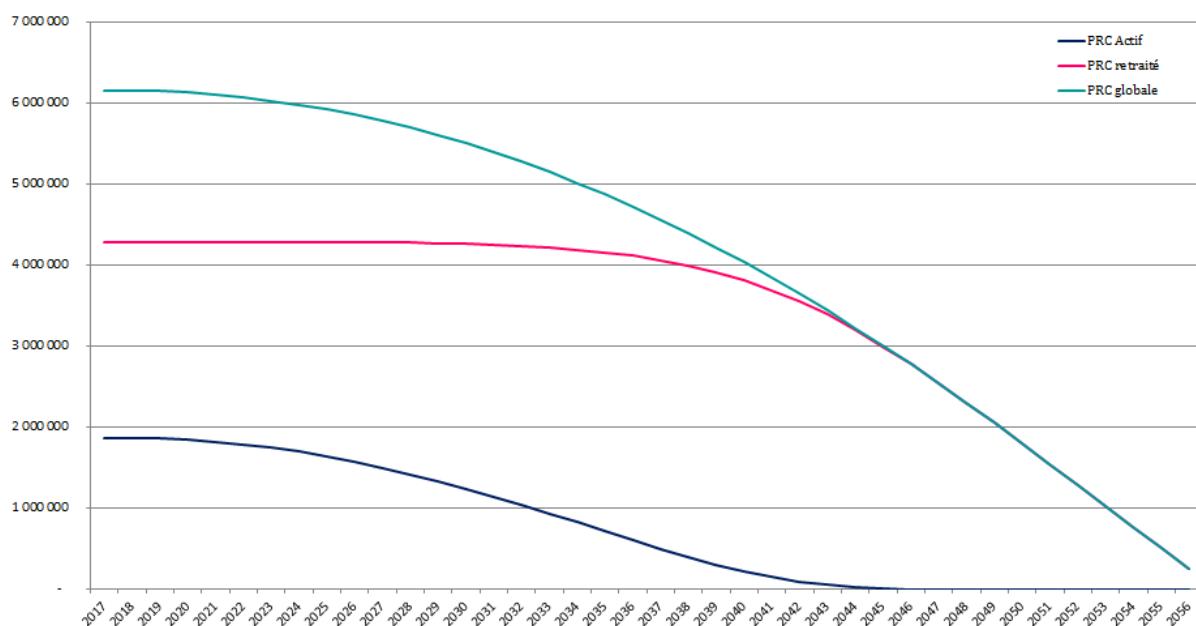


FIGURE 4.3 – Provision pour risque croissant.

Sensibilité de la PRC aux paramètres taux d'actualisation et horizon d'évaluation T

La formulation mathématique que nous avons donnée à la PRC peut être plus affinée suivant les cas par la prise en compte de paramètres spécifiques. Ces paramètres ont tous un impact sur le niveau de la PRC à constituer. Dans ce paragraphe, nous allons nous intéresser uniquement à deux paramètres : le taux d'actualisation et l'horizon de projection des cotisations et des prestations.

Globalement, la provision pour risque croissant est une fonction décroissante du taux d'actualisation et une fonction croissante de l'horizon de calcul (cf. figure 4.4). Cette représentation pose le problème du choix optimal des paramètres i et T lors de l'évaluation de cette provision, car des valeurs faibles du taux d'actualisation et des horizons de calcul importants pourraient conduire à un sur-provisionnement. Inversement, un taux d'actualisation important et un faible horizon de calcul pourrait conduire à un sous-provisionnement. Ces deux paramètres ont donc un fort impact sur le niveau de la provision à constituer, ils doivent être judicieusement choisis. En assurance maladie complémentaire, les coûts des biens et services subissent régulièrement des variations ce qui rend inapproprié des horizons de temps importants. Les horizons les plus utilisés sont 20 ans, 30 ans et 40 ans.

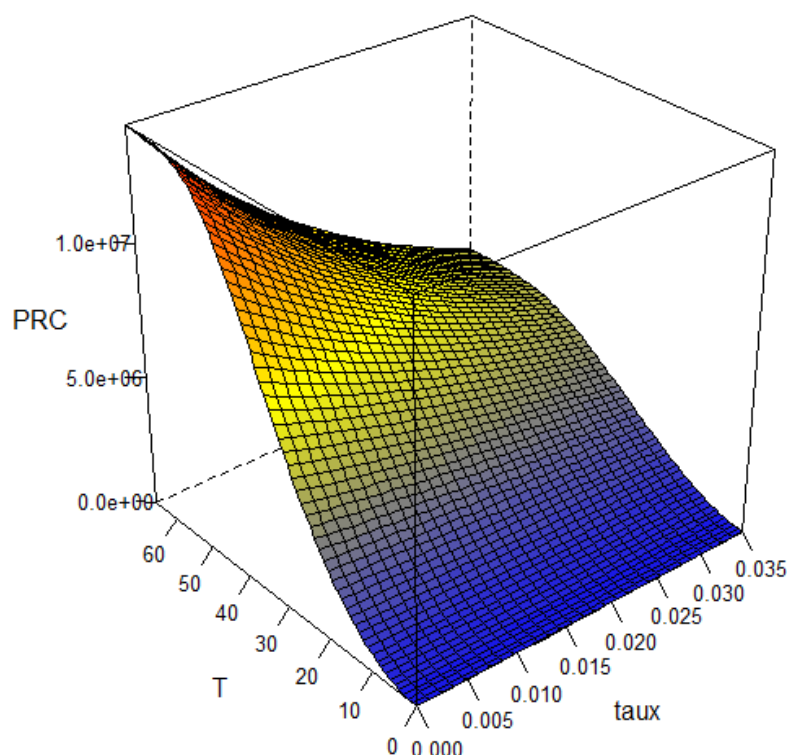


FIGURE 4.4 – Évolution de la PRC avec le taux d'actualisation et l'horizon d'évaluation T.

4.2 Rentabilité et pilotage d'un contrat d'assurance maladie complémentaire

Dans la section précédente, nous avons présenté les principales provisions rencontrées en assurance maladie complémentaire. L'évaluation de ces provisions vise une meilleure estimation du montant des sinistres que l'assureur s'est contractuellement engagé à couvrir et une estimation "juste" du ratio

$$S/P = \frac{\text{prestations} + \text{provisions}}{\text{prime pure}}$$

Ce ratio représente donc le quotient du montant estimé des sinistres rapportés aux cotisations nettes des taxes et des chargements. Ces montants sont relatifs à un contrat et mesurés pour une période de couverture pouvant être mensuelle, trimestrielle, semestrielle ou annuelle.

Le ratio S/P est fondamental dans l'analyse de la rentabilité et le pilotage d'un contrat d'assurance maladie complémentaire, il fournit une aide précieuse quant à la position tarifaire à adopter. Son interprétation est la suivante :

- $S/P > 1$: le montant des sinistres est supérieur à celui des primes pures ou cotisations nettes encaissées, le contrat est déficitaire.
- $S/P < 1$: le montant des sinistres est inférieur à celui des primes pures ou cotisations nettes encaissées, le contrat est rentable.
- $S/P = 1$: le montant des sinistres est égale à celui des primes pures ou cotisations nettes encaissées, le contrat est équilibré.

Le ratio S/P est généralement utilisé pour des besoins de pilotage des contrats déjà en

portefeuille, il permet de déterminer la modification à apporter à la prime pure pour que le contrat soit équilibré. Pour un contrat donné, la prime de l'exercice suivant (ou de l'année suivante $N + 1$) est, toutes choses égales par ailleurs, obtenue de la manière suivante :

$$P_{N+1} = P_N \times (1 + \alpha)$$

où α est une quantité dont le niveau dépend de l'appétence au risque de l'assureur. La quantité α est telle que

$$1 + \alpha \geq S/P \times (1 + d) \times (1 + g)$$

c'est-à-dire

$$\alpha \geq S/P \times (1 + d) \times (1 + g) - 1$$

où g représentent le taux global de changement de garanties et d l'anticipation de la dérive de la consommation médicale entre les exercices N et $N + 1$. Cette dérive dépend principalement de l'inflation des biens et services médicaux. Elle peut aussi prendre en compte un éventuel désengagement de la sécurité sociale et le vieillissement global des assurés.

Remarque : En assurance collective, le redressement des cotisations peut être fait par catégorie objective de salariés ou de manière globale pour l'ensemble des salariés afin de mutualiser l'évolution tarifaire.

Exemple : Considérons une police d'assurance pour laquelle l'arrêté des comptes au 31/12/2017 est celui donnée par le tableau 4.7.

Catégorie	effectif	cotisation nette	prestations	provisions	solde
Cadre	17	22 079,33	24 022,98	336,32	-2 279,97
Non cadre	166	145 220,93	127 897,06	1 790,56	15 533,31
Total	183	167 300,26	151 920,04	2 126,88	13 253,34

TABLE 4.7 – Compte d'un contrat collectif d'assurance maladie complémentaire arrêté au 31/12/2017.

Partant du tableau 4.7, les différents ratios S/P sont de l'exercice 2017 sont :

$$S/P_{cadre} = \frac{24\,022,98 + 336,32}{24\,022} = 110,33\%$$

$$S/P_{non\,cadre} = \frac{127\,897,06 + 1\,790,56}{145\,220,93} = 89,30\%$$

$$S/P_{global} = \frac{151\,920,04 + 2\,126,88}{167\,300,26} = 92,08\%$$

D'après ces niveaux de S/P , le "contrat cadre" est largement déficitaire. Ce déficit est globalement compensé par le "contrat non-cadre" qui est quant à lui bénéficiaire.

En supposant que les garanties souhaitées pour l'exercice 2018 sont les mêmes qu'en 2017 pour chaque catégorie objective de salariés (c'est-à-dire $g = 0\%$) et que la dérive de la consommation médicale est $d=1\%$ on obtient

$$\alpha_{cadre} \geq S/P_{cadre} \times (1 + d) \times (1 + g) - 1 = 11,43\%$$

$$\alpha_{non\ cadre} \geq S/P_{non\ cadre} \times (1 + d) \times (1 + g) - 1 = -9,80\%$$

$$\alpha_{global} \geq S/P_{global} \times (1 + d) \times (1 + g) - 1 = -7,00\%$$

Pour les calculs, nous retiendrons $\alpha_{cadre} = 11,43\%$, $\alpha_{non\ cadre} = -9,80\%$ et $\alpha_{globale} = -7,00\%$.

L'assureur a deux possibilités d'ajustement des tarifs :

- la première consiste en un redressement par catégorie objective de salariés (cf. tableau 4.8)
- et la seconde consiste en un redressement global de l'ensemble des cotisations (cf. tableau 4.9).

Catégorie	effectif	cotisation nette 2017	$1 + \alpha$	Cotisation nette attendue
Cadre	17	22 079,33	111,43%	24 602,89
Non cadre	166	145 220,93	90,20%	130 984,50
Total	183	167 300,26	/	155 587,39

TABLE 4.8 – Tarifs 2018 (alternative 1 : redressement par catégorie objective de salariés).

Catégorie	effectif	cotisation nette 2017	$1 + \alpha$	Cotisation nette attendue
Cadre	17	22 079,33	93,00%	20 533,53
Non cadre	166	145 220,93	93,00%	135 053,86
Total	183	167 300,26	/	155 587,39

TABLE 4.9 – Tarifs 2018 (alternative 2 : redressement global des cotisations).

Ces deux alternatives de redressement tarifaire sont équivalentes en terme d'assiette globale de cotisations.

Conclusion de la première partie

La consommation des biens et services médicaux est passée de 2,5% du PIB en 1950, à 11,54% en 2014, donnant ainsi à la santé, une place de plus en plus importante dans les débats économiques et politiques. Les questions de santé portent sur l'invalidité, sur les accidents du travail et principalement sur la maladie. D'après les comptes de la protection, 81,7% des dépenses de santé en 2015 étaient dues à la maladie, 15,4% à l'invalidité et 3,5% aux accidents du travail. Ces dépenses font de l'assurance maladie (obligatoire et complémentaire) une nécessité : elle offre des prestations dont le but est de couvrir les pertes financières générées par la maladie. En France, un peu plus de 94% de la population bénéficie d'une complémentaire santé. Cette couverture complémentaire est la contrepartie d'une prime d'assurance qui permet d'indemniser les assurés malades. La maladie et son degré de gravité sont aléatoires et imprévisibles, pour les gérer efficacement, les organismes d'assurance maladie complémentaire constituent des provisions.

Deuxième partie

L'open DAMIR, traitements des données et statistiques descriptives

Introduction

Les *open data* sont des données publiques auxquelles tout le monde peut accéder sans autorisation particulière. Ces données ont vocation à être partagées et réutilisées par tous. Elles sont un vecteur d'innovation pour la recherche et les entreprises. Le développement technologique est le principal facteur qui a rendu possible l'ouverture massive des données à laquelle nous assistons aujourd'hui. Cette ouverture des données est confrontée à une difficulté majeure, celle de la protection des informations personnelles. Les producteurs d'open data prennent en compte le risque de présence d'une donnée personnelle, ils ont recours à l'anonymisation ou au recueil du consentement.

Lorsqu'il s'agit de données personnelles, les techniques d'anonymisation ont pour but de rendre impossible l'identification singulière d'un individu et des informations le concernant car une donnée peut être non nominative tout en étant personnelle, ce qui accroît les possibilités de ré-identification. Les méthodes d'anonymisation appliquées doivent donc être irréversibles [2]. Les méthodes les plus utilisées actuellement sont la *k-anonymisation*[36], la *l-diversité*[16] et la *sécurité différentielle*[25].

Les données exploitées dans cette partie sont issue de *l'open DAMIR* (Dépenses d'Assurance Maladie Inter Régimes).

Chapitre 5

Présentation générale des données DAMIR

En actuariat, il est courant d'exploiter des jeux de données de taille relativement importante. Avec le développement des *open data*, les données sont de plus en plus disponibles à des coûts d'acquisition faibles et parfois inexistants. Les actuaires sont donc confrontés à de nouveaux défis en terme de gestion de volumétrie et de variété des données. A ces problèmes il faut rajouter complexité (vitesse d'exécution) des algorithmes à mettre en jeu pour collecter, stocker, traiter et analyser dans des délais courts et parfois en temps réel.

5.1 Présentation générale

L'*open DAMIR* a pour but de décrire les dépenses de santé de la population française. Ses données concernent l'ensemble des prestations réalisées par l'assurance maladie obligatoires pour l'ensemble des régimes de base, par région, par tranche d'âge, par acte de soin, et regroupe plusieurs catégories de risques : *maladie, accident du travail-maladie professionnelle, maternité, invalidité, décès*.

5.1.1 Source des données

Le jeu de données considéré dans ce projet est la base DAMIR (Dépenses Assurance Maladie inter-régime). Cette base est issue du SNIIRAM (Système National d'Information Inter-Régime de l'Assurance Maladie) et mise à disposition du grand public depuis février 2015.

Le SNIIRAM

Le SNIIRAM a été créé en 1999 par la loi de financement de la sécurité sociale (article L. 161-28-1). Ses objectifs principaux sont de contribuer :

- "A la connaissance des dépenses de l'ensemble des régimes d'assurance maladie par circonscription géographique, par nature de dépenses, par catégorie de professionnels responsables de ces dépenses et par professionnel ou établissement ;
- A la transmission en retour aux prestataires de soins d'informations pertinentes relatives à leur activité et leurs recettes, et s'il y a lieu à leurs prescriptions ;
- A la définition, à la mise en œuvre et à l'évaluation de politiques de santé publique (ce troisième objectif date de 2004)."

Les données contenues dans ce système d'information résultent principalement : des fichiers administratifs, des fichiers des services médicaux qui permettent d'avoir des informations sur les bénéficiaires, des répertoires de professionnels qui renseignent sur les prestataires de service médicaux, des feuilles de soins et des remboursements, des bordereaux de facturation des cliniques, des arrêts de travail et des indemnités journalières et des résumés de sortie hospitaliers transmis depuis 2007 par l'Agence Technique d'Information sur l'Hospitalisation (ATIH). Ces informations recueillies contiennent un nombre important de données personnelles des bénéficiaires et des praticiens notamment, les noms, prénoms, dates et lieux de naissance, numéro de sécurité sociale. Ces données personnelles sont supprimées et remplacées par un numéro unique et anonyme qui correspond à un numéro de chaînage issue d'une méthode d'anonymisation irréversible.

Comme le précise le *RAPPORT SUR LA GOUVERNANCE ET L'UTILISATION DES DONNEES DE SANTE* [5], l'anonymisation ne supprime pas le risque de ré-identification car l'intérêt de ce système d'information est sa précision et son exhaustivité. Concernant les données du SNIIRAM, ce rapport précise que les données individuelles qu'elle contient sont "bien anonymes prises une par une, en ce sens qu'elles ne comportent pas l'identité des personnes, mais qu'elles ne peuvent pas être en accès libre parce qu'en croisant certaines informations qui y figurent, on peut identifier des personnes connues par ailleurs (des proches, des collègues ou des célébrités)." L'accès aux données de ce système d'information est encadré par la commission nationale de l'information et des libertés (Cnil)[12]. Les différents niveaux d'accès sont donnés par la figure 5.1.

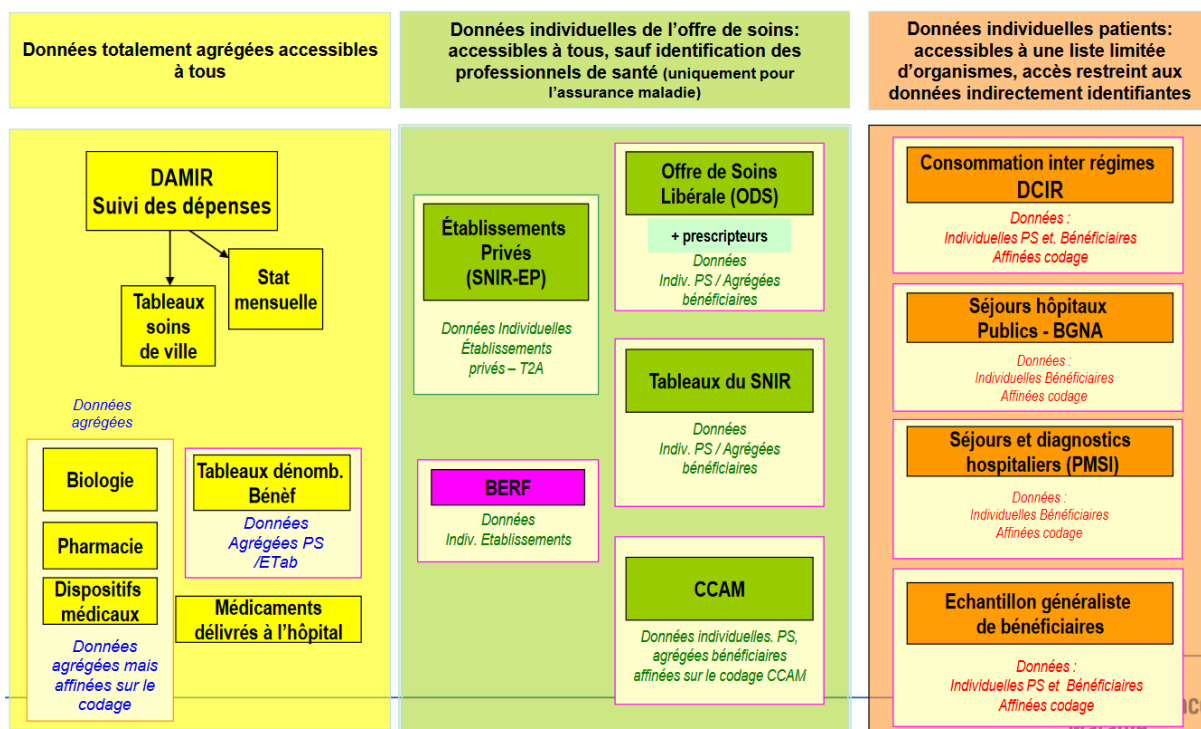


FIGURE 5.1 – Les trois niveaux d'accès aux données du SNIIRAM (source : ameli.fr).

Le DAMIR

La base DAMIR rend compte des prestations de l'assurance-maladie tous régimes confondus. Elle est juridiquement encadrée par les contraintes à l'ouverture de données publiques et est par construction agrégée afin de préserver l'anonymat des bénéficiaires et des prestataires d'actes et services médicaux. A ce jour la couverture temporelle de cette base s'étend du 01/01/2009 au 31/12/2016. La couverture spatiale est quant à elle organisée en neuf zones géographiques (pour les données relatives aux années 2009 - 2014) qui sont des regroupements de régions administratives et en treize zones géographiques (à partir de 2015) proches des grandes régions administratives créées par la réforme territoriale de 2015. Dans cette base les dépenses de santé sont détaillées suivant huit axes d'analyses :

— *La période de traitement*

Cet axe est constitué d'une seule variable qui donne l'année et le mois de traitement informatique et d'enregistrement dans la base des données portée par la même ligne d'observations.

— *Les prestations versées*

Cet axe d'analyse est porté par treize variables. Elles rendent par exemple compte : de la nature de la prestation (consultations, frais de séjour, forfait journalier, etc), de la nature d'assurance (maladie, maternité, accident de travail-maladie professionnelle, décès, invalidité, etc), de la nature de l'accident de travail (accident sur le lieu de travail, maladie professionnelle, accident du trajet), du secteur d'activité du bénéficiaire (public, privé), de la nature du destinataire (assuré, conjoint, conjoint séparé, conjoint divorcé, concubin, pacsé, enfant, correspondant, d'entreprise, mutuelle, etc), du type de remboursement (prestation de référence, complément d'acte, supplément Alsace Moselle, supplément hors Alsace Moselle, ticket modérateur CMU, ticket modérateur hors CMU, forfait CMU) et du taux de remboursement de l'acte de soin considéré.

— *L'organisme de soin*

Cet axe est constitué d'une variable donnant la zone géographique de l'organisme de liquidation des prestations.

— *La période de soin*

Cet axe est constitué de deux variables qui nous renseignent sur l'année et le mois de réalisation de l'acte de soin ou d'un service médical.

— *Le bénéficiaire du soin*

Cet axe est porté par six variables qui nous renseignent sur le sexe du bénéficiaire, sa tranche d'âge au moment du soins, sa qualité (assuré, conjoint et assimilé, enfant, autre ayant-droit), sa zone de résidence, sur sa qualité de bénéficiaire de la CMU et sur le niveau du ticket modérateur (qui dépend du régime auquel est rattaché l'assuré principal).

— *L'exécutant du soin*

L'exécutant est celui qui réalise l'acte ou le service médical prescrit. *L'open* DAMIR le décrit par douze variables qui permettent de connaître : sa catégorie (médecin, fournisseur,

sage femme, infirmier, masseur, etc), sa spécialité (médecine générale, anesthésiologie-réanimation, pathologie cardio-vasculaire, chirurgie, dermatologie et vénéréologie, radiologie, gynécologie obstétrique, etc), son type d'activité, son statut juridique, la région d'implantation du prescripteur et de l'établissement prescripteur, la catégorie de cet établissement (centre hospitalier, hôpital local, centre hospitalier spécialisé, etc), la discipline de cet établissement (dialyse, gynécologie et obstétrique, scanner-IRMN-tomographie, etc) et le mode traitement pratiqué par cet établissement (en externe, à domicile ou en hospitalisation).

— *Le prescripteur du soin*

Le prescripteur est quant à lui décrit par sept variables qui permettent de déterminer, sa catégorie, sa spécialité, son type d'activité, sa zone géographique, son statut juridique, la région d'implantation de l'établissement auquel il appartient et la catégorie de cet établissement.

— *Les indicateurs de montant*

Il s'agit des indicateurs de dépenses de santé. cet axe est organisé en treize variables dont six sont des variables de montant (qui nous renseignent sur les volumes de dépassements, de dépenses réelles et de remboursement de la sécurité sociale), quatre sont des variables de comptage qui renseignent sur les fréquences de prestations, deux sont des variables de poids (qui représentent la part de la prestation considérée dans le total des prestations) et une variable donne le taux de remboursement de chaque acte de soins.

Remarque :

Les douze premières variables évoquées pour le dernier axe d'analyse cité se distinguent deux à deux par leur caractère "préfiltré". Les indicateurs préfiltrés au sens du DAMIR permettent d'étudier le régime obligatoire. Lorsque le remboursement concerne une prestation de référence ou un complément d'acte, les indicateurs préfiltrés donnent le niveau de dépense correspondant. Par contre lorsqu'il s'agit d'une prestation supplémentaire, d'une prise en charge PUMa ou d'une exonération du ticket modérateur du fait de la PUMa, les indicateurs préfiltrés valent zéro. La différence entre les prestations préfiltrées et non préfiltrées est illustrée par le tableau 5.1.

Ainsi, les indicateurs préfiltrés permettent d'étudier le régime obligatoire tandis que les indicateurs non préfiltrés sont utiles pour l'analyse des prestations ne relevant pas du régime obligatoire, notamment la CMU-C, les compléments Alsace-Moselle et certaines prestations de prévention.



	<i>prs_rem_typ</i>	<i>prs_rem_mnt</i>	<i>flt_rem_mnt</i>
code prestation	libellé prestation		
0	prestation de référence	1.958e+08	1.958e+08
1	complément d'acte	2.015e+06	2.015e+06
2	ticket modérateur hors CMU	1.061e+06	0.000e+00
3	supplément hors Alsace Moselle	6.023e+03	0.000e+00
4	supplément Alsace Moselle	7.144e+05	0.000e+00
5	ticket modérateur CMU	2.997e+06	0.000e+00
6	forfait CMU	7.244e+05	0.000e+00
99	valeur inconnue	1.265e+07	1.265e+07

TABLE 5.1 – Exemple de prestation préfiltrée. La variable *prs_rem_typ* représente les types de remboursements, la variable *prs_rem_mnt* représente les montants versés/remboursés et la variable *flt_rem_mnt* correspond aux montants versés/remboursés préfiltrés.

5.1.2 Données et outils

Les différents axes d'analyse sus-présentés forment un ensemble de 55 variables (cf. annexe 1). La présentation détaillée de ces variables est donnée par le descriptif de l'*open DAMIR* [9]. A ce jour les 8 années de données dont nous disposons sont constitués de

- 218 134 047 observations pour l'année de traitement 2009 (91 Go au format sas7bdat),
- 235 479 973 observations pour l'année de traitement 2010 (98 Go au format sas7bdat),
- 240 167 951 observations pour l'année de traitement 2011 (100 Go au format sas7bdat),
- 244 175 492 observations pour l'année de traitement 2012 (102 Go au format sas7bdat),
- 251 646 650 observations pour l'année de traitement 2013 (105 Go au format sas7bdat),
- 253 982 906 observations pour l'année de traitement 2014 (106 Go au format sas7bdat),
- 347 789 914 observations pour l'année de traitement 2015 (145 Go au format sas7bdat),
- 365 356 954 observations pour l'année de traitement 2016 (153 Go au format sas7bdat).

Ici le mot "observations" est employé au sens d'une ligne d'observation (ou de données). Les huit années de *datas* disponibles sont donc constituées de 2 156 733 887 lignes de données qui ont un poids total de 900 Go (Gigas octets) au format sas7bdat. Afin de gérer ce grand volume de données nous avons opté pour un stockage sur un serveur local. Cette solution nous a paru la plus adaptée pour les études exploratoires et les traitements que nous avons effectués. Pour nos travaux nous utilisons les logiciels  (pour la phase exploration) et  (pour la modélisation).

5.2 Forme et type des données

Le jeu de données que nous avons construit en regroupant les bases mensuelles de l'*open DAMIR* forme un "rectangle" de 2Md de lignes et de 55 colonnes qui contient des données de types variés.

5.2.1 Définitions (*Variable aléatoire et type d'une variable*)

Considérons un espace de probabilité (Ω, P) et Φ un ensemble non vide. Ω représente notre rectangle de données et P la probabilité historique sous laquelle est notre jeu de données.

Toute fonction X définie de Ω vers Φ (notation $X : \Omega \longrightarrow \Phi$) est une variable aléatoire.

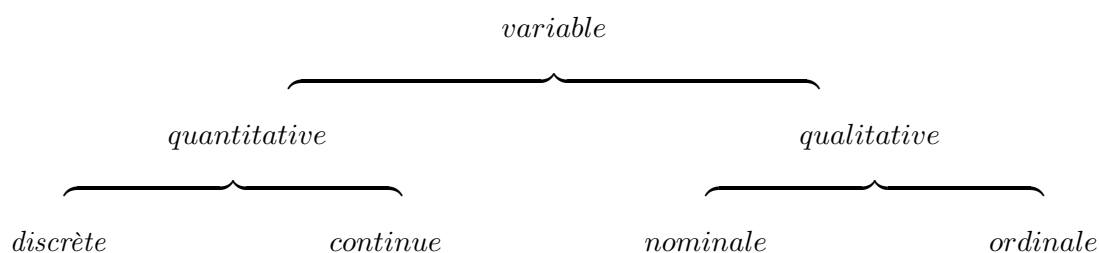
- Lorsque $\Phi \subseteq \mathbb{R}$ la variable X est dite réelle. Les variables réelles sont **quantitatives**, elles peuvent être ordonnées et sommées. En particulier si Φ est dénombrable (c'est-à-dire en bijection avec N) la variable X est **discrète** et Φ correspond à un continuum de valeurs, X est **continue**.

En pratique, une variable quantitative ne peut prendre qu'un nombre fini de valeurs. Ce qui permet de différencier les variables discrètes des variables continues est le nombre de valeurs possibles (par exemple le poids, les taux de remboursement de la sécurité sociale). Lorsque ce nombre est grand on considère que la variable est continue, c'est typiquement le cas pour les variables dont les observations sont issues d'un outil de mesure. Si le nombre de valeurs possibles est faible la variable est discrète. Ces valeurs sont généralement isolées et entières (par exemple le nombre de bénéficiaires d'une police d'assurance, le nombre de sinistre au cours d'une année).

- Lorsque Φ est un ensemble de valeurs représentant des "qualités" appelé modalités, la variable est **qualitative** (ou catégorielle). Ces modalités peuvent s'exprimer de façon littérale ou numérique. Ces variables sont différenciées en variables nominales et ordinales. Une variable est qualitative nominale lorsque ses modalités ne peuvent pas être ordonnées. Par exemple la variable donnant la discipline de prestation (orthopédie, kinésithérapie, pédiatrie, etc) d'un établissement de santé est qualitative nominale. Contrairement aux variables nominales, les variables qualitatives ordinales ont des modalités qui peuvent être ordonnées, les tranches d'âges par exemple (0 - 19 ans, 20 - 29 ans, 30 - 39 ans, etc).

5.2.2 Types des variables de base DAMIR

Les définitions données par la sous-section précédente vont être appliquées aux variables de l'*open DAMIR* afin de les différencier suivant leur type. Le schéma de différenciation est le suivant :



L'application de ce schéma nous a permis d'attribuer un type à chaque variable de notre base de données. Cette étape était nécessaire car les méthodes de traitement et de descriptions des variables dépendent de leurs types. La majorité des variables de notre jeu de données sont qualitatives. La base compte 41 variables qualitatives et 14 variables quantitatives. Parmi les variables qualitatives 37 sont nominales et 4 ordinales. Parmi les variables quantitatives 10 sont continues et 4 discrètes (cf. Annexe 1).

5.2.3 Les valeurs manquantes

La gestion des valeurs manquantes est un sujet qui apparaît très souvent dès qu'il est question de base de données. Ces données absentes ne peuvent pas être ignorées lors d'une analyse

actuarielle. Une valeur manquante peut être due à une non-réponse réelle (cela est typiquement le cas lorsque qu’une fiche de soins transmise au SNIIRAM n’est pas complètement remplie), à une réponse inexploitable (c’est le cas lorsque la fiche de soins transmise est complète mais illisible) ou à un dysfonctionnement du système d’information. Généralement les non-réponses proviennent de phénomènes involontaires, c’est-à-dire qu’il n’existe pas de lien entre une non-réponse et l’origine de cette non-réponse. Lorsque qu’elles dépendent de la nature des variables portant les observations manquantes, le mécanisme de non-réponse est alors volontaire, c’est souvent le cas lorsqu’il s’agit de données sensibles. En santé par exemple, il peut s’agir des données liées à alcool, au tabac, à l’hygiène ou à la sexualité.

Les mécanismes de non-réponse

En 1987 Little et Rubin ont proposé une première catégorisation des données manquantes [18]. Soient $X = (X_{ij})$ une matrice $(n \times p)$ rectangulaire représentant notre jeu de données et $M = (M_{ij})$ une matrice d’indication des valeurs manquantes, $M_{ij} = 1$ si l’observation X_{ij} est absente et $M_{ij} = 0$ si X_{ij} est présente. Little et Rubin proposent 3 mécanismes de non-réponse :

— *Missing Completely At Random (MCAR)*

Une donnée est dite manquante de façon complètement aléatoire, c’est à dire *MCAR* si la probabilité d’absence est la même pour toutes les observations. Cette probabilité ne dépend donc pas des données observées ou non observées.

$$P(M = 1 | X^{obs}, X^{miss}) = P(M = 1)$$

où $X = (X^{obs}, X^{miss})$, $X^{obs} = X1_{\{M=0\}}$ et $X^{miss} = X1_{\{M=1\}}$

— *Missing At Random (MAR)*

Une donnée est dite manquante de façon aléatoire, c’est à dire *MAR* si la probabilité d’absence ne dépend que des données observées (X^{obs}). Cette probabilité ne dépend donc pas des données non-observées conditionnellement aux données observées.

$$P(M = 1 | X^{obs}, X^{miss}) = P(M = 1 | X^{obs})$$

— *Missing Not At Random (MNAR)*

Une donnée est dite manquante de façon non-aléatoire, c’est à dire *MNAR* si la probabilité d’absence dépend des données observées et non-observées. Dans ce cas même en tenant compte des données observées, les raisons pour lesquelles une observation est manquante dépend d’autres observations manquantes.

Selon le mécanisme et la proportion de données manquantes, différentes solutions théoriques sont proposées. Il peut s’agir des méthodes permettant de mener une analyse en présence de valeurs manquantes, de supprimer les variables et/ou les individus présentant des observations manquantes, ou de remplacer aux observations manquantes des valeurs obtenues par les méthodes d’imputations (cf. chapitre suivant). Dans la pratique la visualisation de la base de données permet de faire des hypothèses sur le mécanisme de non-réponse. Dans notre cas, la base de données est très volumineuse (2 156 733 887 de lignes et 55 colonnes) pour être visualisée en entier. Nous nous sommes donc contentés de déterminer le nombre de valeurs manquantes par variables et par années de traitement. Pour les 8 années d’observations, nous avons identifié 5 variables ayant

des valeurs manquantes (cf. tableau 5.2).

	<i>prs_act_nbr</i>	<i>prs_rem_bse</i>	<i>prs_rem_mnt</i>	<i>flt_act_nbr</i>	<i>psp_act_snds</i>
2009	50 061 573	206 245	206 259	46 496 718	0
2010	50 048 159	0	20	46 559 066	0
2011	51 174 965	0	17	47 583 751	22
2012	51 612 023	0	11	47 895 745	0
2013	51 994 291	0	16	48 175 705	0
2014	52 559 805	0	7	48 580 428	0
2015	71 920 388	0	7	66 650 906	0
2016	74 210 602	0	5	68 641 249	0
total	453 581 806	206 245	206 342	420 583 568	22
% de VM	21,030958	0,009563	0,009567	19,500949	0,000001

TABLE 5.2 – Nombre de valeurs manquantes (VM) par année de traitement et par variables concernées (*prs_act_nbr*=dénombrement de la prestation, *prs_rem_bse*=base de remboursement, *prs_rem_mnt*=montant versé/remboursé, *flt_act_nbr*=dénombrement de la prestation préfiltrée, *psp_act_snds*=nature d'activité).

Chapitre 6

Traitement des données manquantes et remarques sur les données atypiques

En présence de données manquantes, la méthode de traitement la plus simple consiste à exclure de la base de données toutes les lignes présentant au moins une valeur manquante. Ce qui permet ensuite d'effectuer des études sur celles dont toutes les valeurs sont présentes. Il s'agit de **l'analyse dite des cas complets**. Cette méthode est implémentée par défaut dans la majorité des logiciels d'analyse statistique mais le fait qu'elle n'utilise pas toute l'information disponible dans la base peut induire une importante imprécision sur les résultats obtenus.

Dans ce chapitre, nous nous intéressons à quelques méthodes de complétion de données manquantes. Ces méthodes seront théoriquement présentées avant d'être mises en œuvre.

6.1 Les méthodes de complétions des données manquantes

6.1.1 Méthode *LOCF* (*Last Observation Carried Forward*)

L'imputation des données manquantes par la méthode *locf* consiste à remplacer une valeur manquante par la dernière valeur observée. Cette méthode repose donc sur l'hypothèse de constance de la dernière valeur observée. Cette méthode est très utilisée pour des données médicales. En absence d'information il est courant de supposer que l'état pathologique du patient considéré n'a pas changé.

6.1.2 Méthode d'imputation par la moyenne

L'idée de cette méthode est de remplacer les valeurs manquantes de chaque variable par la moyenne de celles présentes. Partant de notre rectangle de données X , considérons une colonne $X_{.j}$ ayant des observations manquantes. $X_{.j}$ peut s'écrire

$$X_{.j} = (X_{.j}^{obs}, X_{.j}^{miss}) \in R^{n \times 1}$$

où $X_{.j}^{obs}$ est la composante observée de $X_{.j}$ et $X_{.j}^{miss}$ sa composante manquante. Ces composantes représentent respectivement l'ensemble des n^{obs} valeurs observées de $X_{.j}$ et l'ensemble des n^{miss} valeurs manquantes de $X_{.j}$ ($n^{obs} + n^{miss} = n$). Si toutes les valeurs de $X_{.j}$ étaient observées, sa moyenne serait

$$\bar{X}_{.j} = \frac{1}{n} \sum_i X_{ij} = \frac{1}{n} \left(\sum_{i^{obs}} X_{ij}^{obs} + \sum_{i^{miss}} X_{ij}^{miss} \right)$$

La méthode d'imputation par la moyenne suggère de remplacer chaque valeur manquante par la moyenne de celles observées ($\bar{X}_{.j}^{obs} = \frac{1}{n^{obs}} \sum_{i^{obs}} X_{ij}^{obs}$). La moyenne de $X_{.j}$ après imputation est :

$$\bar{X}_{.j}^* = \frac{1}{n} \left(\sum_{i^{obs}} X_{ij}^{obs} + \sum_{i^{miss}} \bar{X}_{.j}^{obs} \right) = \frac{n^{obs}}{n} \bar{X}_{.j}^{obs} + \frac{n^{miss}}{n} \bar{X}_{.j}^{obs}; \quad e.i \quad \bar{X}_{.j}^* = \bar{X}_{.j}^{obs}$$

La méthode d'imputation par la moyenne conserve la moyenne des valeurs présentes.

Aussi

$$var(X_{.j}) = \frac{1}{n} \sum_i (X_{ij} - \bar{X}_{.j})^2 = \frac{1}{n} \left(\sum_{i^{obs}} (X_{ij}^{obs} - \bar{X}_{.j})^2 + \sum_{i^{miss}} (X_{ij}^{miss} - \bar{X}_{.j})^2 \right)$$

En imputant aux observations manquantes la moyenne de celles présentes, on a

$$var(X_{.j})^* = \frac{n^{obs}}{n} \times \frac{1}{n^{obs}} \sum_{i^{obs}} (X_{ij}^{obs} - \bar{X}_{.j}^*)^2 + \frac{n^{miss}}{n} \times \frac{1}{n^{miss}} \sum_{i^{miss}} \underbrace{(\bar{X}_{.j}^{obs} - \bar{X}_{.j}^*)^2}_0$$

Donc $var(X_{.j})^* = (1 - \alpha^{miss}) var(X_{.j}^{obs})$; avec $\alpha^{miss} = \frac{n^{miss}}{n} = 1 - \frac{n^{obs}}{n}$

La méthode d'imputation des valeurs manquantes par la moyenne modifie la variance des valeurs présentes. Cette méthode a l'avantage d'être simple. Toutefois, la distorsion que subit la variance observée lorsque la proportion de valeurs manquantes est importante peut fortement biaiser les résultats d'une étude.

6.1.3 Méthode d'imputation par la médiane

Comme pour l'imputation par la moyenne, le principe de l'imputation des valeurs manquantes par la médiane consiste à remplacer les valeurs manquantes par la médiane des valeurs observées. La médiane d'une série statistique est la valeur qui divise les observations de cette série en deux parties, telles que 50% des observations lui sont supérieures et 50% lui sont inférieures.

6.1.4 Méthode *LOESS* (*LOcal regrESSion*)

La méthode *LOESS* est une méthode d'imputation des valeurs manquantes par régression locale. Les modèles de régression sont généralement utilisés pour prédire ou expliquer les valeurs d'une variable Y à partir des valeurs de p variables X_1, \dots, X_p dites explicatives. Pour des besoins de complétion, leurs mise en œuvre locale (dans le voisinage d'une valeur manquante) permet d'imputer les valeurs manquantes par les prédictions des modèles. Dans notre cas, nous avons procédé à des régressions linéaires locales. De manière générale, l'équation générique d'un modèle de régression linéaire multiple est de la forme

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \xi_i, \quad \forall i \in [1 : n]$$

où

- Les β_j sont les paramètres inconnus à estimer.
- y_i est la variable à prédire ou à expliquer.
- Les x_{ij} sont des variables explicatives ($j \in [1 : p]$).
- ξ_i est une variable d'erreur telle que $E(\xi_i) = 0$, $Var(\xi_i) = \sigma^2$ et $\forall i \neq j \text{ } cov(\xi_i, \xi_j) = 0$.

En utilisation l'écriture matricielle, l'équation du modèle est

$$Y = X\beta + \xi$$

$$\text{où } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \text{ et } \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}$$

Les hypothèses de ce modèle sont les suivantes :

- $E(\xi) = 0$. Cette hypothèse implique $E(Y) = X\beta$
- $Var(\xi) = \sigma^2 I_n$, I_n est la matrice identité.
- X est de plein rang ($rg(X) = p + 1$; $p + 1 < n$).
- X est déterministe.

Définition (Estimateur des MC)

On appelle estimateur des moindres carrés (MC) noté $\hat{\beta}$ et β la quantité suivante :

$$\hat{\beta} = \underset{\beta_1, \dots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^n \left\{ y_i - \sum_{j=1}^p \beta_j x_{ij} \right\}^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|Y - \beta\|_2^2$$

Pour obtenir $\hat{\beta}$, il est nécessaire que la dérivée première de $S(\beta) = \|Y - \beta\|_2^2$ s'annule. Cette dérivée est donnée par :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta$$

Si $\hat{\beta}$ vérifie $-2X'Y + 2X'X\hat{\beta} = 0$, alors

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

si la dérivée seconde de $S(\beta)$ est une matrice définie positive. Cet estimateur $\hat{\beta}$ est la valeur qui minimise $S(\beta)$.

En effet

$$\frac{\partial^2 S(\beta)}{\partial^2 \beta} = 2X'X$$

X étant de rang plein, $X'X$ est inversible et n'a pas de valeur propre nulle. La matrice $X'X$ est donc définie et de plus $\forall z \in R^{p+1} \ z'X'Xz = \|Xz\|_2^2 \geq 0$. $X'X$ est donc bien définie positive. $\hat{\beta}$ est donc la quantité qui minimise $S(\beta)$.

Remarque : $\hat{\beta}$ est un estimateur sans biais de β car $E(\hat{\beta}) = E((X'X)^{-1}X'Y) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$.

Comme nous l'avons déjà souligné dans les lignes précédentes, un des objectifs de la régression est de donner des prévisions de la variable d'intérêt Y . L'imputation des valeurs manquantes est un cas particulier des problèmes de prévision. Supposons que notre variable Y à une valeur y_{n+1} manquante. Schématiquement, notre jeu de données peut être représenté par le tableau 6.1.

y_1	$x_{1,1}$	\cdots	$x_{1,p}$
y_2	$x_{2,1}$	\cdots	$x_{2,p}$
\vdots	\vdots	\vdots	\vdots
y_n	$x_{n,1}$	\cdots	$x_{n,p}$
NA	$x_{n+1,1}$	\cdots	$x_{n+1,p}$

TABLE 6.1 – Exemple théorique de tableau de données avec une valeur manquante.

A partir de $x'_{n+1} = (1, x_{n+1,1}, x_{n+1,2}, \dots, x_{n+1,p})$ on obtient la valeur qui servira à l'imputation.

$$y_{n+1}^* = x'_{n+1} \hat{\beta}$$

y_{n+1}^* et y_{n+1} ont la même espérance.

En effet

$$E(y_{n+1}^*) = x'_{n+1} E(\hat{\beta}) = x'_{n+1} \beta = x'_{n+1} E(\beta) = E(x'_{n+1} \beta + \xi) = E(y_{n+1}).$$

La variance de l'erreur d'imputation est donnée par

$$Var(y_{n+1} - y_{n+1}^*) = var(x'_{n+1} \beta + \xi_{n+1} - x'_{n+1} \hat{\beta}) = \sigma^2 + x'_{n+1} Var(\hat{\beta}) x_{n+1}$$

or $Var(\hat{\beta}) = var((X'X)^{-1}X'Y) = (X'X)^{-1}X'var(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$
d'où

$$Var(y_{n+1} - y_{n+1}^*) = \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}).$$

Aussi

$$Var(y_{n+1} - y_{n+1}^*) = E[y_{n+1} - y_{n+1}^* - E(y_{n+1}) + E(y_{n+1}^*)]^2 = E(y_{n+1} - y_{n+1}^*)^2$$

La variance de l'erreur d'imputation est donc mesurée par l'erreur quadratique moyenne.

Dans la pratique les bases de données présentent plus d'une valeur manquante. Lorsque ces valeurs manquantes sont univariées, la prédiction est multiple, la composante Y^{miss} de Y est imputée par :

$$Y^{imp} = X^{imp} \hat{\beta}$$

avec $X^{imp} = (X_{i^{miss}_j}); j \in [1, p]$ et $i^{miss} \in [1^{miss}, n^{miss}] = \text{ensemble des numéros de lignes pour lesquelles } Y \text{ a des valeurs manquantes.}$

Lorsque les valeurs manquantes sont arbitrairement réparties entre plusieurs variables de la base de données, la méthode de régression locale est alors utilisée. Elle consiste à définir autour de chaque valeurs manquantes (ou suite de valeurs manquantes) un voisinage constitué de k lignes et l colonnes. On obtient ainsi un problème de régression linéaire classique : résolution de l'équation des moindres carrées et estimation de y_{n+1}^* (ou Y^{imp}).

6.1.5 Méthode kNN (k -Nearest Neighbors)

La méthode d'imputation des valeurs manquantes dite des kNN (k plus proches voisins) repose sur l'utilisation d'une distance.

Soit E un ensemble quelconque. Par définition, une distance sur E est une application

$$d : E \times E \longrightarrow [0; +\infty[$$

$$(x, y) \longmapsto d(x, y)$$

vérifiant les propriétés suivantes : $\forall x, y, z \in E$

- $d(x, x) = 0 \iff x = y$ "séparation"
- $d(x, y) = d(y, x)$ "symétrie"
- $d(x, z) \leq d(x, y) + d(y, z)$ "inégalité triangulaire"

La distance à utiliser dans pour mettre en œuvre la méthode kNN est laissée au libre choix du statisticien. Toutefois, la distance à choisir doit être adaptée aux données l'étude.

Lorsque la base de données est uniquement constituée de variables quantitatives, il est courant d'utiliser la distance euclidienne définie par :

$$d_e : R^n \times R^n \longrightarrow [0; +\infty[$$

$$(x, y) \longmapsto d_e(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Lorsque la base de données est uniquement constituée de variables qualitatives, il est courant d'utiliser la distance de Hamming définie par :

$$d_1 : E \times E \longrightarrow [0; +\infty[$$

$$(a, b) \longmapsto d_h(a, b) = \sum_{i=1}^n \mathbf{1}_{a_i \neq b_i}$$

où E est un ensemble constitué de suites de longueurs n à valeur dans l'ensemble des lettres d'un alphabet.

La question du choix de la distance est plus compliquée lorsque la base de données considérée est mixte, c'est-à-dire constituée de variables quantitatives et qualitatives. C'est le cas de la base DAMIR. Soit Ω une telle base de données. Afin de construire une métrique sur cet ensemble nous le divisons en deux sous ensembles supplémentaires Ω^{quanti} et Ω^{quali} représentant respectivement l'ensemble des variables quantitatives et l'ensemble des variables qualitatives de la base. On note $\Omega = \Omega^{quanti} \oplus \Omega^{quali}$. Ainsi, tout vecteur de Ω s'écrit de manière unique comme somme d'un vecteur de Ω^{quanti} et d'un vecteur de Ω^{quali} :

$$\forall x \in \Omega, \exists! (x^{quanti}, x^{quali}) \in \Omega^{quanti} \times \Omega^{quali}, \quad x = x^{quanti} + x^{quali}$$

où x^{quanti} et x^{quali} sont respectivement l'ensemble des composantes quantitatives et qualitatives de x . Dans le cas particulier de la base DAMIR, Ω est de dimension $p = 55$, Ω^{quanti} de dimension $q = 14$ et Ω^{quali} de dimension $p - q = 41$.

Partant de cette décomposition, nous définissons sur Ω l'application suivante :

$$d : \Omega \times \Omega \longrightarrow [0; +\infty[$$

$$(x, y) \longmapsto d(x, y) = d_e(x^{quanti}, y^{quanti}) + d_h(x^{quali}, y^{quali})$$

où d_e est la distance euclidienne sur Ω^{quanti} , d_h la distance de *Hamming* sur Ω^{quali} et (x, y) un couple de vecteurs ligne dans l'espace Ω des variables. Cette application d est une distance sur Ω .

Preuve : soient $x, y, z \in \Omega$,
— *Séparation*

Montrons que $x = y \implies d(x, y) = 0$.

$$\begin{aligned} x = y \implies d(x, y) &= d(x, x) = d_e(x^{quanti}, x^{quanti}) + d_h(x^{quali}, x^{quali}) \\ &= \sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - x_{j^{quanti}})^2} + \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq x_{j^{quali}}} = 0 \end{aligned}$$

d'où $x = y \implies d(x, y) = 0$

Montrons que $d(x, y) = 0 \implies x = y$.

$$\begin{aligned} d(x, y) = 0 &\implies d_e(x^{quanti}, y^{quanti}) + d_h(x^{quali}, y^{quali}) = 0 \\ \implies d_e(x^{quanti}, y^{quanti}) &= 0 \text{ et } d_h(x^{quali}, y^{quali}) = 0 \text{ car } d_e \text{ et } d_h \text{ sont positives par définition.} \\ \implies \sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - y_{j^{quanti}})^2} &= 0 \text{ et } \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} = 0 \\ \implies (x_{j^{quanti}} - y_{j^{quanti}})^2 &= 0 \quad \forall j^{quanti} \text{ et } \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} = 0 \quad \forall j^{quali} \\ \implies x_{j^{quanti}} &= y_{j^{quanti}} \text{ et } x_{j^{quali}} = y_{j^{quali}} \\ \text{d'où } x &= (x^{quanti}, x^{quali}) = (y^{quanti}, y^{quali}) = y \\ \text{donc } d(x, y) &= 0 \implies x = y. \end{aligned}$$

Comme $x = y \implies d(x, y) = 0$ et $d(x, y) = 0 \implies x = y$ alors $d(x, y) = 0 \iff x = y$.

— *Symétrie*

$$\begin{aligned} d(x, y) &= d_e(x^{quanti}, y^{quanti}) + d_h(x^{quali}, y^{quali}) \\ &= \sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - y_{j^{quanti}})^2} + \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} \\ &= \sqrt{\sum_{j^{quanti}} (y_{j^{quanti}} - x_{j^{quanti}})^2} + \sum_{j^{quali}} \mathbf{1}_{y_{j^{quali}} \neq x_{j^{quali}}} \\ &= d_e(y^{quanti}, x^{quanti}) + d_h(y^{quali}, x^{quali}) \end{aligned}$$

d'où

$$d(x, y) = d(y, x)$$

— *Inégalité triangulaire*

Montrons que $d(x, z) \leq d(x, y) + d(y, z)$.

$$d(x, z) = d_e(x^{quanti}, z^{quanti}) + d_h(x^{quali}, z^{quali})$$

$$d(x, z) \leq \sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - z_{j^{quanti}})^2} + \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}}$$

On a

$$\sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - z_{j^{quanti}})^2} \leq \sqrt{\sum_{j^{quanti}} (|x_{j^{quanti}} - y_{j^{quanti}}| + |y_{j^{quanti}} - z_{j^{quanti}}|)^2}$$

"inégalité de Minkowski"

$$\begin{aligned} &\leq \sqrt{\sum_{j^{quanti}} (x_{j^{quanti}} - y_{j^{quanti}})^2} + \sqrt{\sum_{j^{quanti}} (y_{j^{quanti}} - z_{j^{quanti}})^2} \\ &= d_e(x^{quanti}, y^{quanti}) + d_e(y^{quanti}, z^{quanti}) \end{aligned}$$

Aussi $\forall j^{quali}$,

$$\begin{aligned} x_{j^{quali}} = z_{j^{quali}} &\Rightarrow \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} = 0 \leq \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} + \mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}}, \\ x_{j^{quali}} \neq z_{j^{quali}} &\Rightarrow \begin{cases} \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} = 1 \leq \underbrace{\mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}}}_0 + \underbrace{\mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}}}_1 & \text{si } x_{j^{quali}} = y_{j^{quali}}, \\ \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} = 1 \leq \underbrace{\mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}}}_1 + \underbrace{\mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}}}_0 & \text{si } z_{j^{quali}} = y_{j^{quali}}, \\ \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} = 1 \leq \underbrace{\mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}}}_1 + \underbrace{\mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}}}_1 & \text{si } x_{j^{quali}} \neq y_{j^{quali}} \text{ et } z_{j^{quali}} \neq y_{j^{quali}} \end{cases} \end{aligned}$$

$$\text{donc } \forall j^{quali} \quad \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} \leq \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} + \mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}}.$$

d'où

$$\begin{aligned} \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq z_{j^{quali}}} &\leq \sum_{j^{quali}} \mathbf{1}_{x_{j^{quali}} \neq y_{j^{quali}}} + \sum_{j^{quali}} \mathbf{1}_{y_{j^{quali}} \neq z_{j^{quali}}} \\ &= d_h(x^{quali}, y^{quali}) + d_h(y^{quali}, z^{quali}) \end{aligned}$$

On obtient ainsi,

$$\begin{aligned} d(x, z) &\leq d_e(x^{quanti}, y^{quanti}) + d_e(y^{quanti}, z^{quanti}) + d_h(x^{quali}, y^{quali}) + d_h(y^{quali}, z^{quali}) \\ &= d_e(x^{quanti}, y^{quanti}) + d_h(x^{quali}, y^{quali}) + d_e(y^{quanti}, z^{quanti}) + d_h(y^{quali}, z^{quali}) \\ &= d(x, y) + d(y, z) \end{aligned}$$

On a donc $d(x, z) \leq d(x, y) + d(y, z)$

L'application d vérifie la séparation, la symétrie et l'inégalité triangulaire, elle est donc une distance sur Ω . C'est cette distance que nous utiliserons pour la mise en œuvre du modèle kNN .

L'idée du modèle kNN est d'identifier les k lignes d'observations plus proches de la ligne présentant la valeur absente à imputer. La valeur d'imputation est déterminée en prenant la moyenne des observations correspondantes des k plus proches voisins. Considérons par exemple une table de données de la forme suivante :

$y_{1,1}$	\cdots	$y_{1,j-1}$	$y_{1,j}$	$y_{1,j+1}$	\cdots	$y_{1,p}$
\vdots			\vdots			\vdots
$y_{i-1,1}$	\cdots	$y_{i-1,j-1}$	$y_{i-1,j}$	$y_{i-1,j+1}$	\cdots	$y_{i-1,p}$
$y_{i,1}$	\cdots	$y_{i,j-1}$	NA	$y_{i,j+1}$	\cdots	$y_{i,p}$
$y_{i+1,1}$	\cdots	$y_{i+1,j-1}$	$y_{i+1,j}$	$y_{i+1,j+1}$	\cdots	$y_{i+1,p}$
\vdots			\vdots			\vdots
$y_{n,1}$	\cdots	$y_{n,j-1}$	$y_{n,j}$	$y_{n,j+1}$	\cdots	$y_{n,p}$

Chaque ligne de cette table peut être vue comme un vecteur à p composantes dans l'espace Ω des variables (les vecteurs colonnes). Dans cet ensemble de données, la ligne i présente une valeur manquante. En supprimant la colonne j , toutes nos lignes passent en dimension $p - 1$, la ligne i n'a plus d'observation manquante, on peut alors déterminer la distance qui la sépare des autres lignes de la base afin d'identifier ses k plus proches voisins. La valeur d'imputation est :

$$y_{i,j}^* = \frac{1}{k} \sum_{i^k} y_{i^k,j}$$

Il s'agit de la moyenne des j^{eme} observations des k lignes les plus proches de i au sens de la métrique d que nous avons défini.

La démarche que nous venons de présenter lorsque la ligne considérée présente une seule valeur manquante est la même lorsqu'il y a plus d'une valeur manquante à imputer. Si q est ce nombre de valeurs manquantes, il suffit de se ramener à un espace de dimension $p - q$ dans lequel la ligne d'intérêt ne présente pas de valeurs manquantes. Cela permet de calculer les distances qui la sépare des autres lignes de la base dans l'espace de dimension $p - q$.

Remarque : En fonction du taux de valeurs manquantes et de leur répartition dans la base de données, la méthode des kNN peut être plus ou moins coûteuse en temps. En résumé, l'algorithme d'imputation est le suivant :

- Fixer un entier $k : 1 \leq k \leq n$ (dans notre modèle $k = 10$).
- Choisir une métrique d .
- Calculer les distances $d(i, l); \forall l \neq i$
- Retenir les observations j des k lignes les plus proches de i .
- Affecter à la valeur manquante la moyenne des k observations j .

6.1.6 Méthode *MissForest*

La méthode *MissForest* est une méthode d'imputation des valeurs manquantes par les forêts aléatoires (Breiman [2001]). Tout comme les modèles *LOCF* et kNN , le modèle *MissForest* n'a pas de restriction en matière de types de données. Il peut être utilisé sur une base de données mixte (quantitative et qualitative) et tient compte des liaisons qui existent entre toutes les variables. Cette méthode tolère la présence de variables dépendantes dans le jeu de données.

Partant d'un échantillon d'apprentissage, E_n (l'ensemble des valeurs observées) et d'une méthode de prédiction (par exemple une prédiction par arbre de régression) qui permet de définir sur E_n un prédicteur $\hat{p}(E_n)$, le principe permettant de construire une forêt aléatoire, consiste à générer B échantillons bootstrap $E_n^{b_1}, \dots, E_n^{b_B}$ et d'appliquer la méthode de prédiction à chaque échantillon, de façon à obtenir la suite de prédicteurs $(\hat{p}(E_n^{b_l}))_{l \in [1:B]}$. La construction de la forêt se termine par l'agrégation des B estimateurs.

Lorsque la suite de prédicteurs est quantitative, l'agrégation se fait en considérant la moyenne des prédicteurs

$$\hat{p}_{RF} = \frac{1}{B} \sum_{l=1}^B \hat{p}(E_n^{b_l})$$

Lorsque la suite de prédicteurs est qualitative, l'agrégation des B estimateurs se fait par vote majoritaire, c'est-à-dire qu'on retient la qualité la plus représentée

$$\hat{p}_{RF} = p^*, \sum_{l=1}^B \mathbf{1}_{\hat{p}(E_n^{b_l})=p^*} \geq \sum_{l=1}^B \mathbf{1}_{\hat{p}(E_n^{b_l})=p} \quad \forall p \in (\hat{p}(E_n^{b_l}))_{l \in [1:B]}$$

A la différence des modèles *CART* que nous présentons dans la 3^e partie de ce mémoire, le modèle *MissForest* ne nécessite pas de fixer à l'avance un critère d'arrêt. L'algorithme qui définit ce modèle s'arrête lorsque, suite à l'ajustement d'un nombre important de forêt aléatoire (ensemble d'arbres de régression), on obtient une dégradation du modèle. Cette dégradation est donnée par l'erreur *Out-Of-Bag* (*OOB*). Lorsque la base ne contient qu'un seul type de données, l'algorithme s'arrête dès la première augmentation de l'erreur *OOB*. La prédiction obtenue avant cet arrêt est celle qui sera utilisée pour la complétion des données manquantes. Lorsque la base de données est mixte, l'algorithme s'arrête lorsque chacune des deux erreurs *OOB* que nous notons *OOB^{quanti}* et *OOB^{quali}* (pour les ensembles de variables quantitatives et qualitatives respectivement) augmentent pour la première fois.

Pour une variable quantitative X de taille n et sa prédiction X^{imp} , l'erreur *OOB^{quanti}* est obtenue par

$$OOB^{quanti} = \sqrt{\frac{\frac{1}{n} \sum_{i=1}^n (X_i^{imp} - X_i)^2}{var(X)}}$$

Lorsque X est qualitative, l'erreur *OOB^{quali}* est donnée par

$$OOB^{quali} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i^{imp} \neq X_i}$$

6.2 Comparaison et choix des modèles

Les méthodes d'imputation des données manquantes que nous avons présenté dans les sections précédentes vont être testées sur un échantillon de données issu d'une extraction d'un peu plus de 1 *Md* (1 119 235 004) de lignes complètes (sans valeurs manquantes) de notre base d'étude.

6.2.1 Échantillonnage des données

Pour Jean Vaillant [2005], «La définition d'échantillon représentatif diffère selon que le plan d'échantillonnage est probabiliste ou non probabiliste :

- Un plan probabiliste fournit un échantillon représentatif dès lors que chaque individu de la population a une probabilité connue et non nulle d'être inclus dans l'échantillon.
- Un plan non probabiliste fournit un échantillon représentatif si la structure de l'échantillon pour certaines variables clés est similaire à celle de la population cible. Par exemple, on peut vouloir construire un échantillon pour lequel les proportions de catégories d'individus soient similaires dans l'échantillon à celles de la population cible (c'est le principe de la méthode dite des quotas)» [37].

Cette méthode des quotas est celle qui a guidé le choix de notre démarche d'échantillonnage. L'objectif des données DAMIR étant d'expliquer les dépenses de santé, nous considérons qu'un échantillon est représentatif s'il conserve la moyenne et la variance de la variable *flt_pai_mnt* qui représente les montants de dépenses des prestations préfiltrées. En notant par $Y = (Y_i)_{i \in [1:N]}$ cette variable, sa moyenne est donnée par $\mu = \frac{1}{N} \sum_{i=1}^N Y_i$ et sa variance par $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2$, avec $N = 1\text{ Md}$.

La moyenne d'un échantillon $y = (y_1, \dots, y_n)$ de Y est $\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$. Sous l'hypothèse que les Y_i sont indépendants et de même loi, lorsqu'on augmente la taille n de l'échantillon la loi des grands nombres garantit la convergence de \bar{y}_n vers l'espérance mathématique de Y qui a μ pour estimateur sans biais. Ainsi, $\bar{y}_n \xrightarrow[n \rightarrow N]{} \mu$ et par le théorème centrale limite, $\frac{\bar{y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[n \rightarrow N]{} N(0, 1)$. L'intervalle

$$IC_{1-\alpha}(\bar{y}_n) = [\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

est celui dans lequel doit se trouver \bar{y}_n pour que l'échantillon de taille n considéré soit représentatif (au sens du critère d'égalité des moyennes) avec une erreur α ($\alpha \in [0; 1]$ et $z_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite).

Partant de l'échantillon $y = (y_1, \dots, y_n)$ de Y , on estime σ^2 par $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2$. En admettant que Y est gaussienne (ce qui induit les sous suites y de Y sont également gaussiennes), on a

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}_n}{\sigma} \right)^2 \stackrel{\text{Loi}}{=} \frac{\sigma^2}{n-1} \chi_{(n-1)}^2$$

et $E(S_n^2) = \frac{\sigma^2}{n-1} E(\chi_{(n-1)}^2) = \sigma^2$. S_n^2 est un estimateur sans biais de σ^2 . Avec les quantiles d'ordre $\alpha/2$ et d'ordre $1 - \alpha/2$ ($\alpha \in [0; 1]$) de la loi $\chi_{(n-1)}^2$ respectivement notés $q_{\alpha/2}$ et $q_{1-\alpha/2}$ on obtient

$$P(q_{\alpha/2} \leq \frac{n-1}{\sigma^2} S_n^2 \leq q_{1-\alpha/2}) = 1 - \alpha$$

$$P(q_{\alpha/2} \frac{\sigma^2}{n-1} \leq S_n^2 \leq q_{1-\alpha/2} \frac{\sigma^2}{n-1}) = 1 - \alpha$$

On a ainsi l'intervalle

$$IC_{1-\alpha}(S_n^2) = [q_{\alpha/2} \frac{\sigma^2}{n-1}; q_{1-\alpha/2} \frac{\sigma^2}{n-1}]$$

dans lequel doit se trouver S_n^2 pour que l'échantillon de taille de n correspondant soit considéré comme représentatif au sens de l'égalité des variances des S_n^2 et σ^2 avec une erreur α .

Mise en œuvre du procédé d'échantillonnage

Partant de notre base complète de données, nous avons extrait des échantillons de tailles de plus en plus importantes. Chaque échantillon est obtenu par un tirage aléatoire sans remise des lignes de la base complète, elles ont toutes la même probabilité d'être sélectionnées. Ces tirages

sont réalisés sous **Sas** à l'aide de la méthode *SRS* (*Simple Random Sampling*) de la procédure *SURVEYSELECT*.

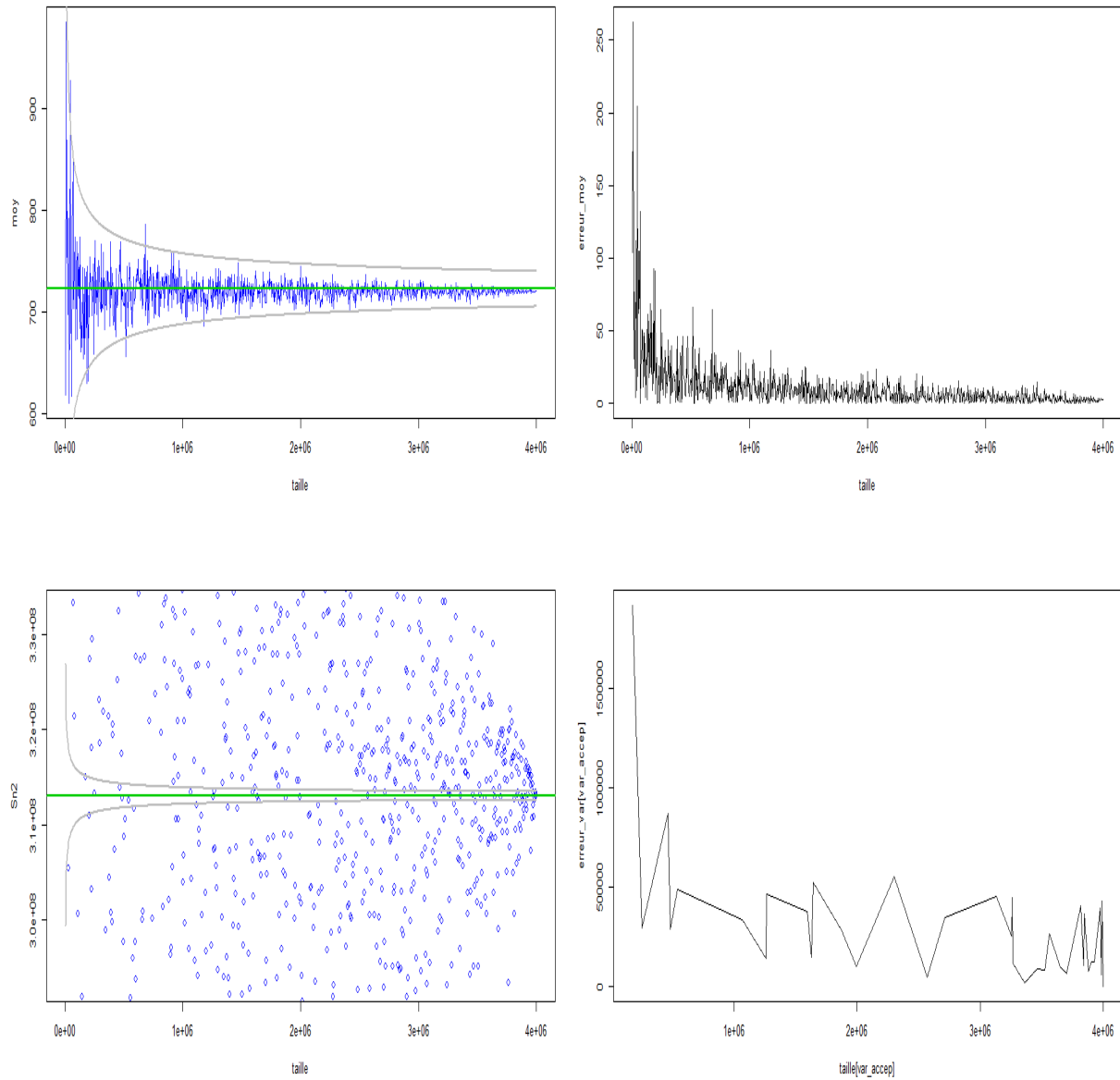


FIGURE 6.1 – En haut estimation de la moyenne de l'échantillon (à gauche moyenne et intervalle de confiance à 95%, à droite erreur d'estimation). En bas estimation de la variance de l'échantillon (à gauche variance et intervalle de confiance à 95%, à droite erreur d'estimation des 38 échantillons acceptables).

Nous avons réalisé 1000 échantillons de la taille n allant de 4 000 à 4 000 000 par pas de 4000. Pour chaque échantillon nous estimons \bar{y}_n , $IC_{1-\alpha}(\bar{y}_n)$, S_n^2 et $IC_{1-\alpha}(S_n^2)$. Les résultats de ces estimations sont donnés par la figure 6.1 :

- En haut à gauche, nous avons en vert la valeur $\mu = 722,156$ vers laquelle converge la suite $(\bar{y}_n)_n$ (en bleu) lorsqu'on augmente la taille n de l'échantillon. Pour presque toutes les valeurs de n les moyennes \bar{y}_n appartiennent à l'intervalle de confiance $IC_{1-\alpha}(\bar{y}_n)$ représenté en gris, ce qui garantit le respect du critère d'égalité des moyennes (avec une erreur de 5%) pour presque tous les échantillons sélectionnés.

- En haut à droite, nous avons représenté (en noir) l'évolution des écarts absolus entre \bar{y}_n et la moyenne $\mu = 722,156$ lorsque n devient grand. Ces écarts diminuent avec l'augmentation de la taille n de l'échantillon, ce qui signifie que plus n est grand, plus l'échantillon de taille n est représentatif au sens du critère d'égalité des moyennes.
- En bas à gauche, nous avons en vert la valeur $\sigma^2 = 313\,051\,765$ et en bleu les estimations S_n^2 de σ^2 . Ces estimations sont majoritairement à l'extérieur de $IC_{1-\alpha}(S_n^2)$ (en gris) : parmi les 1000 estimations, 962 valeurs n'appartiennent pas à $IC_{1-\alpha}(S_n^2)$. Nous avons donc obtenu 38 échantillons représentatifs au sens du critère d'égalité des variances avec une erreur de 5%.
- En bas à droite, nous avons représenté (en noir) les écarts absolus entre les 38 valeurs acceptables de S_n^2 et la variance $\sigma^2 = 313\,051\,765$. Bien que l'écart le plus faible soit donné par l'échantillon de taille 4 000 000, nous retenons parmi les 38 échantillons acceptables celui de taille la plus petite ($n = 176000$).

Avec une taux d'erreur $\alpha = 5\%$, nous considérons comme représentatif l'échantillon de taille $n = 176000$ pour lequel nous avons les suivantes statistiques de la variable *flt_pai_mnt* :

- $\bar{y}_n = 730,8229 \in IC_{1-\alpha}(\bar{y}_n) = [639,4950 ; 804,8167]$
- $S_n^2 = 314\,968\,911 \in IC_{1-\alpha}(S_n^2) = [310\,986\,786 ; 315\,123\,483]$

6.2.2 Méthode de sélection du modèle

L'échantillon de données que nous avons construit dans la sous-section précédente va être utilisé pour sélectionner le modèle qui sera utilisé pour la complétion des données manquantes de notre base. Sur notre échantillon de données, nous allons créer des données manquantes artificielles. Ces données artificiellement manquantes seront complétées par les valeurs générées par les modèles. Le modèle à retenir s'obtiendra en comparant les résultats de la complétion avec les données retirées.

Afin de choisir le modèle le plus adapté pour la complétion, les valeurs manquantes sont artificiellement créées dans les mêmes proportions que les données manquantes de la base DAMIR et ses données sont portées par les mêmes variables (cf. table 5.2). La figure 6.2 donne une représentation des données des valeurs manquantes que nous avons créées : 63% de lignes sont complètes et 38% présentent au moins une valeur manquante. Ce sont des valeurs manquantes non monotones. Nous renvoyons le lecteur intéressé par les types de répartition des valeurs manquantes à *Imputation de données manquantes* [10].

Remarque : La majorité des modèles de complétion privilégient les données quantitatives. Parmi les modèles que nous avons présenté, les méthodes *LOCF*, *kNN* et *MissForest* permettent d'imputer les variables qualitatives et quantitatives alors que les méthodes *moyenne*, *médiane* et *LOESS* sont uniquement utilisées pour des données quantitatives.

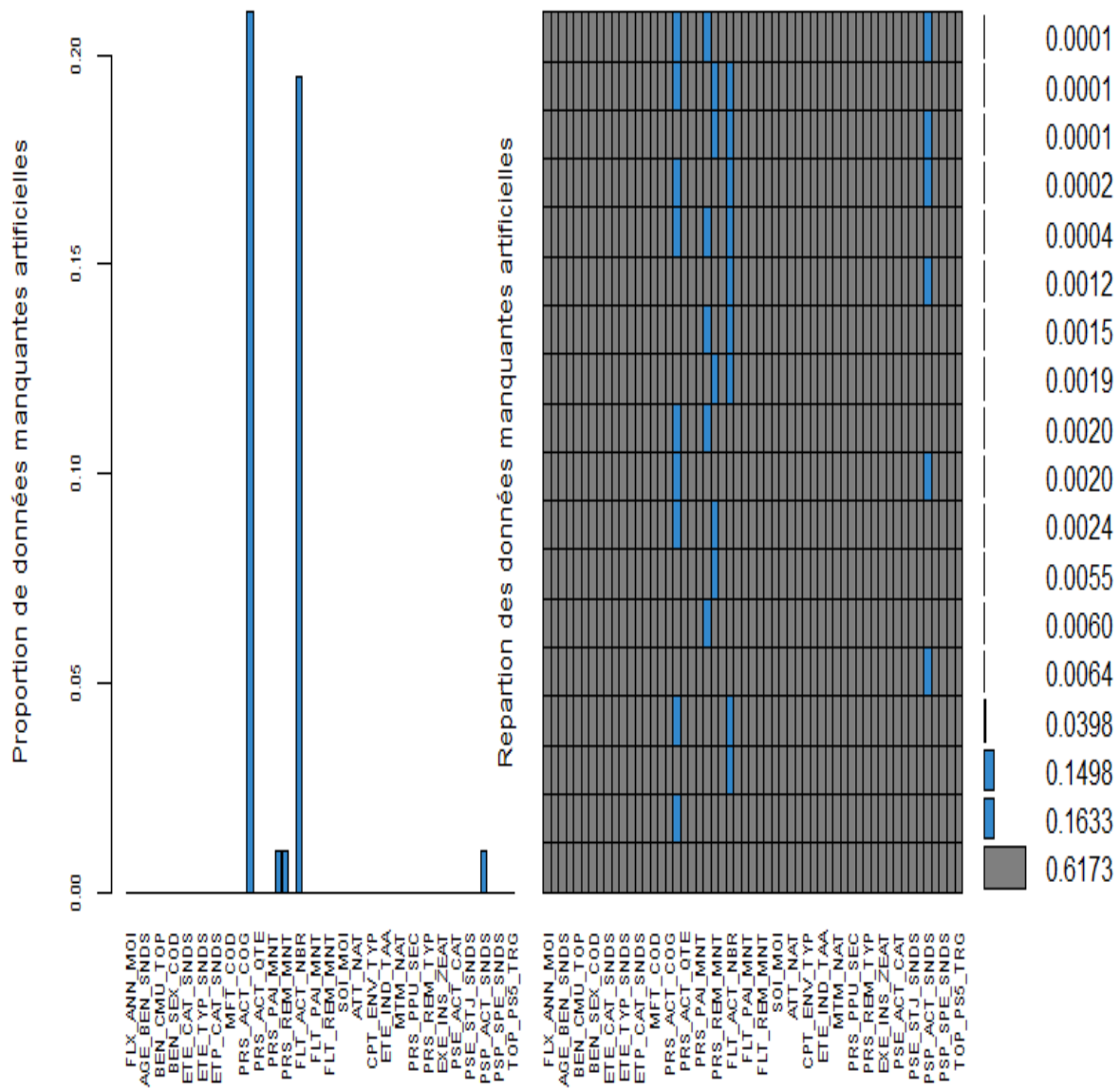


FIGURE 6.2 – A gauche, la proportion des valeurs manquantes. A droite, la répartition des données de l'échantillon (en bleu les valeurs manquantes, en gris les valeurs observées).

Résultat de l'imputation

La mise en œuvre des 6 modèles d'imputations de données manquantes présentés nous a fourni les résultats donnés par le tableau 6.2. Pour chacune des variables à compléter, ce tableau donne une statistique qui quantifie la qualité de la complétion obtenue par les différentes méthodes envisagées. Le meilleur modèle est celui qui donne la statistique la plus faible.

Pour les variables quantitatives *prs_act_nbr* (dénombrement des prestations), *flt_act_nbr* (dénombrement des prestations préfiltrées), *prs_rem_bse* (base de remboursement), *prs_rem_mnt* (montant versé/remboursé) les statistiques sont données par la moyenne des écarts absolus entre les données réelles et les données de complétion :

$$d_1(X^{imp}, X) = \frac{\sum_{i^{imp}} |X_{i^{imp}}^{imp} - X_{i^{imp}}|}{N^{imp}}$$

où N^{imp} le nombre de données à compléter, X^{imp} les données de complétion et X les données réelles.

Pour la variable qualitative *psp_act_snds* (nature d'activité du professionnel de santé prescripteur) les statistiques sont données par la distance de *Hamming* normalisée par le nombre de valeurs à imputer.

$$d_2(X^{imp}, X) = \frac{\sum_{i^{imp}} \mathbf{1}_{X_{i^{imp}}^{imp} \neq X_{i^{imp}}}}{N^{imp}}$$

variable	locf	moyenne	médiane	kNN	loess	missForest
<i>prs_act_nbr</i>	35.80	754.61	17.53	15.84	28.02	4.26
<i>flt_act_nbr</i>	36.29	751.52	20.17	18.38	30.26	3.82
<i>prs_rem_bse</i>	929.44	721.48	217.20	207.94	520.28	14.63
<i>prs_rem_mnt</i>	1606.64	1052.58	552.57	514.61	805.82	74.68
<i>psp_act_snds</i>	0.08	/	/	0.03	/	0.01

TABLE 6.2 – Distance entre les données réelles et les données simulées pour la complétion. Pour les quatre premières variables (quantitatives) les distances sont obtenues par la moyenne des écarts absolus. Pour la cinquième variable (qualitative) les distances sont données par la métrique de Hamming normalisée.

Un regard sur le tableau 6.2 suggère que la méthode *MissForest* est la plus précise pour la complétion des valeurs manquantes de nos variables quantitatives. Pour l'unique variable qualitative présentant des valeurs manquantes, ce modèle reste le plus adapté.

Robustesse des méthodes de complétion

Les résultats précédents ont été obtenus sur un échantillon de données présentant en proportion la même quantité de valeurs manquantes (par variable) que notre jeu de données d'intérêt. Nous allons tester la pertinence de ces modèles en faisant varier la proportion des données manquantes de 10 à 80% pour chacune des 5 variables à imputer.

Les résultats obtenus sont donnés par la figure 6.3. Pour les variables *prs_act_nbr* et *flt_act_nbr* la méthode d'imputation par la moyenne est de loin la moins précise. Pour cette méthode, les valeurs obtenues par d_1 (valeurs données par l'axe vertical droit) sont en moyenne de 772 pour *prs_act_nbr* et de 765 pour *flt_act_nbr*. Ces valeurs sont très au dessus de celles relatives aux autres modèles pour lesquelles les statistiques données par d_1 varient entre 0 et 50 (axe vertical gauche). Pour ces deux variables, quelque soit la proportion de valeurs manquantes, le modèle le plus stable et le plus précis est le modèle *MissForest*. Ce modèle est également celui qui fournit la meilleure complétion pour les variables *prs_rem_bse* et *prs_rem_mnt*. Ainsi, nous retenons la méthode *MissForest* pour l'imputation des valeurs manquantes des variables quantitatives *prs_act_nbr*, *flt_act_nbr*, *prs_rem_bse* et *prs_rem_mnt*.

Les statistiques obtenues pour la variable qualitative *psp_act_snds* par la distance d_2 montrent que la méthode *MissForest* est moins stable que les méthodes *LOCF* et *kNN*. Toutefois, la méthode *MissForest* est la plus précise pour la complétion des valeurs manquantes de la variable

psp_act_snds. C'est cette méthode que nous retenons pour le traitement des valeurs manquantes de cette variable. Sous **SAS** l'imputation des valeurs manquantes par les forêts aléatoires s'effectue à l'aide de la *proc imstat* avec la méthode *RANDOMWOODS* et l'option *Impute*. Le parallèle de cette procédure en **R** est la fonction *MissForest* du package qui porte le même nom. Dans le cadre de ce mémoire, l'imputation a été réalisée sous **SAS** pour des raisons de volumétrie.

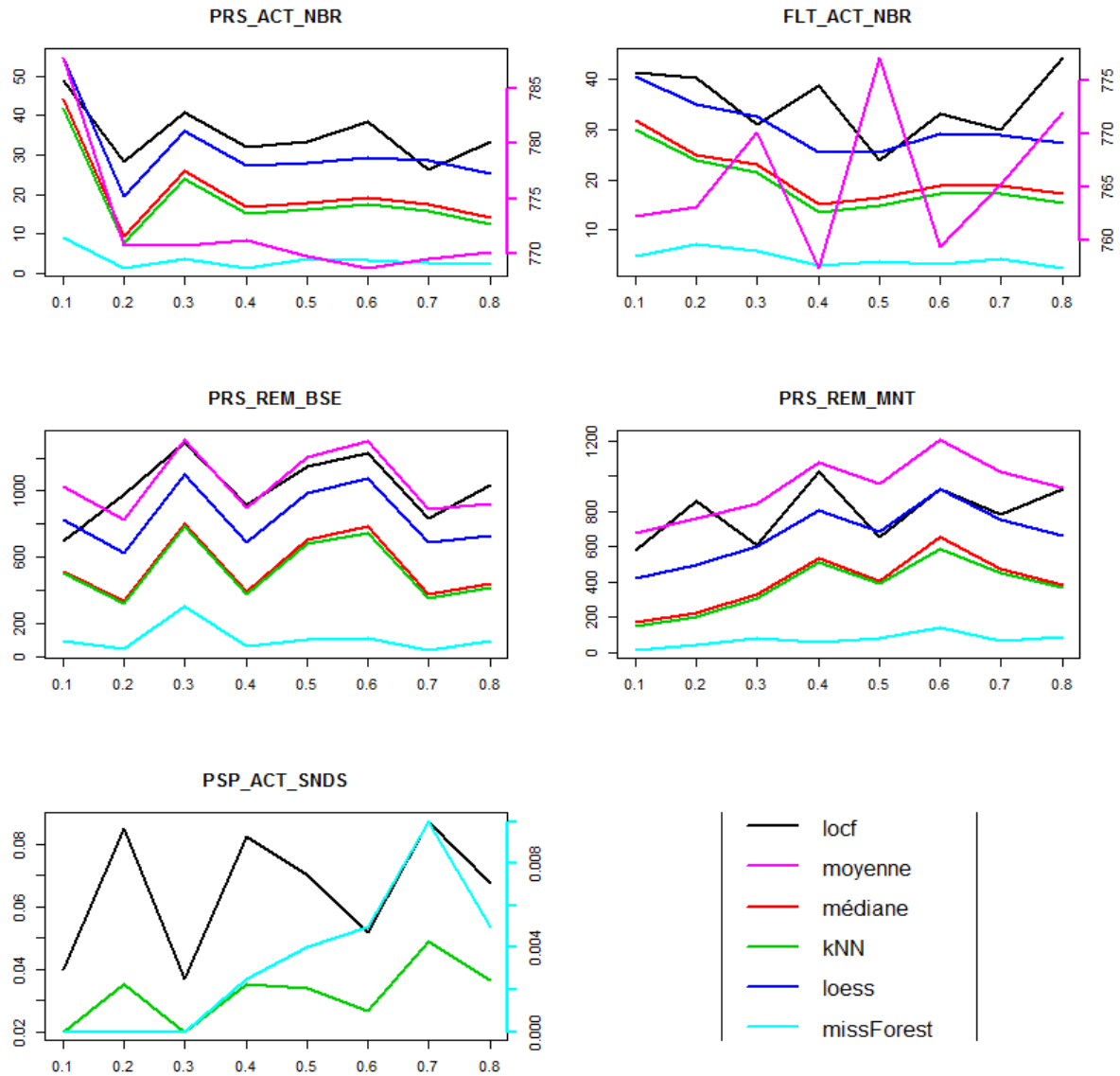


FIGURE 6.3 – Représentation graphique de la pertinence des méthodes de complétion lorsque la proportion de données manquantes augmente de 10 à 80% de la taille de l'échantillon.

Remarque : Les procédés de traitement des données manquantes que nous avons présentés dans ce chapitre ont été choisis pour leur tolérance en matière de colinéarité et normalité des données. Comme nous le verrons dans les chapitres suivants, la base de données mixtes (quantitatives et qualitatives) que nous étudions présente de fortes dépendances entre certaines paires de variables : il s'agit des paires d'indicateurs préfiltrées et non préfiltrées (*prs_act_nbr* et *flt_act_nbr* par exemple).

Les méthodes *EM* (*Expectation Maximisation*) et *MCMC* (*Monte Carlo par Chaîne de Markov*) présentées en annexe sont également très utilisées pour la complétion de données. Le lecteur intéressé par une présente détaillée de ses modèles peut se référer à *L'algorithme EM : une courte présentation* (Frédéric Santos [2015]) et à *Méthode de Monte Carlo par Chaînes DE Markov* (Christian Robert [1996]). Ces publications présentent respectivement les méthodes *EM* et *MCMC* de façon très précise. La majorité des logiciels statistiques qui proposent ces méthodes font l'hypothèse que les données sont gaussiennes et incorréllées, ce qui n'est généralement pas le cas dans la réalité. Le préalable à l'utilisation de ces modèles est donc une étude visant à déterminer la nature des distributions de probabilités qui offrent la meilleure adéquation aux données. Cette étape préalable permet de déterminer les formules des estimateurs calculés à chaque itération. Une implémentation de ces algorithmes sous **Ssas** est possible avec la *proc mi*.

6.3 Les données atypiques

Comme nous l'avons précisé dans les sections précédente, la base de données à laquelle nous nous intéressons est constituée de données mixtes. S'agissant des variables catégorielles, il est peu courant de qualifier l'une de leurs observations d'atypique. Cette difficulté est due aux problèmes que posent la définition d'une modalité de référence à laquelle les autres modalités doivent être comparées et le choix du critère de comparaison.

Il est toutefois courant de qualifier d'atypique les observations "valeur inconnue" et "sans objet" présentées par certaines variables qualitatives. Dans de nombreuses, études, il est courant de considérer ces observations comme des données manquantes et de leur imputer une valeur issue d'un modèle de complétion de données. Une autre méthode consiste à supprimer les lignes de données présentant une de ces deux observations, ce qui entraîne une perte d'information car leur présence dans la base constitue en elle même une information. Dans le cadre de ce mémoire nous choisissons de conserver ces observations.

Concernant les variables quantitatives, il est courant qu'une valeur atypique soit qualifiée d'aberrante. Les valeurs atypiques sont généralement des valeurs extrêmes. Il est important de noter qu'une valeur peut être extrême du fait de sa variabilité naturelle. Dans ce cas, la donnée est certes atypique, mais vraie. S'agissant des 14 variables quantitatives de notre base de données, elles se répartissent comme l'illustre la figure 6.4.

La représentation donnée par la figure 6.4 montre que les variables quantitatives de notre jeu de données présentent des valeurs extrêmes (valeurs au delà des frontières hautes des différentes boîtes à moustaches). Toutefois comme nous l'avons déjà souligné, la base DAMIR est construite par de nombreux regroupements dont le but est de rendre ses données anonymes. Certaines de ces grandes valeurs que nous observons sont donc le résultat des différents agrégations faites lors de la construction de la base. Qualifier toutes ces grandes valeurs d'aberrantes et les supprimer reviendrait à nous priver d'une importante masse d'informations. La figure 6.4 nous permet également de remarquer que toutes les variables quantitatives de notre base, sauf la variable *prs_rem_tau* (taux de remboursement de la sécurité sociale), présentent des valeurs négatives. Lorsqu'il est question de variables de dénombrement ou de prestation, les valeurs négatives peuvent représenter des corrections ou des régularisations dues par exemple à des sur-remboursements. Toutefois, la variables *prs_rem_tau* présente des observations supérieures à 100, ce qui signifierait que les actes de soins concernés ont été remboursés au delà du taux maximal de remboursement de la sécurité sociale : le taux de remboursement de la sécurité sociale varie entre 0% et 100% de la base de remboursement.

Remarque : Les valeurs prises par la variable *prs_rem_tau* sont en pourcentage de la base de

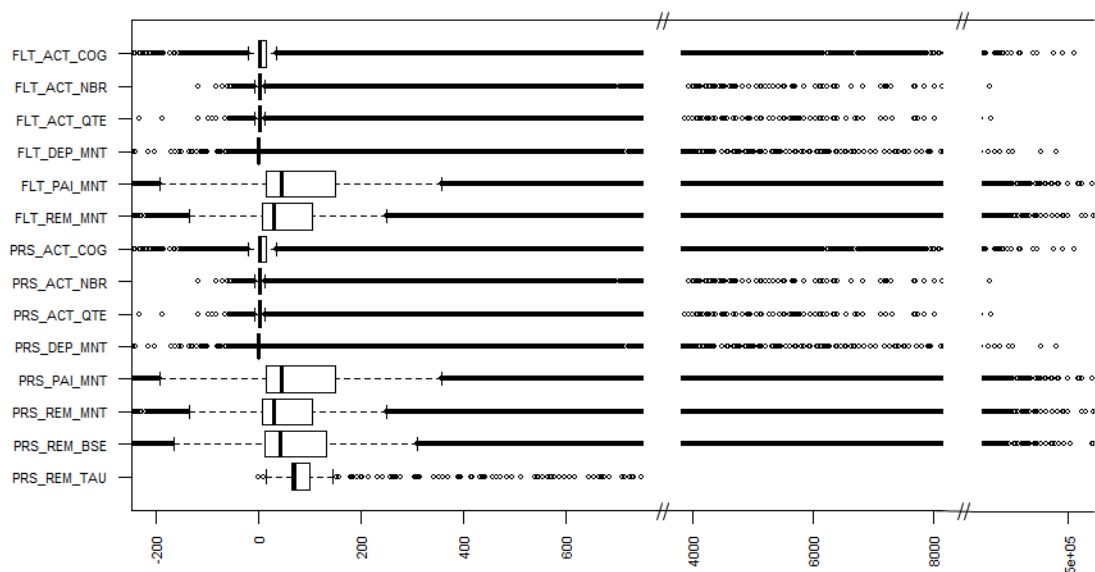


FIGURE 6.4 – Boîtes à moustaches des 14 variables quantitatives de la base de données.

remboursement de la sécurité sociale. Les observations de cette variable qui sont supérieures à 100 sont aberrantes. Ces observations ne sont pas à confondre avec des valeurs correspondantes à des sur-remboursements. Par exemple, pour le régime général, le taux de remboursement de la sécurité sociale pour la consultation d'un médecin généraliste est de 70% de la BR (base de remboursement). Un remboursement à 75% de la BR est un sur-remboursement alors qu'un remboursement à 102% est aberrant.

Nous avons ainsi considéré comme aberrantes les 4452 lignes de notre base de données qui présentent des taux de remboursement supérieures à 100% de la base de remboursement. Ces 4452 lignes représentent 0,00217 ‰ de notre jeu de données.

Chapitre 7

Statistiques descriptives de la base de données

La présentation statistique que nous proposons dans ce chapitre est réalisée à partir des données traitées : imputation des valeurs manquantes et suppression des valeurs aberrantes. Comme nous l'avons déjà souligné, la base DAMIR met en relation un nombre important de variables de nature diverse telles que le sexe du bénéficiaire, sa tranche d'âge, sa zone de résidence, la nature d'assurance, le montant des prestations, le montant du remboursement fait par la Sécurité Sociale, etc. En raison du grand volume de données, nous nous intéressons dans ce chapitre aux variables qui décrivent les bénéficiaires et aux indicateurs préfiltrés qui permettent une étude chiffrée des prestations de l'AMO (Assurance Maladie Obligatoire).

7.1 Description qualitative

Considérons une variable qualitative X pour laquelle nous avons n observations et m modalités. Soit x^i $i \in [1 : m]$, une modalité de X . Cette modalité a n_i occurrences sur l'ensemble des observations de x . Sa fréquence f_i est donnée par $f_i = \frac{n_i}{n}$, où $n = \sum_{i=1}^m n_i$.

La description d'une variable qualitative se fait généralement à l'aide d'un tableau (cf. table 7.1) qui précise ses différentes modalités, leurs effectifs et leurs fréquences.

x^i	n_i	f_i
-------	-------	-------

TABLE 7.1 – Forme du tableau descriptif d'une variable qualitative

Par habitude, les tableaux descriptifs sont accompagnés d'une représentation graphique sous forme :

- d'un **diagramme circulaire** (*camembert*) qui est un cercle divisé en secteurs dont les angles sont proportionnels aux fréquences. Pour une modalité x^i , l'angle α_i est défini par $\alpha_i = f_i \times 360^\circ$.
- d'un **diagramme en barre ou cylindrique** qui à une modalité x^i associe une barre ou un cylindre dont la hauteur h_i correspond à l'effectif n_i ou à la fréquence f_i . Ces diagrammes sont en général verticaux ou horizontaux.

Les données fournies par la base DAMIR concernent plusieurs risques d'assurance. Nous rappelons que ces données ont pour but d'expliquer les dépenses de santé de l'ensemble de la population française. Dans cette sous-section, nous nous intéressons aux différents risques de santé

expliqués par notre jeu de données, à l'âge, au sexe et à la région de résidence du bénéficiaire d'un acte de soin, à sa qualité (assuré, conjoint, enfant, etc) et aux types de couverture de santé dont ils bénéficient (bénéficiaires ou non de la CMU-C).

7.1.1 Nature d'assurance

<i>asu_nat</i>	2009	2010	2011	2012	2013	2014	2015	2016
<i>maladie</i>	205420	222550	226840	231550	240890	243280	331510	349890
<i>maternité</i>	4085,5	4439,1	4519,9	4661,8	4997,1	5049,8	6237,0	6577,1
<i>at et mp</i>	5176,0	5011,7	5080,5	4994,5	4961,6	4896,6	6119,7	6296,5
<i>décès</i>	4,0	3,9	3,9	3,9	3,8	3,5	3,7	3,6
<i>invalidité</i>	25,2	25,2	25,5	26,4	26,5	26,2	41,2	40,4
<i>prévention maladie</i>	511,2	520,3	548,5	516,9	516,6	495,2	3629,6	2166,6
<i>presta supplémentaire</i>	2708,8	2924,5	3146,2	2420,3	254,2	231,6	243,5	234,3
<i>valeur inconnue</i>	207,0	0,7	0,5	0,4	0,3	0,3	9,2	149,0

TABLE 7.2 – Effectif (en milliers) des différentes natures d'assurance par année de données

La base DAMIR s'intéresse principalement à l'assurance maladie, 95% des actes de santé qu'elle explique sont relatifs à la maladie (cf. figure 7.1).

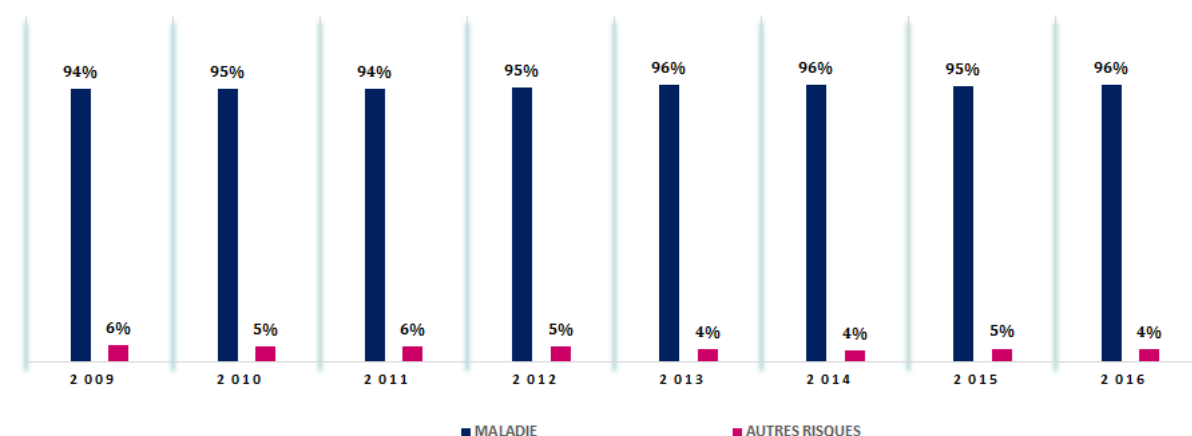


FIGURE 7.1 – Fréquences par année de données des différents risques (maladie et autres)

L'objet de ce mémoire étant l'assurance maladie, les statistiques présentées dans la suite de ce document concernent uniquement les données relatives au risque maladie.

7.1.2 Qualité du bénéficiaire

S'agissant de la qualité du bénéficiaire, en moyenne 78% des bénéficiaires de prestations d'assurance maladie sont les assurés principaux, 12% des bénéficiaires sont des conjoints, 9% des enfants et 1% les autres ayants droit (cf. figure 7.2).

<i>ben_qlt_cod</i>	2009	2010	2011	2012	2013	2014	2015	2016
<i>assuré</i>	159370	173340	176640	180730	188400	190530	261970	278660
<i>conjoint et assimilé</i>	25455	26758	26977	27010	27697	27879	36992	37371
<i>enfant</i>	18657	20168	20748	21240	22046	22007	28882	30446
<i>autre ayant-droit</i>	1931	2286	2468	2534	2722	2860	3661	3406
<i>inconnue</i>	4	3	11	37	22	4	1	5

TABLE 7.3 – Effectif (en milliers) des différentes qualités de bénéficiaires par année de données

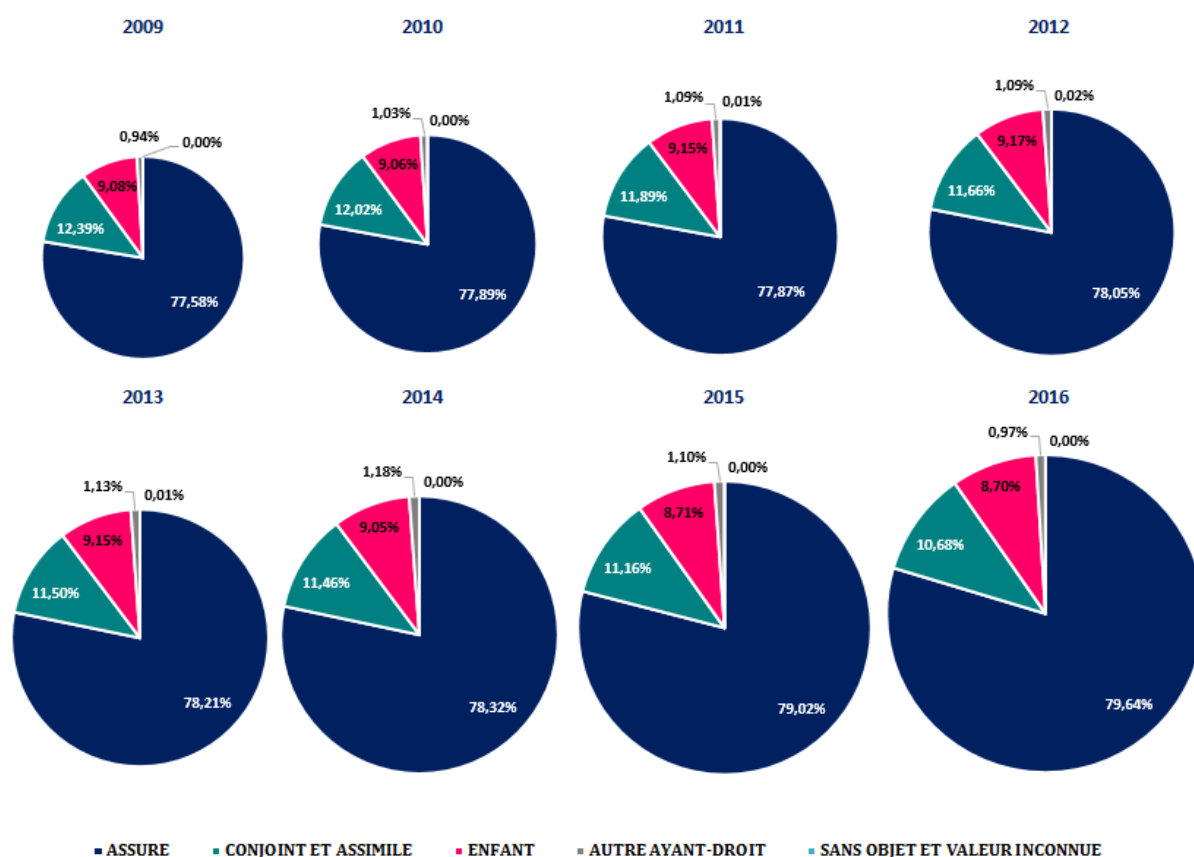


FIGURE 7.2 – Fréquences par année de données des différentes qualités de bénéficiaires

7.1.3 Âge du bénéficiaire

D'après les données DAMIR, en 2009 et 2010 la tranche d'âge des bénéficiaires de l'assurance maladie obligatoire la plus représentée était celle des 50 - 59 ans. Elle correspondait à 14,85% de bénéficiaires en 2009 et à 14,68% en 2010. De 2011 à 2016, la tranche d'âge la plus représentée était celle des 60 - 69 ans avec une proportion de bénéficiaires allant de 14,76% en 2011 à 16,25% en 2016. Quelque soit l'année, plus de la moitié des bénéficiaires ont entre 30 et 69 ans.

<i>age_ben_snds</i>	2009	2010	2011	2012	2013	2014	2015	2016
0 - 19 ans	20250	22752	23464	24188	25628	25795	33977	35959
20 - 29 ans	20741	24206	24885	25569	26856	26713	34872	36270
30 - 39 ans	22742	23684	24106	24392	25218	25390	33930	35432
40 - 49 ans	25960	27244	27651	28012	28972	29143	39569	41155
50 - 59 ans	30511	32671	32995	33354	34386	34688	47925	50364
60 - 69 ans	29525	32410	33482	34617	36384	37274	53008	56854
70 - 79 ans	28601	30275	30226	30542	31426	31730	43964	47290
80 ans et +	25743	28024	28756	29605	30819	31354	43186	46038
inconnu	1345	1289	1279	1272	1197	1193	1076	528

TABLE 7.4 – Effectif (en milliers) des différentes tranches d'âge par année de données

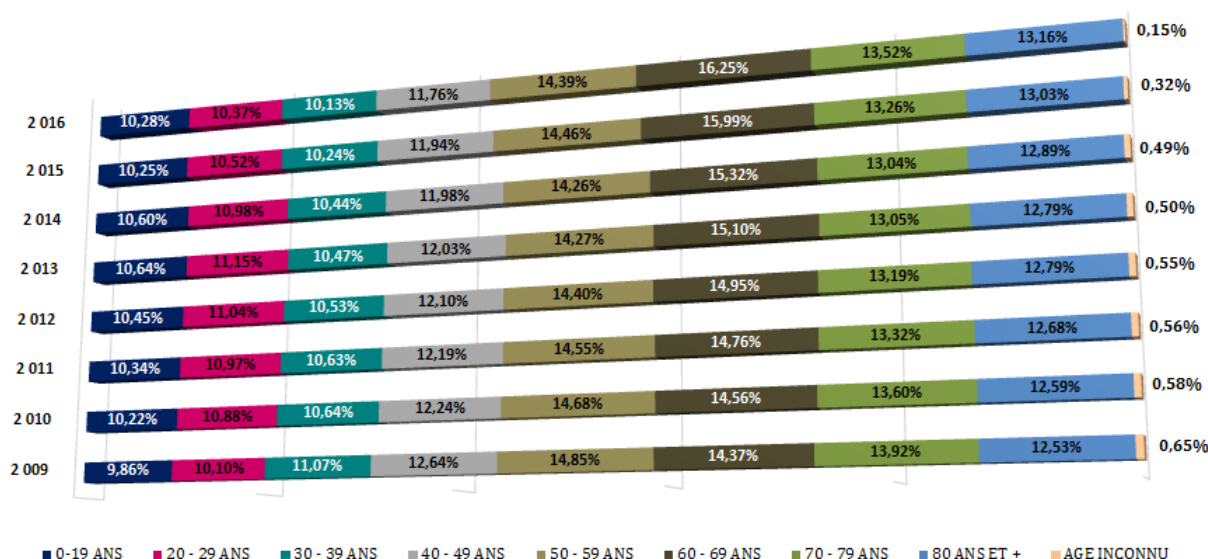


FIGURE 7.3 – Fréquences par année de données des différentes tranches d'âge de bénéficiaires

7.1.4 Couverture du bénéficiaire (CMU-C ou non)

<i>ben_cmu_top</i>	2009	2010	2011	2012	2013	2014	2015	2016
<i>non bénéficiaire de la cmu-c</i>	95964	104600	106350	109280	115130	115640	168200	177750
<i>bénéficiaire de la cmu-c</i>	15354	16308	16834	17367	18741	19853	28125	29519
<i>inconnue</i>	94099	101646	103659	104904	107016	107787	135181	142621

TABLE 7.5 – Effectif (en milliers) des différentes couverture de bénéficiaires par année de données

Le taux de bénéficiaires de la couverture maladie universelle complémentaire est de 7% entre 2009 et 2011. De 2012 à 2016, ce taux est de 8%.

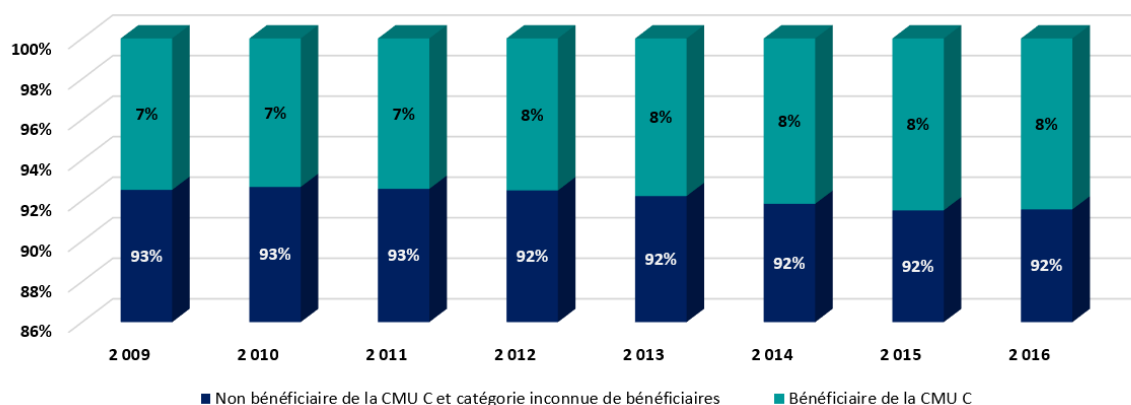


FIGURE 7.4 – Fréquences par année de données des différentes couvertures de bénéficiaires (CMU-C ou non)

7.1.5 Sexe du bénéficiaire

<i>ben_sex_cod</i>	2009	2010	2011	2012	2013	2014	2015	2016
<i>masculin</i>	89339	96539	98614	100420	104020	104910	144010	152940
<i>féminin</i>	116070	126010	128220	131010	135720	136370	185980	196940
<i>inconnu</i>	7	6	9	121	1147	2000	1516	9

TABLE 7.6 – Effectif (en milliers) des différents genres par année de données

En ce qui concerne le genre des bénéficiaires de l'assurance maladie, les femmes sont plus nombreuses que les hommes.

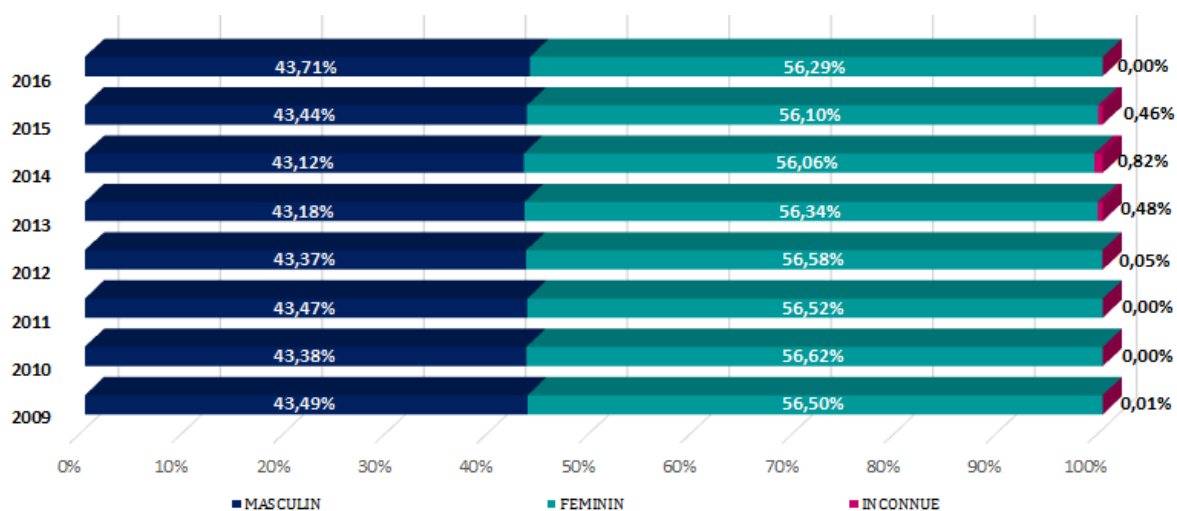


FIGURE 7.5 – Fréquences par année de données des différents genres de bénéficiaires

7.1.6 Région de résidence du bénéficiaire

Libellé ZEAT de Résidence du Bénéficiaire	2009	2010	2011	2012	2013	2014
Région Parisienne	30042	33342	33769	34594	35291	35215
Bassin Parisien	29507	31789	32424	33310	34644	35197
Nord	12815	13691	14174	14539	15000	15226
Est	20353	21950	22448	22891	24129	24616
Ouest	22051	23703	24214	24878	25891	25671
Sud-Ouest	21559	23729	24113	25128	26526	26877
Centre-Est	21581	23223	23803	24727	26140	26367
Méditerranée	27851	30319	30800	31383	32745	32976
Régions et Départements d'outre-mer	861	845	925	951	947	955
Inconnu	18796	19963	20175	19151	19573	20179

TABLE 7.7 – Effectif (en milliers) des différentes zones d'études et d'aménagement du territoire (ZEAT) par année de données (de 2009 à 2014)

Concernant la région de résidence des bénéficiaires, de 2009 à 2014, la Région Parisienne a la plus forte occurrence, les zones de résidence les moins représentées sont les Régions et Départements d'outre-mer (cf. tableau 7.7).

Les données 2015 et 2016 fournissent le même résultat : la région de résidence la plus représentée est l'Île-de-France (ou région parisienne) et les moins représentées sont les régions et Départements d'outre-mer (cf. figure 7.6).

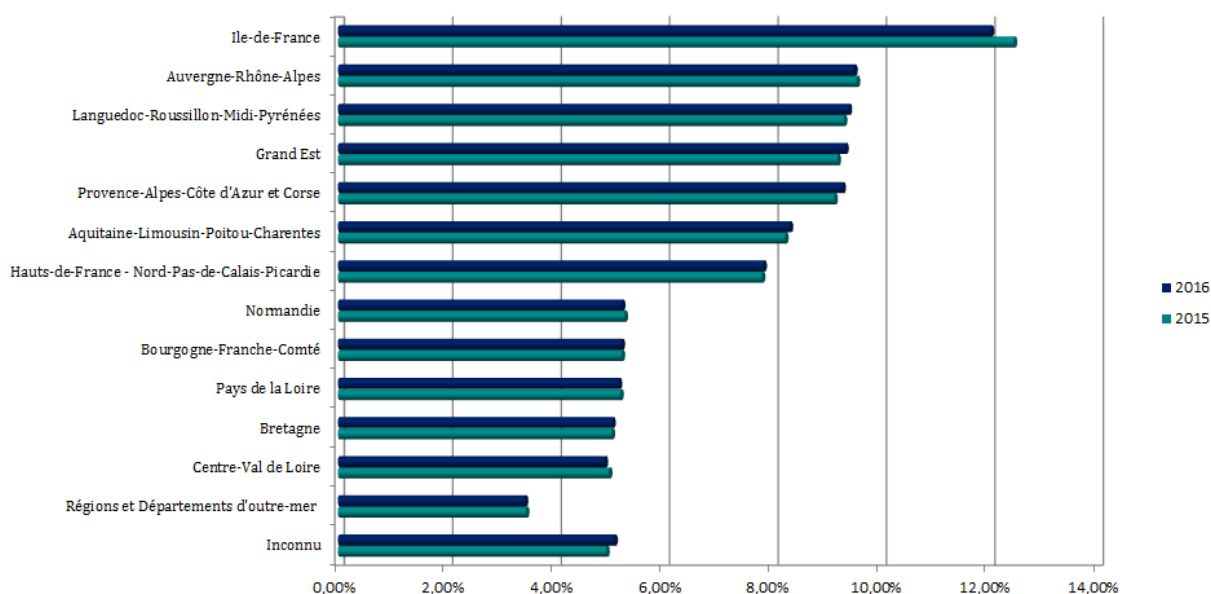


FIGURE 7.6 – Fréquences par année de données (2015 et 2016) des différentes régions de résidence des bénéficiaires

Région de Résidence du Bénéficiaire	2015	2016
Régions et Départements d'outre-mer	11499	12061
Ile-de-France	41319	42129
Centre-Val de Loire	16575	17212
Bourgogne-Franche-Comté	17348	18306
Normandie	17525	18330
Hauts-de-France - Nord-Pas-de-Calais-Picardie	25920	27459
Grand Est	30552	32735
Pays de la Loire	17265	18122
Bretagne	16733	17700
Aquitaine-Limousin-Poitou-Charentes	27343	29157
Languedoc-Roussillon-Midi-Pyrénées	30939	32959
Auvergne-Rhône-Alpes	31724	33322
Provence-Alpes-Côte d'Azur et Corse	30357	32558
Inconnu	16407	17841

TABLE 7.8 – Effectif (en milliers) des différentes régions de résidence des bénéficiaires par année de données (2015 et 2016)

La représentativité des régions de résidence des bénéficiaires de prestations d'assurance maladie donnée par la base DAMIR est cohérente avec la répartition régionale de la population française. En effet, les régions à fort effectif de population sont les plus représentées. La figure 7.7 donne la répartition par région des 66 millions de personnes résidant légalement sur le territoire français en 2015 et la représentativité des régions de résidence des bénéficiaires la branche maladie de la sécurité sociale pour cette même année. Nous observons une évolution quasi identique.

Afin de quantifier la liaison qui existe entre la répartition par région de la population et la représentativité des régions de résidence des personnes ayant bénéficié d'une prestation (en 2015), nous utilisons une mesure de corrélation des rangs : le *Tau de Kendall*. La définition la plus simple de cette mesure est

$$\tau = \frac{(\text{nombre de paires concordantes}) - (\text{nombre de paires discordantes})}{\frac{1}{2} \cdot n \cdot (n - 1)}$$

où n est le nombre total de paires. Dans le cas présent, il s'agit du nombre de régions ($n = 13$).

Deux paires (x_1, y_1) et (x_2, y_2) sont dites concordantes lorsque $x_1 < x_2$ et $y_1 < y_2$ ou $x_1 > x_2$ et $y_1 > y_2$. Elles sont dites discordantes lorsque $x_1 > x_2$ et $y_1 < y_2$ ou $x_1 < x_2$ et $y_1 > y_2$. Lorsque $x_1 = x_2$ ou $y_1 = y_2$, les paires ne sont ni concordantes, ni discordantes. Ainsi nous comptons 68 paires concordantes et 10 paires discordantes, soit un *Tau de Kendall* égal à $\tau_{2015} = 74,36\%$. De même, en considérant la répartition régionale de population française en 2016 et la représentativité des régions de résidence des bénéficiaires de prestations de l'assurance maladie fournie par les données DAMIR 2016, nous obtenons un *Tau de Kendall* égal au précédent ($\tau_{2016} = 74,36\%$). Ainsi, il existe une dépendance positive entre la représentativité des régions de résidence des bénéficiaires de prestations et la répartition régionale de la population française.

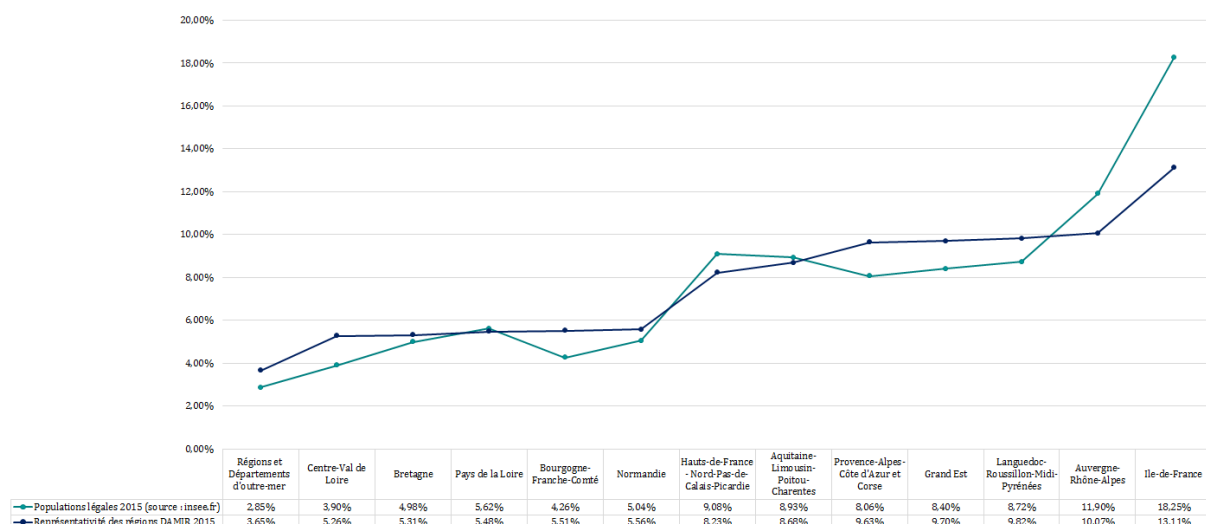


FIGURE 7.7 – Représentativité des régions de résidence des bénéficiaires pour l'année 2015 vs répartition par régionale de la population française en 2015

7.2 Description quantitative

La description proposée par cette section a pour but de chiffrer les dépenses de santé, par tranches d'âge, par genre ou par région. Nous rappelons que les différentes statistiques que nous donnons concernent uniquement l'assurance maladie obligatoire. Pour cela, les statistiques que nous donnons sont estimées à partir des variables préfiltrées :

- *flt_act_nbr* (dénombrement des prestations),
- *flt_pai_mnt* (montant de la dépense de prestations) : les frais réels (FR),
- *flt_rem_mnt* (montant versé/remboursé) : remboursement sécurité sociale (RSS).

Nous nous servirons également de la variable *prs_rem_bse* dont les observations correspondent aux bases de remboursements (BR ou BRSS) permettant d'estimer le montant versé par la sécurité sociale pour un acte de soin donné.

7.2.1 Les dépenses contre la maladie de 2009 à 2016

D'après les données DAMIR, en 2016 le total des dépenses contre la maladie étaient de 99,7 milliards d'euros, ce qui représente une hausse de 2,5% par rapport à 2015 et de 25,4% depuis à 2009 (cf. figure 7.8). Sur cette même période, la population française a cru de 3,7%, passant ainsi de 64,30 millions en 2009 à 66,69 millions en 2016. Les frais réels de maladie ont donc une croissance 6,8 fois plus importante que l'évolution de la population pour les 8 années considérées. Ce qui suggère qu'en plus de la croissance démographique, d'autres facteurs influent sur l'évolution des dépenses de santé.

Une différenciation par genre des frais réels de santé révèle que 66,4 milliards d'euros ont été dépensés en 2016 par les femmes, soit 1,13 fois les 58,7 milliards d'euros dépensés par les hommes cette même année. En 2009 les dépenses faites par les femmes étaient de 53,4 milliards d'euros contre 46,4 milliards par les hommes (cf. figure 7.9). Quels que soient le genre et l'année, les dépenses sont croissantes de 20 à 79 ans. Au-delà de 79 ans, cette croissance est maintenue pour les femmes mais pas pour les hommes. Cette baisse des dépenses masculines pourrait s'expliquer par le décès. De 2009 à 2012, les dépenses des femmes ont augmenté de 15,2%, une croissance plus forte que celle des dépenses masculines sur cette même période. Depuis 2013, la

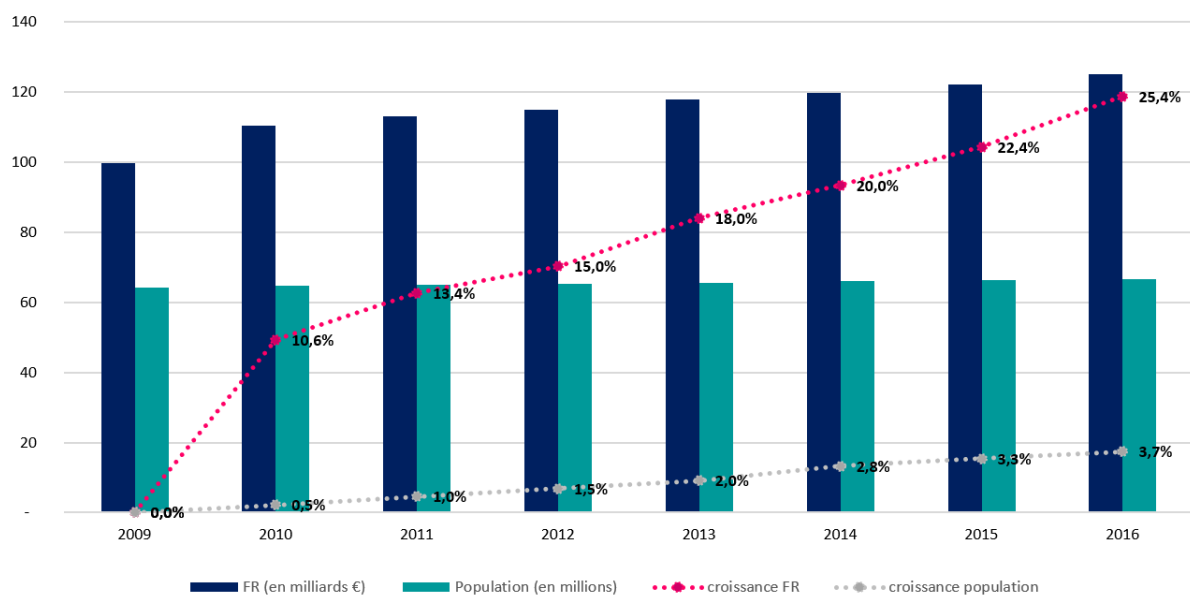


FIGURE 7.8 – Croissance des dépenses (en milliards d'euros) de maladie et hausse de la population française (en millions) de 2009 à 2015

hausse des dépenses des hommes est plus importante. De 2009 à 2016, les dépenses ont globalement augmenté de 26,6% pour les hommes, une hausse de 2,1 points par rapport à celles des femmes (+24,5%). Cette hausse des dépenses de santé est soutenue par de nombreux facteurs : la croissance démographique, le poids croissant des affections liées aux comportements et à l'environnement, le vieillissement de la population et des charges liées au progrès de la technologie médicale qui connaît depuis quelques années une accélération et une évolution remarquable.

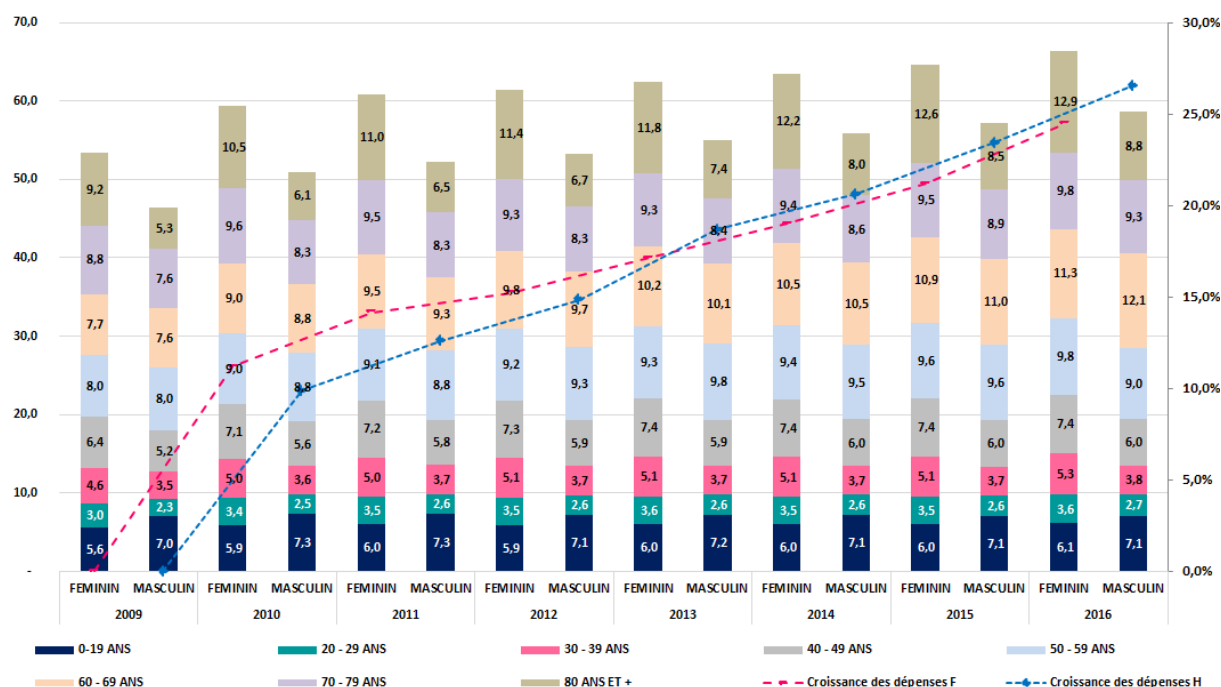


FIGURE 7.9 – Évolution des dépenses (en milliards d'euros) de maladie par genre et tranche d'âge de 2009 à 2016

Les dépenses de maladie varient également, fortement d'une région à l'autre. L'Île-de-France est la région qui affiche le plus important niveau de dépenses. En 2016, le montant total des

dépenses de cette région était de 20,22 milliards d'euros contre 19,82 milliards d'euros en 2015, soit une hausse de 2% (cf. figure 7.11). Avec ses 20,22 milliards de d'euros de frais de santé, l'île-de-France a dépensé en 2016 un peu plus de 4,7 fois que le Centre-Val de Loire et un peu 5,1 fois les Régions et départements d'outre-mer. En France métropolitaine, de 2015 à 2016, la plus importante hausse de dépenses est enregistrée par le Grand Est (+3,3%), suivi de la région Auvergne-Rhône-Alpes (+3,1%). La hausse la plus faible est celle du Centre-Val de Loire (+0,5%).

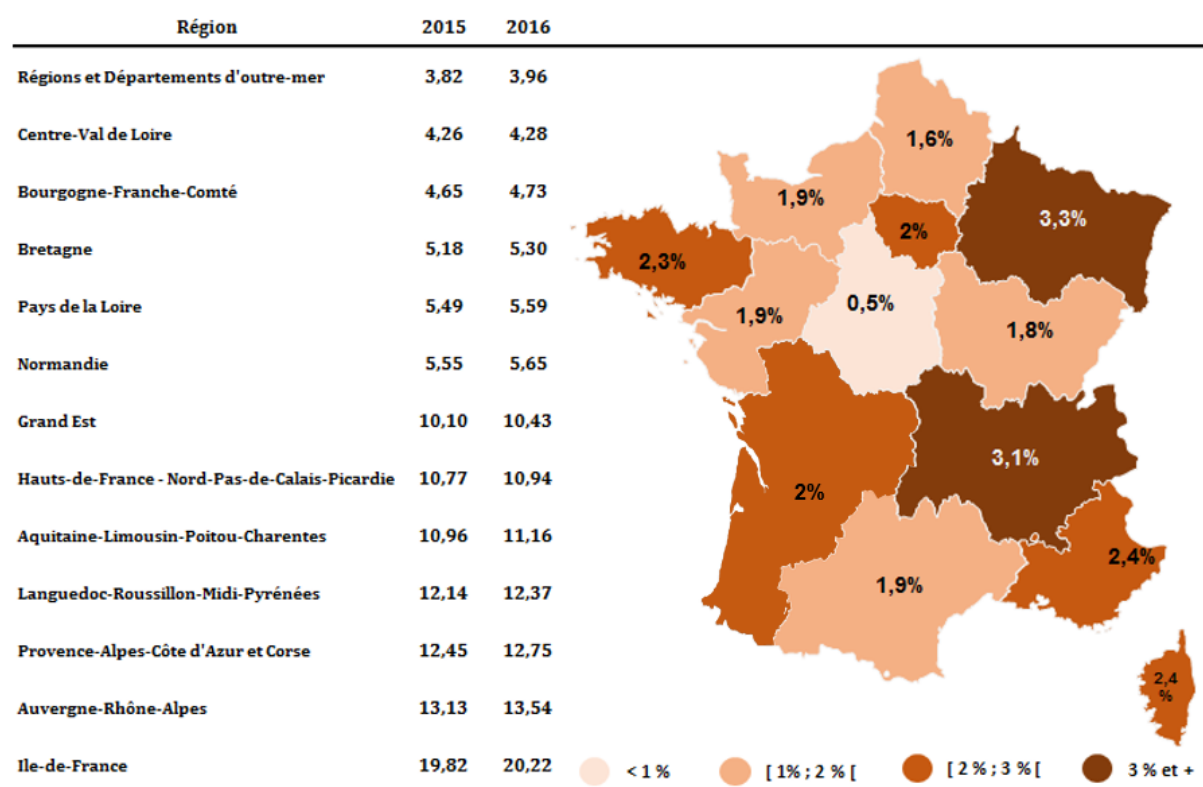


FIGURE 7.10 – Évolution et variation des dépenses de maladie (en milliards d'euros) par régions entre 2015 et 2016

En rapportant les variations des dépenses de santé aux variations du nombre d'actes de soins, on obtient, pour chaque région, une grandeur sans dimension (sans unité) : il s'agit de l'élasticité. Les dépenses de santé étant le prix des soins et le nombre d'actes de soins étant la quantité de soins demandée, la théorie économique de l'élasticité nous permet d'analyser l'effet de la demande de soins sur les frais de santé.

L'élasticité du coût des soins par rapport à la demande de soins mesure la sensibilité du coût des soins à une variation 1% de la demande de soins.

Soient U et V deux quantités. L'élasticité de U par rapport à V est donnée par :

$$e = \frac{\frac{\Delta U}{U}}{\frac{\Delta V}{V}} = \frac{\Delta \log(U)}{\Delta \log(V)}$$

Lorsque la quantité $|e| > 1$, U et V sont fortement élastiques, une hausse de 1% de V entraîne une importante variation de U . Si $|e| = 1$, U et V évoluent à la même vitesse. Lorsque $0 < e < 1$, une augmentation de V entraîne une hausse de U dans une proportion moindre. Si $e = 0$, U ne varie ni à la hausse ni à la baisse lorsque V augmente. Si $-1 < e < 0$, U diminue faiblement lorsque V augmente.

Les résultats obtenus pour chaque région sont donnés par le tableau 7.9. Pour chaque région, les dépenses de santé sont faiblement élastiques par rapport à la demande de soins (le nombre d'actes de soins). Ainsi, quelque soit la région, une hausse de 1% de la demande de soins entraîne une augmentation du coût des soins dans des proportions moindres. Le tableau 7.9 suggère que les régions pour lesquelles les dépenses de santé sont moins sensibles à la demande de soins sont le Centre-Val de Loire et le Hauts-de-France. Les régions et départements d'outre-mer sont les zones où la sensibilité des frais de maladie à la demande de soins est la plus importante. Une hausse de 1% de la demande de soins fait croître les dépenses de santé de 0,6% dans les régions et départements d'outre-mer.

Régions	$\frac{\Delta \text{ dépenses }}{\text{dépense}}$	$\frac{\Delta \text{ nombre d'acte }}{\text{nombre d'acte}}$	Élasticité
Aquitaine-Limousin-Poitou-Charentes	1,9%	8,2%	0,2
Auvergne-Rhône-Alpes	3,1%	8,3%	0,4
Bourgogne-Franche-Comté	1,8%	9,0%	0,2
Bretagne	2,3%	8,3%	0,3
Centre-Val de Loire	0,5%	7,2%	0,1
Grand Est	3,3%	10,8%	0,3
Hauts-de-France - Nord-Pas-de-Calais-Picardie	1,6%	11,2%	0,1
Île-de-France	2,0%	7,9%	0,3
Languedoc-Roussillon-Midi-Pyrénées	1,9%	8,2%	0,2
Normandie	1,9%	8,4%	0,2
Pays de la Loire	1,9%	8,1%	0,2
Provence-Alpes-Côte d'Azur et Corse	2,4%	7,6%	0,3
Régions et Départements d'outre-mer	3,8%	6,4%	0,6

TABLE 7.9 – Élasticité du coût des soins par rapport à la demande de soins par région, calculée à partir des données (2015 et 2016)

7.2.2 Les prestations de l'assurance maladie obligatoire

En 2016, les remboursements de l'assurance maladie obligatoire étaient de 98,4 milliards d'euros, ce qui correspond à une hausse de 2,7% par rapport au montant des prestations versées en 2015 (95,8 milliards d'euros) et une augmentation de 26,3% par rapport aux prestations de l'assurance maladie obligatoire de 2009 (77,9 milliards d'euros). En rapportant ces montants de remboursements aux nombres d'actes pris en charge par l'assurance maladie obligatoire, on constate une diminution continue de la prise en charge moyenne de l'assurance maladie obligatoire par actes (cf. figure 7.11). D'après les données DAMIR, de 2010 à 2012, le déremboursement moyen de l'assurance maladie obligatoire par acte de soins est de 3%, en 2013, le mouvement de baisse s'accroît, il est de -10,5%. Globalement, de 2010 à 2016, la diminution moyenne annuelle de cette prise en charge est de -5,9%.

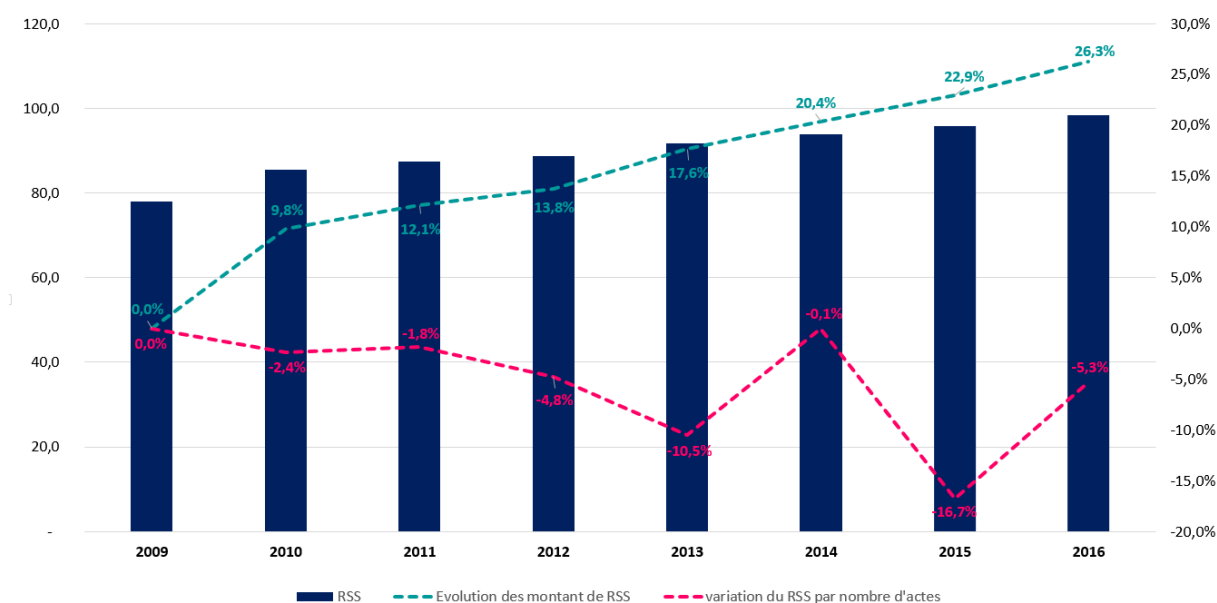


FIGURE 7.11 – Montant (en milliards d'euros) et évolution (en vert) des remboursements de l'assurance maladie obligatoire de 2009 à 2016. En rose, variation du remboursement de l'assurance maladie obligatoire par acte de 2009 à 2016

Une distinction des montants de remboursement par genre permet de mettre en évidence le fait que quelque soit l'année, plus de la moitié des remboursements ont pour bénéficiaires des femmes. En effet, en 2016, 51,5% du montant remboursé par l'assurance maladie obligatoire a pour bénéficiaires des femmes (50,7 milliards d'euros), contre 51,3% en 2015. En moyenne, de 2010 à 2016, le remboursement des montants versés par l'assurance maladie aux femmes a augmenté de 2,3% par an contre un remboursement en hausse de 2,5% par an (de 2010 à 2016) pour les hommes. De 2009 à 2016, les prestations versées aux femmes ont donc progressé de 26,3% et celles versées aux hommes ont progressé de 26,4% (cf. figure 7.12).

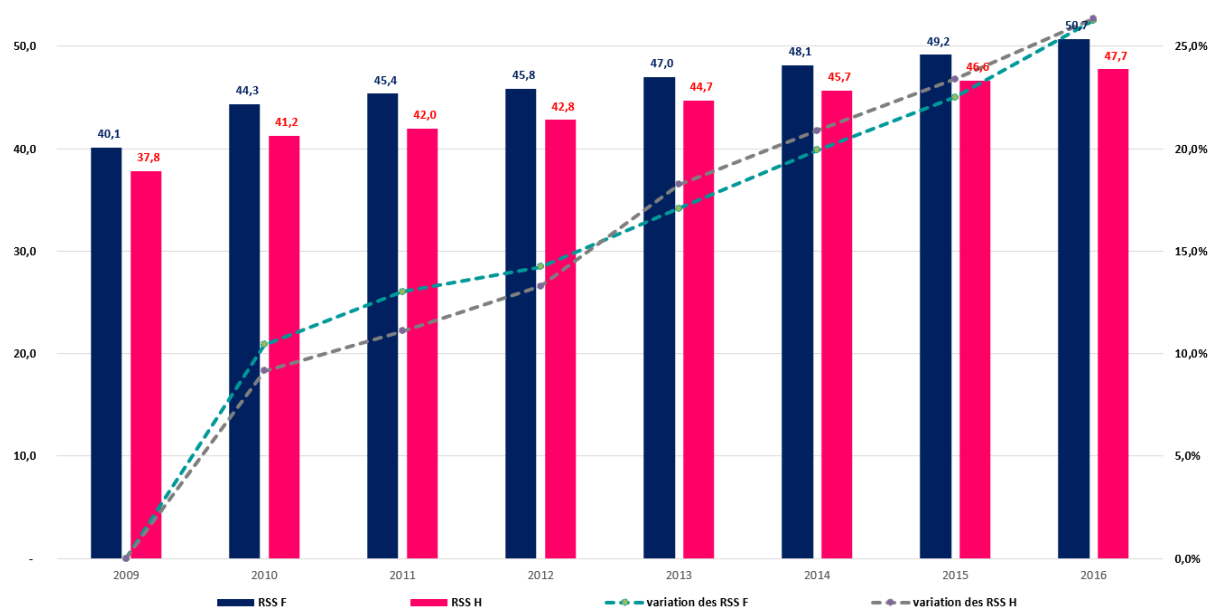


FIGURE 7.12 – Montant (en milliards d'euros) et évolution par genre des remboursements de l'assurance maladie obligatoire de 2009 à 2016.

Une répartition des dépenses par région permet de mettre en évidence la disparité des montants versés par l'assurance maladie obligatoire d'une région à une autre. En 2016, 13,6 milliards d'euros sont versées aux bénéficiaires d'Île-de-France, une hausse de 2,2% par rapport à 2015 (cf.figure 7.13), contre 3,1 milliards d'euros pour la région du Centre-Val de Loire, elle même ayant un remboursement qui est de 0,7 point au dessus de celui des régions et départements d'outre-mer. Le tableau 7.10 donne (en milliards) pour les années 2015 et 2016 les remboursements de l'assurance maladie obligatoire par région.

Régions	2015	2016
Aquitaine-Limousin-Poitou-Charentes	8,09	8,24
Auvergne-Rhône-Alpes	9,22	9,53
Bourgogne-Franche-Comté	3,36	3,40
Bretagne	3,82	3,90
Centre-Val de Loire	3,08	3,09
Grand Est	7,28	7,47
Hauts-de-France - Nord-Pas-de-Calais-Picardie	7,95	8,08
Île-de-France	13,30	13,59
Languedoc-Roussillon-Midi-Pyrénées	9,25	9,39
Normandie	4,09	4,18
Pays de la Loire	3,90	3,97
Provence-Alpes-Côte d'Azur et Corse	9,51	9,74
Régions et Départements d'outre-mer	2,95	3,07
Inconnu	9,99	10,75
Total général	95,79	98,41

TABLE 7.10 – Montants en milliards d'euros des remboursements par région de l'assurance maladie obligatoire pour les années 2015 et 2016

Le tableau 7.10 montre une hausse des prestations de l'assurance maladie obligatoire par région de 2015 à 2016. Pour chaque région, les variations de ces prestations sont données par la figure 7.13 (image de gauche). En rapportant ces variations à celles des dépenses réelles de maladie, nous obtenons par région, l'élasticité des remboursements de l'assurance-maladie obligatoire par rapport aux coûts réels de maladie (cf. image de droite de la figure 7.13). L'élasticité moyenne de l'ensemble des régions (hors régions et départements d'outre-mer) est de 1, c'est-à-dire qu'une hausse de 1% des dépenses de maladie induit une égale augmentation du niveau des remboursements de la sécurité sociale. Il existe toutefois une faible variabilité de la sensibilité des remboursements aux dépenses de maladie par région. En effet, en Normandie, en Île-de-France et dans la région Auvergne-Rhône-Alpes, les remboursements sont fortement élastiques par rapport aux dépenses de santé. Ils croissent un peu plus que les dépenses car une hausse de 1% des dépenses entraîne, pour ces trois régions, une augmentation de 1,1% du niveau des remboursements. Pour les régions Grand Est, centre-Val de Loire, Bourgogne Franche-Coté et Languedoc-Roussillon Midi-Pyrénées, la sensibilité des remboursements à une hausse de 1% des dépenses de santé est de 0,8. Une évolution de 1% des montants des dépenses de santé induit, pour ces 4 régions une augmentation de 0,8% du niveau des remboursements.

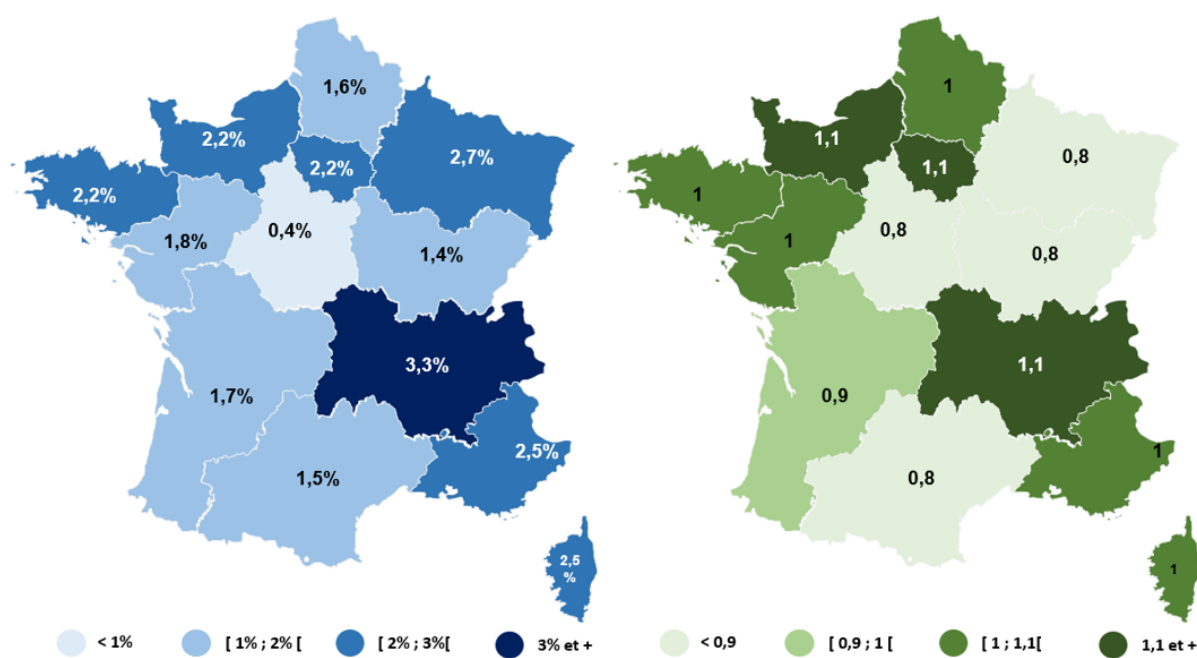


FIGURE 7.13 – A gauche : évolution des remboursements de l'assurance maladie obligatoire par région de 2015 à 2016. A droite : sensibilité (élasticité) de ces remboursements aux dépenses réelles de maladie.

Conclusion de la deuxième partie

Les 900 gigas octets de données que nous exploitons dans ce mémoire sont constituées de 41 variables catégorielles et de 14 variables quantitatives qui permettent d'expliquer les dépenses de santé de 2009 à 2016. Ces données constituent une base dont les valeurs manquantes ont été traitées par la mise en œuvre d'un modèle d'imputation basé sur les forêts aléatoires. Nous avons considéré comme aberrantes les lignes de données qui présentaient des taux de remboursement supérieurs à 100% de la base de remboursement. Ces lignes ont été supprimées. Partant des données traitées, nous donnons dans cette partie, une description d'ensemble des dépenses de réelles de maladie et des remboursements l'assurance maladie obligatoire par genre, par tranche d'âge et par région.

Troisième partie

Le reste à charge après le remboursement de la sécurité sociale

Introduction

Cette dernière partie du mémoire propose quelques analyses basées sur l'exploitation des données DAMIR. Ces analyses permettent de lever un pan de voile sur la pertinence de *l'open* DAMIR et sur son apport à la maîtrise des dépenses de santé. Elle s'intéresse principalement à un sujet d'actualité, celui du reste à charge après remboursement de la sécurité sociale. Par genre, tranche d'âge et régions de résidence des bénéficiaires de soins, nous donnons dans cette partie quelques statistiques du reste à charge sécurité sociale. Nous nous intéressons également à l'importance de la tranche d'âge, de la région de résidence et du genre de bénéficiaire vis-à-vis du reste à charge sécurité sociale et nous traitons de l'impact de ce dernier sur la demande d'actes de soins. Cette partie s'achève par une analyse chronologique des restes à charge moyen des paires de verres optiques.

Chapitre 8

Le reste charge des dépenses de santé

Ce chapitre continue la description de la base DAMIR commencée dans la partie précédente. Il contribue à affiner la description des données DAMIR en s'intéressant à la part des dépenses qui restent à la charge de l'assurance maladie complémentaire et des ménages. Il est ici question de la charge financière qu'il reste à couvrir après le remboursement de l'assurance maladie obligatoire. Ce reste à charge est constitué de tickets modérateurs, franchises, participations forfaitaires et dépassesments d'honoraires. Comme dans le chapitre précédent, les statistiques données concernent uniquement les actes de soins relevant de la branche maladie de la sécurité sociale.

8.1 Évolution et répartition du reste à charge

8.1.1 Progression annuelle du reste à charge

Les données DAMIR révèlent qu'en 2016, la branche maladie de la sécurité sociale a financé 78,6% des dépenses de maladie. Ce taux de prise en charge est de 0,1% plus important que celui de 2015, et de 0,5 point supérieur à celui de 2009. Cette hausse de financement ne parvient tout de même pas à maintenir constante, ou à ralentir, l'augmentation globale de la charge de l'assurance maladie complémentaire et des ménages. De 2009 à 2016, la part des dépenses de maladie restant à la charge de l'assurance complémentaire et des ménages n'a cessé d'augmenter (cf. figure 8.1). Avec une évolution moyenne de 1,3% depuis 2010, les montants de reste à charge sont passés de 21,8 milliards d'euros en 2009 à 26,7 milliards d'euros en 2016 (une hausse cumulée de 22,4%).

8.1.2 Le reste à charge par genre

Une distinction par genre des montants de reste à charge met en évidence l'existence d'un important écart entre la proportion de reste à charge revenant aux femmes et celle relative hommes. En 2016, 15,7 milliards d'euros de reste à charge est dû aux actes de santé dont ont bénéficié les femmes, contre 11 milliards d'euros pour les hommes. En 2009, les montants des restes à charges étaient respectivement de 13,2 et 8,6 milliards d'euros pour les femmes et les hommes.

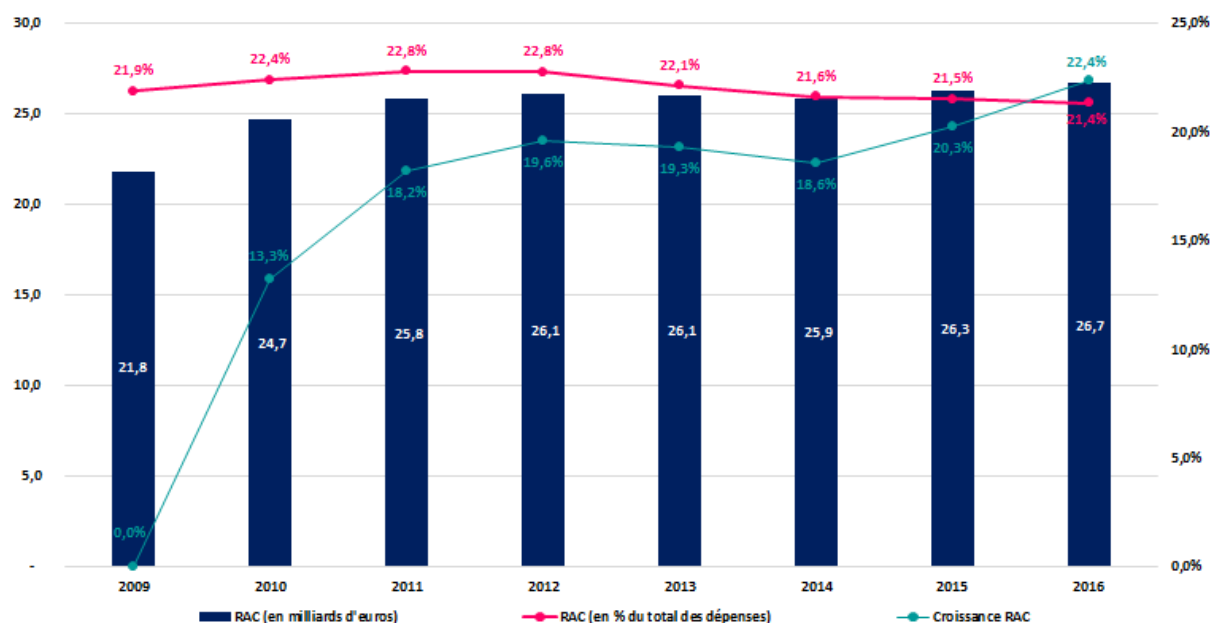


FIGURE 8.1 – Évolution (en milliards d'euros) des restes à charge après le remboursement de la sécurité sociale.

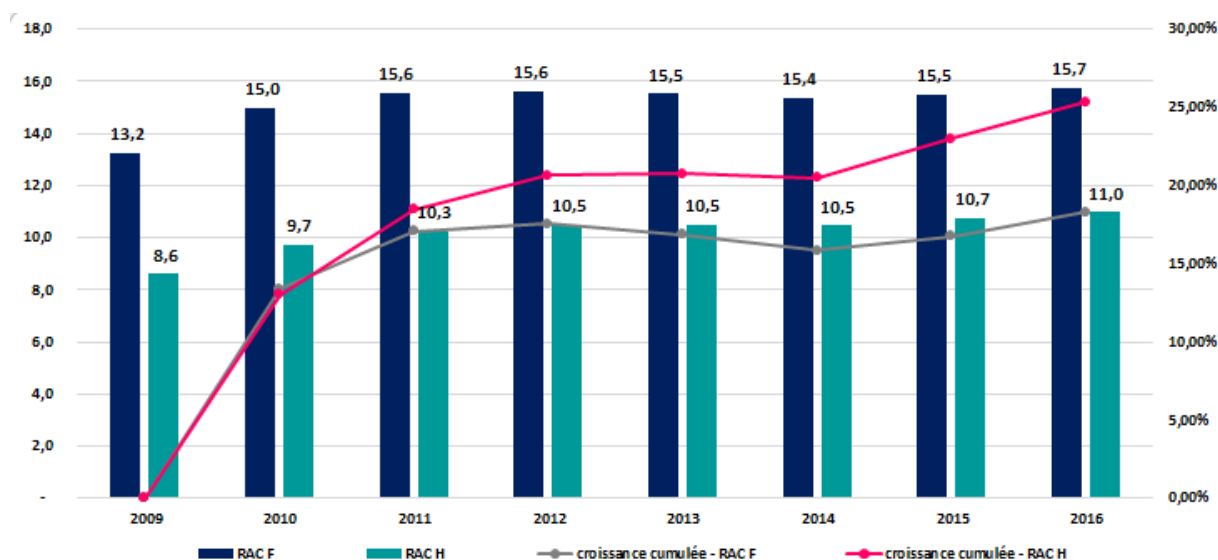


FIGURE 8.2 – Évolution (par genre et en milliards d'euros) des restes à charge après le remboursement de la sécurité sociale..

Cette forte différence des montants de reste à charge s'explique par la répartition homme/femme de la population française et principalement par le gap qui existe entre la demande de soins féminine et celle masculine. En effet la répartition moyenne de la population est de 51% de femmes pour 49% d'hommes. Cette faible différence ne suffit pas pour expliquer les écarts présentés par la figure 8.2. Une explication vraisemblable est donnée par le ratio obtenu en rapportant le volume d'actes de soins dont bénéficient les femmes à celui dont bénéficient les hommes. Pour les 8 années de données que nous considérons, le nombre d'actes de soins dont ont bénéficié les femmes est de 1,5 fois celui dont ont bénéficié les hommes, ce qui explique le fait qu'en distinguant les restes à charge par genre, la part relative aux femmes soit en moyenne de 1,5 celle relative aux hommes (cf. tableau 8.1).

Année	2009	2010	2011	2012	2013	2014	2015	2016
ratio RAC-F / RAC-H	1,5	1,5	1,5	1,5	1,5	1,5	1,4	1,4
ratio Nb d'acte-F / Nb d'acte-H	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5

TABLE 8.1 – Ratios des restes à charges et des nombres d'actes de soins par genre de 2009 à 2016

8.1.3 Le reste à charge par tranche d'âge, par nature d'acte et par région

De 2009 à 2016, la tranche d'âge ayant le volume de reste à charge le plus important est celle des 60 - 69 ans. En 2016, le montant des restes à charge de cette tranche d'âge était de 5,5 milliards d'euros, contre 5,4 milliards d'euros en 2015. Toujours en 2016, un peu moins de la moitié (48,2%) du volume global des restes à charge concerne les personnes ayant entre 30 et 69 ans. La situation est la même pour les autres années (2009 à 2015).

Tranches d'âges	2009	2010	2011	2012	2013	2014	2015	2016
0 - 19 ans	3,6	3,9	4,2	4,2	4,3	4,3	4,4	4,6
20 - 29 ans	1,1	1,3	1,4	1,5	1,5	1,4	1,4	1,4
30 - 39 ans	1,7	1,8	1,8	1,9	1,8	1,7	1,6	1,6
40 - 49 ans	2,4	2,7	2,8	2,9	2,9	2,7	2,6	2,6
50 - 59 ans	2,8	3,2	3,3	3,4	3,3	3,2	3,2	3,2
60 - 69 ans	4,0	4,7	5,0	5,1	5,1	5,2	5,4	5,5
70 - 79 ans	3,7	4,1	4,1	4,0	3,9	4,0	4,1	4,3
80 ans et +	2,7	3,0	3,2	3,3	3,2	3,3	3,5	3,6
Total général	21,8	24,7	25,8	26,1	26,1	25,9	26,3	26,7

TABLE 8.2 – Restes à charge (en milliards d'euros) par tranches d'âge et par année.

Une répartition par nature d'acte du volume global des dépenses de maladie restant à la charge de l'assurance complémentaire et des ménages en 2016, montre que sur les 862 actes que compte la base DAMIR, seulement 15 expliquent un peu moins de 68% de ce volume (cf. tableau 8.3).

Une répartition des montants annuels des restes à charge par région met en évidence une importante variabilité. En effet, en 2016, l'Île-de-France représentait 19,9% du total des restes à charge, une baisse de 0,2 point par rapport à 2015 (cf. figure 8.3). Pour ces deux années, les régions et départements d'outre-mer ont le niveau de reste à charge le plus bas : 2,7% du total des restes après remboursement de la sécurité sociale.

En ventilant les volumes de reste à charge de chaque région par tranche d'âge, nous observons pour chaque région que la tranche ayant la proportion de reste à charge la plus faible est celle des 20 - 29 ans, celle ayant la part de reste à charge la plus élevée est celle des 60 - 69 ans sauf en Île-de-France et dans les régions et département d'outre-mer où la tranche d'âge ayant la plus importante part de reste à charge est celle des 50 - 59 ans. La figure 8.4 illustre la répartition par tranche d'âge et par région des volumes de reste à charge en 2016.

Code	Libellé	poids	poids cumulé
3533	Verres	12,4%	12,4%
3313	Pharmacie PH7 (à 65)%	7,9%	20,2%
1476	Prothèse fixe céramique	7,2%	27,4%
1111	Consultation cotée C	5,5%	32,8%
3532	Monture/Lunette pour enfant de moins de 18 ans	5,5%	38,3%
1112	Consultation cotée CS	3,7%	42,0%
3211	Actes de biologie	3,3%	45,3%
1424	Traitement Orthodontie	3,2%	48,5%
3125	Actes de Kinésithérapie ostéo-articulaire	3,0%	51,5%
3312	Pharmacie PH4 (à 30%)	3,0%	54,5%
3541	Appareils électroniques de surdité	3,0%	57,5%
1352	Actes techniques médicaux (hors imagerie)	2,9%	60,4%
1321	Actes de Chirurgie	2,7%	63,1%
3111	Actes en AMI (Acte Médico-Infirmier)	2,5%	65,6%
1474	Prothèse amovible définitive métallique	2,1%	67,6%

TABLE 8.3 – Répartition par acte du reste à charge : liste et poids des 15 actes les plus importants en termes de volume de reste à charge sécurité sociale en 2016.

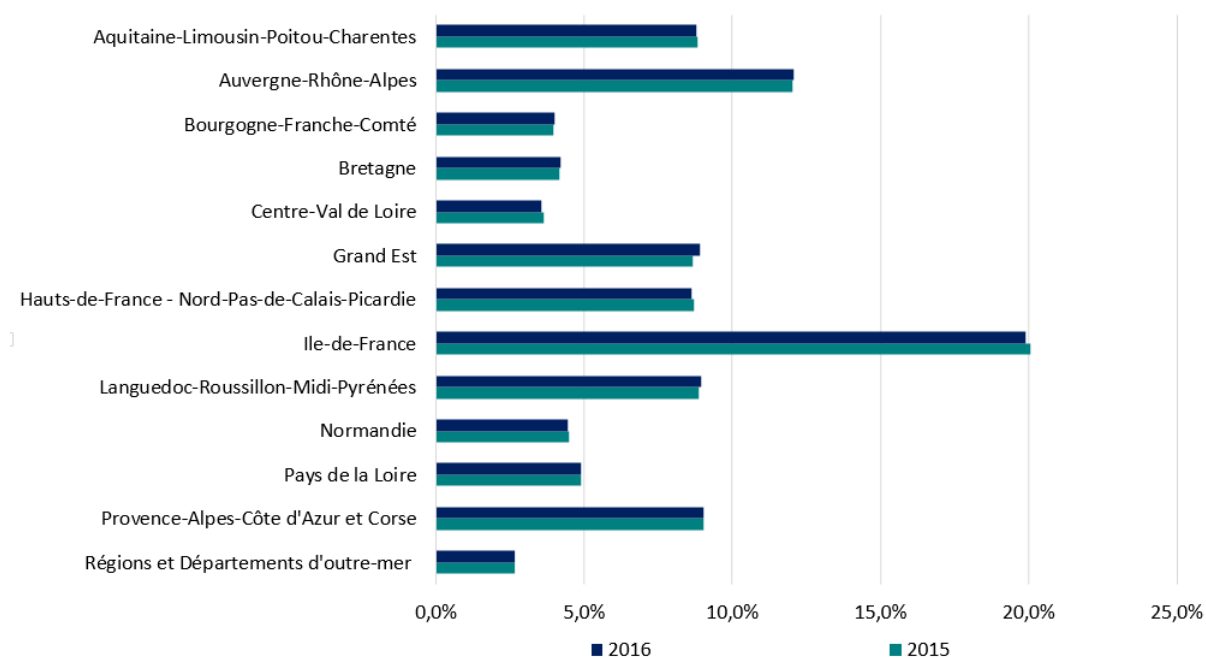


FIGURE 8.3 – Proportion de reste à charge par région pour les années 2015 et 2016.

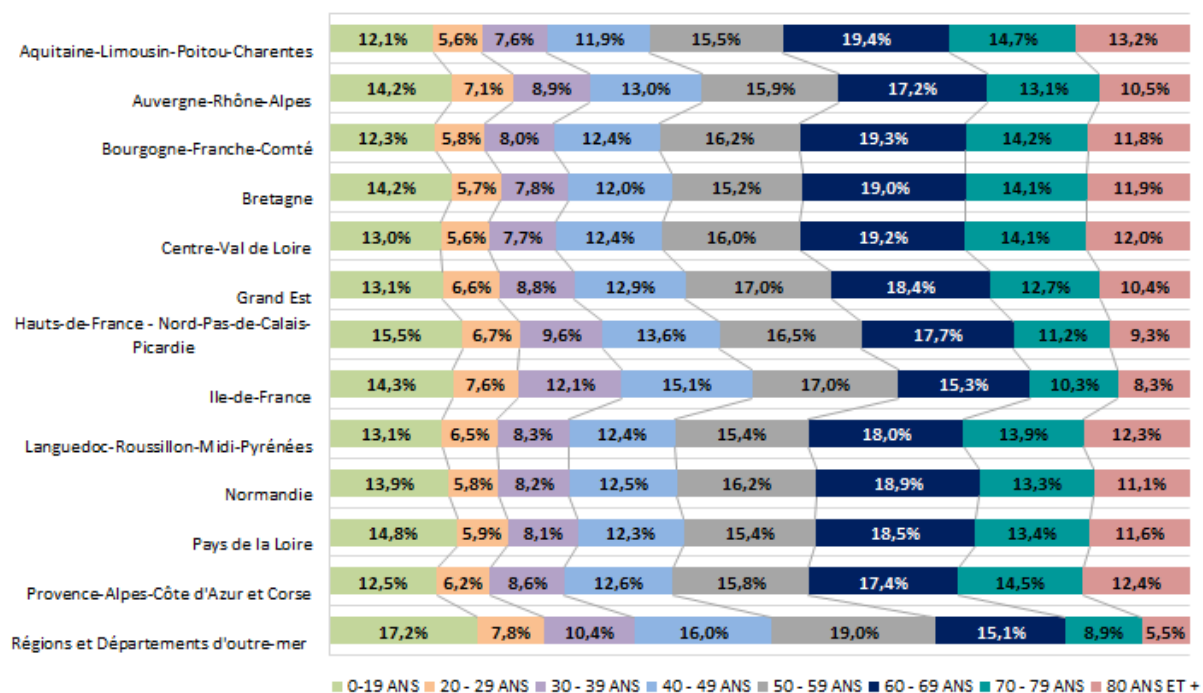


FIGURE 8.4 – Proportion de reste à charge par tranche d'âge pour chaque région en 2016.

8.2 Quelques mesures d'influences

8.2.1 Impact de la zone géographique, de la tranche d'âge et du genre sur le reste à charge

Le but de cette partie est de rendre compte des effets de la région de résidence, de l'âge (ou de la tranche d'âge) et du genre du bénéficiaire sur les niveaux de restes à charge après remboursement de l'assurance maladie obligatoire.

La description faite dans la section précédente suggère l'existence d'une forte corrélation entre le reste à charge et les variables ben_res_reg (région de résidence du bénéficiaire), age_ben_snds (tranche âge du bénéficiaire) et ben_sex_cod (genre du bénéficiaire). Est-ce réellement le cas ? Pour répondre à cette question nous utilisons le V de *Cramer*.

Le V de *Cramer* est une statistique adaptée à la quantification des liaisons qui existent sur un jeu de données mixtes, c'est-à-dire un jeu de données comportant des variables quantitatives et des variables qualitatives. Parmi les 4 variables considérées dans cette sous section, seule le reste à charge est quantitatif, les 3 autres variables sont qualitatives (ou catégorielles). Le V de *Cramer* peut être considéré comme une normalisation de la statistique du test d'indépendance du χ^2 de *Pearson*.

Considérons deux variables $Y = (y_1, \dots, y_l, \dots, y_L)$ et $X = (x_1, \dots, x_c, \dots, x_C)$, auxquelles nous associons le tableau de contingence 8.4.

Pour chaque couple d'indices (l, c) le tableau de contingence donne l'effectif $n_{l,c}$ correspondant au couple de modalité (y_l, x_c) .

L'idée de la statistique χ^2 du test d'indépendance de *Pearson* est de comparer les effectifs observés $n_{l,c}$ aux effectifs théoriques $e_{l,c} = \frac{n_{l.} \times n_{.c}}{n}$ que l'on obtiendrait si X et Y étaient indé-

$Y \times X$	x_1	\cdots	x_c	\cdots	x_C	$Total$
y_1	n_{11}	\cdots	n_{1c}	\cdots	n_{1C}	$n_{1\cdot}$
\vdots			\vdots			\vdots
y_l	n_{l1}	\cdots	n_{lc}	\cdots	n_{lC}	$n_{l\cdot}$
\vdots			\vdots			\vdots
y_L	n_{L1}	\cdots	n_{Lc}	\cdots	n_{LC}	$n_{L\cdot}$
$Total$	$n_{\cdot 1}$	\cdots	$n_{\cdot c}$	\cdots	$n_{\cdot C}$	n

TABLE 8.4 – Forme générale d'un tableau de contingence ; $n = \sum_{l,c} n_{l,c}$

pendantes. L'hypothèse nulle H_0 du test de *Pearson* est l'indépendance. Si cette hypothèse est vraie, le tableau de contingence 8.4 serait alors totalement défini par ses marges car sous H_0

$$P(Y = y_l, X = x_c) = P(Y = y_l) \times P(X = x_c)$$

La statistique χ^2 quantifie la distance (l'écart) entre les effectifs observés et les effectifs théoriques. Elle est donnée par :

$$\chi^2 = \sum_{l=1}^L \sum_{c=1}^C \frac{(n_{l,c} - e_{l,c})^2}{e_{l,c}}$$

Cette statistique permet d'arbitrer entre la dépendance et l'indépendance. Lorsque $\chi^2 = 0$, X et Y sont indépendantes. Elles sont dépendantes lorsque $\chi^2 > 0$. Cette interprétation du χ^2 soulève un problème, celui de l'intensité maximale d'une liaison quantifiée par cette statistique. La statistique χ^2 est une somme de fonctions quadratiques des écarts $|n_{l,c} - e_{l,c}|$, elle prend ses valeurs dans $[0, +\infty)$. Aussi, elle est sensible aux dimensions L et C du tableau de contingence ainsi qu'à l'effectif total n .

La statistique ***V de Cramer*** solutionne ce problème en proposant une normalisation du χ^2 qui permet d'obtenir une mesure dont on connaît la valeur maximale qui est atteinte en cas de liaison parfaite. Cette statistique est définie par :

$$V = \sqrt{\frac{\chi^2}{n(\min(L, C) - 1)}}$$

En remarquant que $\chi^2 = \sum_{l=1}^L \sum_{c=1}^C \frac{(n_{l,c} - e_{l,c})^2}{e_{l,c}} = n[(\sum_{l=1}^L \sum_{c=1}^C \frac{n_{l,c}^2}{n_{l\cdot} \times n_{\cdot c}}) - 1]$, il est facile de vérifier que $0 \leq V \leq 1$. La dépendance est parfaite lorsque $V = 1$. Elle est inexistante pour $V = 0$.

A partir des données 2015 et 2016, nous avons estimé le V de *Cramer* pour chacune des paires formées par nos 4 variables d'intérêts. Les résultats donnés par le tableau 8.5 confirment les corrélations suggérées par la description de la section précédente : le reste à charge après remboursement de la sécurité sociale est fortement lié à la région de résidence, à la tranche d'âge et au genre du bénéficiaire, ces trois dernières étant quasiment indépendantes entre elles.

	<i>ben_res_reg</i>	<i>age_ben_snds</i>	<i>ben_sex_cod</i>	RAC
<i>ben_res_reg</i>	1.00000000	0.02705072	0.03557852	0.9956143
<i>age_ben_snds</i>	0.02705072	1.00000000	0.08043496	0.9941136
<i>ben_sex_cod</i>	0.03557852	0.08043496	1.00000000	0.9978662
RAC	0.99561429	0.99411359	0.99786617	1.0000000

TABLE 8.5 – V de *Cramer*. La variable *ben_res_reg* représente la région de résidence du bénéficiaire, la variable *age_ben_snds* représente sa tranche d'âge et la variable *ben_sex_cod* représente son genre.

Modélisation du reste à charge par arbre de régression et importance des variables du modèle

Soient $Y \in R^+$ une variable qui représente le reste à charge après remboursement de l'assurance maladie obligatoire et X les variables explicatives catégorielles *ben_res_reg* (région de résidence du bénéficiaire), *age_ben_snds* (tranche d'âge du bénéficiaire) et *ben_sex_cod* (genre du bénéficiaire), appartenant à un espace Q . Le modèle de régression est alors de la forme :

$$Y = f(X) + \epsilon$$

La résolution de ce problème consiste à déterminer la fonction de régression $f : Q \rightarrow R^+$ à partir des observations (X_i, Y_i) . Ce cadre général de régression par arbre repose sur l'hypothèse que conditionnellement à X , le bruit ϵ est centré : $E(\epsilon|X) = 0$.

Remarque : Dans le cadre d'une régression linéaire la fonction f est donnée par $f(X) = X\beta$, le paramètre inconnu à déterminer étant β .

Parmi les différents modèles de régression par arbre exposés dans la littérature, nous nous intéressons à la méthode CART (*Classification And Regression Trees*) qui doit sa popularité à *Breiman & al.*[1984]. Cette méthode consiste en un partitionnement binaire récursif de l'espace Q . Les itérations s'arrêtent lorsque la partition optimale de Q est obtenue. Ainsi pour un modèle CART à n_0 classes, l'image d'un individu x par la fonction f est donnée par :

$$f(x) = \sum_i^{n_0} m_i \mathbf{1}_{\{x \in C_i\}}$$

où les C_i sont les différentes classes qui forment la partition optimale de Q et chaque m_i est une constante associée à une classe C_i . Résoudre l'équation de régression consiste alors à déterminer le nombre optimal n de classes C_i et à déterminer dans un second temps les m_i . Pour chaque m_i un estimateur \hat{m}_i sans biais est obtenue par la méthode des moindres carré. La valeur estimée d'une observation y_x de Y est alors

$$\hat{y}_x = \sum_i^{n_0} \hat{m}_i \mathbf{1}_{\{x \in C_i\}}$$

La mise en œuvre d'un modèle CART se fait en deux grandes étapes : la construction d'un arbre maximal et l'élagage de cet arbre afin d'obtenir des modèles optimaux.

L'arbre maximal

Considérons un échantillon d'apprentissage $E_n = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Les y_i représentent les modalités de Y et les x_i représentent les individus (lignes) de X dans l'espace Q

que nous supposons de dimension p (dans notre cas $p = 3$). La construction débute en divisant le nœuds racine C_1 auquel appartient tous les individus en deux nœuds fils (deux sous ensembles complémentaires) C_D et C_G . Chaque nœud fils contenant au moins deux individus est ensuite divisé de la même manière que le nœud racine, jusqu'à l'obtention des nœuds terminaux (ou feuilles). Cette procédure aboutit à la formation de l'arbre maximal qui est une partition de Q . Cette partition est optimale lorsque à chaque nœud, parmi toutes les divisions possibles on retient celle qui minimise l'impureté des nœuds fils, ce qui revient de maximiser la pureté (ou l'homogénéité) de chaque nœud de l'arbre.


- Lorsque Y est catégorielle (ce qui n'est pas le cas du reste à charge) l'impureté est mesurée par l'indice de diversité de Gini $I(a) = \sum_{b=1} p_a^b(1 - p_a^b)$, p_a^b est le taux d'observation du nœud fils b dans le nœud parent a . Ainsi, pour tout nœud C et toutes les paires admissibles (C_D, C_G) , la division optimale est celle qui maximise la quantité

$$I(C) - \frac{|C_G|}{|C|}I(C_G) - \frac{|C_D|}{|C|}I(C_D)$$


Remarque : En pratique, ce critère de division est également utilisé lorsque Y est quantitative.

- Lorsque Y est quantitative (ce qui est le cas du reste à charge), l'impureté d'un nœud C est définie par la variance : $var(C) = \frac{1}{|C|} \sum_{i, x_i \in C} (y_i - \bar{y}_C)^2$, \bar{y}_C est la moyenne des y_i des individus présents au nœud C . La division optimale est celle qui minimise la variance intra-classe

$$\frac{1}{n} \sum_{i, x_i \in C_D} (y_i - \bar{y}_{C_D})^2 + \frac{1}{n} \sum_{i, x_i \in C_G} (y_i - \bar{y}_{C_G})^2$$

Le package **rpart** du logiciel  permet de réaliser des modèles du type CART. La *data frame* constituant les 4 variables retenues pour cette sous-section a été séparée en deux jeux de données : un échantillon d'apprentissage (75% des données) et un échantillon de validation (25% des données).

Partant des données d'apprentissage, nous avons construit deux arbres maximaux. Pour l'un nous avons utilisé la mesure d'hétérogénéité de Gini et pour l'autre le critère de pureté basé sur la variance intra-classe. La figure 8.5 fait une comparaison de ces deux arbres maximaux, elle montre que les deux arbres (*fit.gini* et *fit.anova*) ont été entraînés à partir du même jeu de données d'apprentissage, qu'ils ont conduit à la formation des mêmes nœuds, mais qu'il existe une différence entre les tables de complexité *Cptable* des deux modèles.

La comparaison des tables de complexités *Cptable* (cf. figure 8.5) met en évidence la différence entre les deux modèles maximaux : ils ne conduisent pas aux mêmes taux de mauvaise classification (*xerror*). Ce taux de mauvaise classification est mesuré par validation croisée à partir de la base d'apprentissage. Partant de l'échantillon d'apprentissage E_n , K sous-échantillons disjoints et de données de taille voisine sont produits par des tirages sans remise. Par défaut $K = 10$ sous  (paramètre *xval* de la fonction *rpart.control*). Notons E_1 l'un des K sous-échantillons et \bar{E}_1 son complémentaire dans E (la réunion des $K - 1$ autres sous-échantillons). \bar{E}_1 est utilisé pour déterminer un estimateur $\hat{Y}_{\bar{E}_1}$ de $Y_{\bar{E}_1}$, l'échantillon test étant E_1 dans ce cas. Une fois cette estimation réalisée pour les autres E_k ($k \in [2 : K]$) sous-jeux de données, l'erreur de classification obtenue par cette procédure de validation croisée est

$$\frac{1}{n} \sum_{k=1}^K \sum_{i, x_i \in E_k} (y_i - \hat{y}_{\bar{E}_1}(x_i))^2$$

En considérant comme meilleur modèle celui qui minimise cette erreur de classification, l'arbre maximal retenu est celui dont le critère de pureté est fondé sur *la variance intra-classe* (cf. figure 8.6), ce qui conforte l'utilisation de cet indicateur (*la variance intra-classe*) de pureté lorsque la variable à expliquer est quantitative. L'écart absolu moyen entre les *xerrors* des deux modèles maximaux est 0,01345456. L'annexe 2 donne une représentation de l'arbre maximal retenu.

```
> # ----- Arbre max gini vs Arbre max anova ----- #
>
> all.equal(fit.gini$frame, fit.anova$frame)
[1] TRUE
> all.equal(fit.gini$splits, fit.anova$splits)
[1] TRUE
> all.equal(fit.gini$csplit, fit.anova$csplit)
[1] TRUE
> all.equal(fit.gini$where, fit.anova$where)
[1] TRUE
> all.equal(fit.gini$cptable, fit.anova$cptable)
[1] "Mean relative difference: 0.01395394"
>
> # ----- gini Cptable vs anova Cptable ----- #
>
> all.equal(fit.gini$cptable[,1], fit.anova$cptable[,1]) # Complexité (CP)
[1] TRUE
> all.equal(fit.gini$cptable[,2], fit.anova$cptable[,2]) # profondeur (nsplit)
[1] TRUE
> all.equal(fit.gini$cptable[,3], fit.anova$cptable[,3]) # erreurs d'apprentissages (relerror)
[1] TRUE
> all.equal(fit.gini$cptable[,4], fit.anova$cptable[,4]) # erreurs de classification(xerror)
[1] "Mean relative difference: 0.01345456"
> all.equal(fit.gini$cptable[,5], fit.anova$cptable[,5]) # ecart type de xerror (xstd)
[1] "Mean relative difference: 0.02797586"
```

FIGURE 8.5 – Comparaison des arbres maximaux.

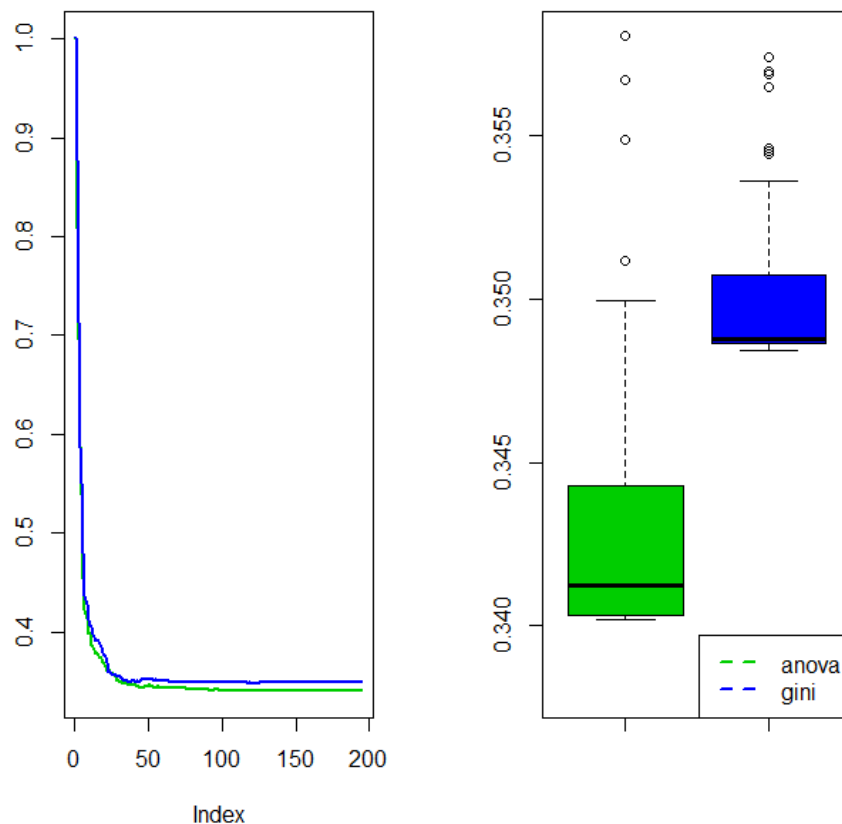


FIGURE 8.6 – A gauche, décroissance des erreurs de validations des deux arbres maximaux avec la profondeur des modèles. A droite, boîte à moustache de ces erreurs.

Élagage de l'arbre maximal

L'élagage de l'arbre maximal (*pruning*) est la deuxième grande étape de la modélisation CART. Le but de cette étape est d'extraire de l'arbre maximale le meilleur sous-arbre au sens de l'erreur de généralisation (ou erreur de prédiction). En effet, l'élagage permet de limiter le risque de sur-apprentissage (*overfitting*) car le risque lié au modèle maximal est sa spécificité aux données d'apprentissage, ce qui le rend difficilement extrapolable à de nouvelles données. L'enjeu du *pruning* est donc de déterminer le sous-arbre qui offre le meilleur compromis entre la complexité de l'arbre et sa capacité à rendre compte de la réalité qu'on souhaite appréhender.

Étant donné un arbre maximal A_{max} et A_1, \dots, A_L une suite croissante (suite emboîtée) de sous-arbres de A_{max} obtenue de manière itérative en supprimant progressivement les branches d'une étape à l'autre (de l'étape L à l'étape 1). Pour un quelconque arbre A de cette suite et $(C_k)_k$ l'ensemble de ses nœuds, l'erreur d'ajustement est

$$erreur(A) = \frac{1}{n} \sum_{C_k} \sum_{i, x_i \in C_k} (y_i - \hat{y}(x_i))^2$$

L'erreur d'ajustement étant par construction une fonction décroissante de la complexité d'un arbre, on a : $erreur(A_L) \leq erreur(A_{L-1}) \leq \dots \leq erreur(A_1)$ et $|A_1| \leq |A_2| \leq \dots \leq |A_L|$, avec $|A|$ la complexité donnée par le nombre de feuilles ou de nœuds terminaux d'un arbre A .

Une stratégie de construction de la suite A_1, \dots, A_L consiste à considérer un paramètre α ($\alpha \geq 0$) qui pénalise la complexité de l'arbre tel que pour chaque valeur de α , il existe un unique sous-arbre de A_{max} noté A_α défini par :

$$A_\alpha = \underset{\{A \text{ sous-arbre de } A_{max}\}}{\operatorname{argmin}} erreur(A) + \alpha|A|$$

avec $\alpha|A|$ qui s'interprète comme le coût de la complexité.



L'unicité de A_α est prouvée par Breiman & al.[1984]. La procédure d'élagage induit la construction progressive d'une suite croissante de paramètre de pénalité α . Pour $\alpha = 0$, $A_\alpha = A_{max}$. En prenant $A_L = A_{max}$, le terme en $L - 1$ de notre suite s'obtient en élaguant le nœud C de A_L , pour lequel

$$erreur(A_C) + \alpha_1|A_C| = erreur(C) + \alpha_1$$

où A_C est une branche de A_L qui provient de C et α_1 est donné par

$$\alpha_1 = \min_{\{C \text{ nœud interne à } A_L\}} \frac{erreur(C) - erreur(A_C)}{|A_C| - 1}$$

Une fois le nœud C élagué, on obtient alors A_{L-1} qui n'est rien d'autre que le A_L privé de A_C c'est-à-dire que $A_{L-1} = A_L - A_C$. La figure 8.7 donne une illustration précise du procédé d'élagage : l'arbre élagué $T - T_{t_2}$ (c) est obtenu en supprimant au nœud t_1 de l'arbre T (a), la branche T_{t_2} (b) qui représente les descendants gauches de t_1 .

Partant de A_{L-1} on obtient par le même procédé A_{L-2} . La construction de la suite A_1, \dots, A_L s'achève par l'obtention du nœud racine A_1 pour lequel $\alpha = +\infty$ (convention d'écriture). Parmi les L sous-arbre ainsi obtenus, l'arbre optimal est celui qui minimise l'erreur de classification ($xerror$ sous ). Les erreurs $xerrors$ des sous-arbres obtenus par le procédé d'élagage sont donnés par la figure 8.8. Le sous-arbre optimal est celui pour lequel le paramètre de complexité α (ou cp sous ) vaut 0.000997304. Une représentation graphique de l'arbre optimal obtenu par élagage est donnée par l'annexe 3. Cet arbre optimal compte 41 nœuds, 166 nœuds de moins que l'arbre maximal.

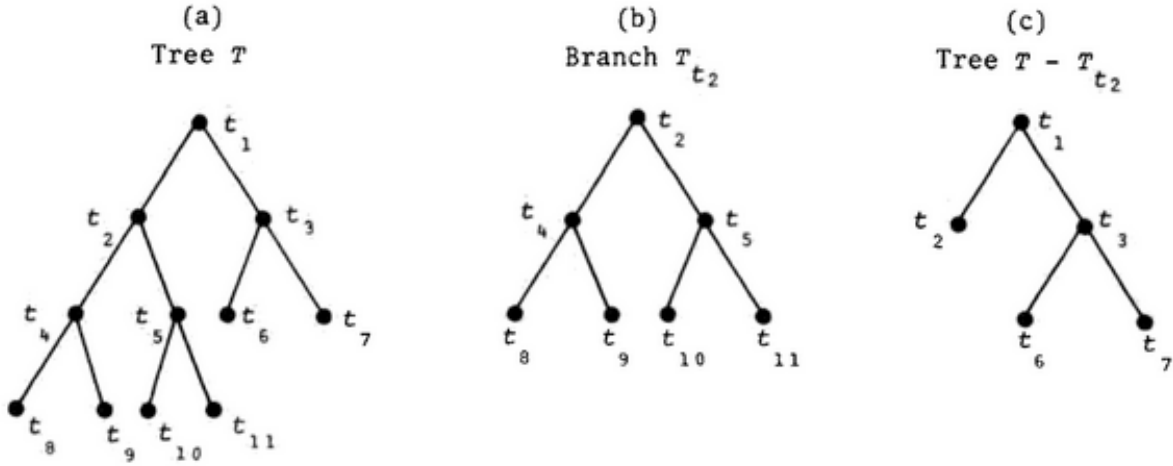


FIGURE 8.7 – Représentation du procédé d’élagage (source : *Classification and Regression Trees*[1984]).

Afin de chiffrer la qualité de l’arbre élagué, nous utilisons l’échantillon de validation. Partant de cet échantillon, du modèle maximal retenu et de l’arbre élagué, nous estimons les différentes valeurs des restes à charge. La qualité de chaque modèle est quantifiée par l’écart absolu normalisé entre les restes à charge estimés et les vraies valeurs de restes à charge. La normalisation de l’écart absolu est faite par la somme des observations réelles des restes à charge : $\text{écart} = \frac{1}{S_n} \sum_{i=1}^n (\hat{y}_i - y_i)$, y est le vecteur de vraies valeurs, \hat{y} est une estimation de y , n est la taille de y et $S_n = \sum_{i=1}^n y_i$. Pour l’arbre maximal, l’erreur d’estimation est $\text{écart}(\text{max}) = 2,615$ et pour l’arbre élagué cette erreur est de $\text{écart}(\text{élagué}) = 1,231$. Ainsi, l’élagage de l’arbre maximal a réduit l’erreur d’estimation de 53%.

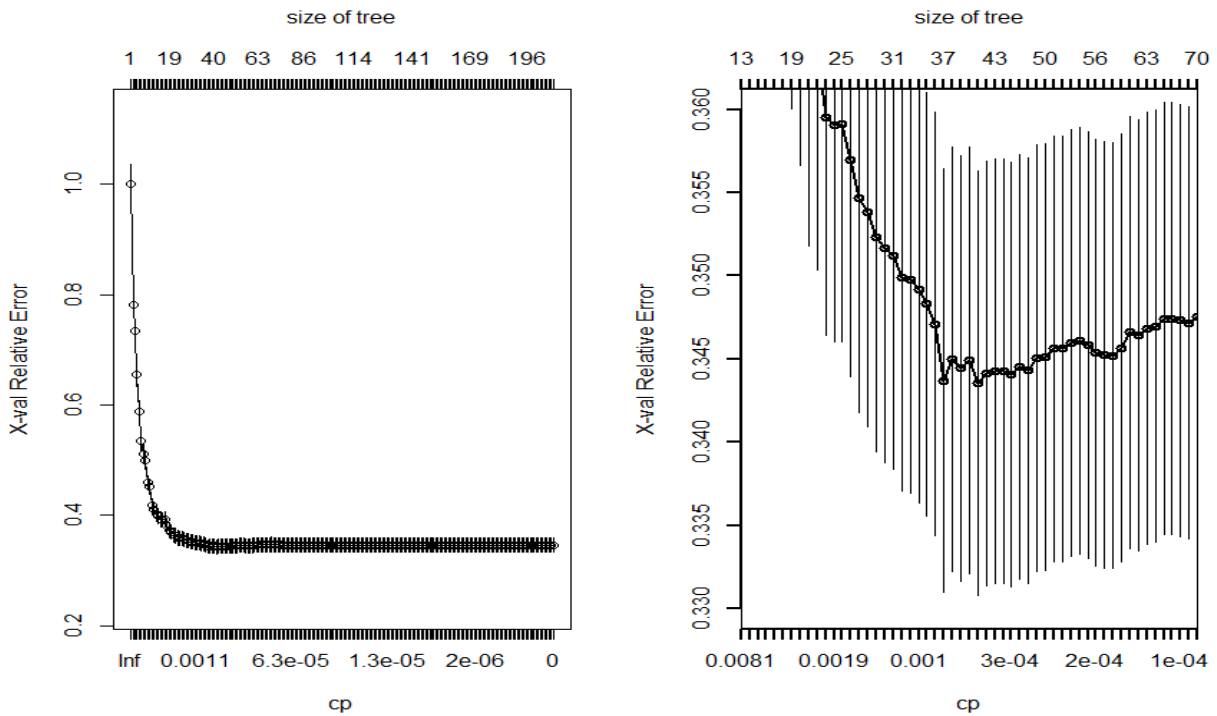



FIGURE 8.8 – Erreur de classification en fonction du paramètre de pénalisation α (ou cp sous ). A gauche, évolution de cette erreur avec le paramètre de pénalisation. A droite, zoom sur partie inférieure gauche de l’image de droite.

Importance des variables

L'arbre optimal que nous avons retenu présente une hiérarchie des variables explicatives. D'après la représentation donnée par l'annexe 3, la hiérarchie des variables explicatives, de la plus importante à la moins importante (vis-à-vis de la variable reste à charge) est : région de résidence du bénéficiaire, genre du bénéficiaire et tranche d'âge du bénéficiaire. Toutefois, la hiérarchie présentée par un arbre CART n'est pas toujours le reflète de l'importance de chacune des variables explicatives pour la variable à expliquer ou à prédire. En effet lorsque le nombre de variables explicatives est significatif, certaines variables peuvent ne pas apparaître dans la structure de l'arbre, elles sont alors inactives, les variables actives étant celles présentées par l'arbre de régression. Comme nous l'avons déjà souligné, chaque nœud C d'un arbre à plusieurs divisions admissibles. Parmi elles, une seule est optimale au nœud C , il s'agit de celle qui maximise la pureté. La division qui donne le deuxième maximum est appelée : *la première division concurrente*. Les autres divisions admissibles sont elles aussi des divisions concurrentes. Chaque division admissible est basée sur l'une des variables explicatives. A un nœud C , c'est la variable associée à la division optimale qui est active. Les variables associées aux autres divisions concurrentes sont alors inactives ou cachées au nœud C .

Remarque : La variable active à un nœud C peut également être associée à une division concurrente à ce même nœud.

Mesurer l'importance des variables ne se limite pas à la prise en compte des variables actives. A chaque nœud C de l'arbre, l'importance "locale" d'une variable X^j est donnée par :

- $imp(X^j, C) = I_{X^j}(C) - \frac{|C_G|}{|C|} I_{X^j}(C_G) - \frac{|C_D|}{|C|} I_{X^j}(C_D)$ si Y est qualitative
- et par $imp(X^j, C) = var_{X^j}(C) - var_{X^j}(C_G) - var_{X^j}(C_D)$ si Y est quantitative.

$I_{X^j}(C)$ est la mesure d'impureté donnée par *l'indice de diversité de Gini* au nœud C pour la division admissible (optimale ou concurrente) associée à la variable X^j . De même $var_{X^j}(C)$ est la variance du nœud C pour la division admissible associée à la variable X^j .

L'importance d'une variable X^j dans une modélisation par arbre est la somme des mesures d'importances locales de cette variable, c'est-à-dire des mesures d'importance de cette variable à chaque nœud C de l'arbre. L'importance d'une variable X^j pour un arbre A est donc donnée par :

$$imp_A(X^j) = \sum_{C \in A} imp(X^j, C)$$

Il est courant que cette mesure d'importance soit normalisée de manière à être comprise entre 0 et 100. Pour X^1, \dots, X^p une suite de variables explicatives et $M = \max_{j \in [1:p]} imp_A(X^j)$, la mesure d'importance normalisée d'une variable X^j , $j \in [1:p]$ est

$$\frac{1}{M} \times imp_A(X^j) \times 100$$

Par cette normalisation, la mesure associée à la variable la plus importante est 100. Pour nos données, l'importance (vis-à-vis du reste à charge sécurité sociale) des trois variables explicatives est fournie par le tableau 8.6.

<i>ben_res_reg</i>	<i>age_ben_snds</i>	<i>ben_sex_cod</i>
100.00000	69.41319	60.06306

TABLE 8.6 – Importance de la région de résidence (*ben_res_reg*), de la tranche d'âge (*age_ben_snds*) et du genre (*ben_sex_cod*) du bénéficiaire vis-à-vis du reste à charge.

Remarque : Dans la littérature il est exposé de nombreux avantages et inconvénients des modèles CART. Comme le soulignent la quasi-totalité des articles qui s'intéressent à ces modèles, leur intérêt et leur succès sont principalement dus à leur interprétabilité et au fait qu'ils permettent (de façon relativement simple) de mesurer l'importance des variables considérées dans la modélisation. Toutefois, le principal inconvénient de cette famille de modèles est leur sensibilité à une légère modification des données. Bien que cette instabilité soit atténuée par le procédé d'élagage, la volonté d'obtenir des modèles robustes a conduit au développement des forêts aléatoires, une famille de modèles mise en œuvre dans la partie précédente de ce mémoire pour traiter les valeurs manquantes.

8.2.2 Impact du reste à charge sur la fréquence d'actes

L'idée de cette partie est de rendre compte de l'effet d'une modification du niveau de reste à charge sur le nombre d'actes de soins. Dans cette partie nous mettons en évidence le lien entre ces deux variables grâce aux modèles linéaires généralisés (GLM ou *Generalized Linear Models*). Soit N le nombre d'actes de santé et X le reste à charge après remboursement de la sécurité sociale. Dans cette sous section, nous proposons une modélisation d'une espérance conditionnelle :

$$E(N/X) = \sum_i n_i P(N = n_i/X)$$

Comme le précisent Charpentier et Dutang dans *L'Actuariat avec R*, en actuariat il est courant de raisonner par classes de risques, c'est-à-dire de supposer que les variables explicatives sont qualitatives. En général on procède à une discrétisation des variables explicatives continues. L'approche que nous proposons dans cette sous-section se détache de ce cadre habituel de modélisation, nous utilisons une variable explicative continue (le reste à charge sécurité sociale). La méthode d'ajustement à laquelle nous nous intéressons est introduite par *Nelder and Wedderburn* [1972]. Dans le contexte d'un GLM, nous considérons que la relation suivante existe entre les variables N et X :

$$g(E(N/X)) = X\beta + \epsilon$$

Comme N est une variable de comptage, il est alors courant de supposer que le terme d'erreur ϵ suit une loi binomiale (sous-dispersée), une loi de Poisson (équidispersée) ou une loi binomiale négative (sur-dispersée). Comme pour un modèle linéaire classique, X est déterministe et β est le paramètre inconnu à déterminer. Il est important de remarquer que dans le cas présent X est constitué d'une unique variable (le reste à charge sécurité sociale) car X représente généralement une matrice de variables explicatives qui doivent être indépendantes deux à deux pour que le modèle soit identifiable.

De manière assez générale, un modèle linéaire généralisé a 3 composantes : une composante aléatoire N dont la distribution appartient à la famille exponentielle conditionnellement à X , une composante systématique $\eta(X) = X\beta$ et une troisième composante qui exprime la liaison entre les deux premières composantes. Cette troisième composante est la fonction g , elle est strictement monotone et dérivable et par elle, $\mu = E(N/X)$ dépend de $\eta(X)$: $g(E(N/X)) = g(\mu) = \eta(X)$.

Remarque : La distribution de probabilité d'une variable N appartient à la famille exponentielle si sa fonction de densité peut se mettre sous la forme

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

où a , b et c sont des fonctions à préciser et θ et ϕ sont des paramètres tels que :

$$E(N) = b'(\theta) \text{ et } \text{var}(N) = b''(\theta)a(\phi) \text{ avec } \theta = b'^{-1}(E(N)) = b'^{-1}(g^{-1}(X\beta))$$

La variable N (nombre d'actes) étant une variable de comptage, nous nous plaçons dans le cadre classique d'une modélisation standard : une modélisation poissonnienne avec une fonction lien logarithmique. Nous faisons l'hypothèse que conditionnellement à X , N a une distribution de Poisson de paramètre $\lambda = \lambda(X)$, sa fonction de densité est alors donnée par :

$$f_{\lambda}(N = y|X) = \exp(-\lambda) \frac{\lambda^y}{y!} = \exp(y \times \log(\lambda) - \lambda - \log(y!)), \quad y \in \mathbb{N}$$

Une comparaison de cette fonction de densité avec la forme générale précédente met en évidence l'appartenance de f_{λ} à la famille exponentielle, avec $\theta = \log(\lambda)$, $a(\phi) = \phi$, $\phi = 1$, $b(\theta) = \exp(\theta) = \lambda$, $c(y, \phi) = -\log(y!)$, $\mu = E(N/X) = \lambda$ et $\text{var}(N/X) = \lambda$.

Partant de la forme générale d'une fonction de densité appartenant à la famille exponentielle et en supposant que les n observations de N sont indépendantes, la log-vraisemblance du modèle est donnée par

$$L(\beta) = \sum_{i=1}^n \log(f_{\theta,\phi}(N = y_i)) = \sum_{i=1}^n \left\{ \frac{y_i \times \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

Par la formule de dérivée par chaîne,

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^n \frac{\partial \log(f_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta}$$

avec $\frac{\partial \log(f_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$, $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta) = \frac{\text{var}(N)}{a(\phi)}$ et $\frac{\partial \eta_i}{\partial \beta} = x_i$. Ainsi, l'équation de vraisemblance est

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_i}{\text{var}(N)} \times \frac{\partial \mu_i}{\partial \eta_i} = 0$$

La résolution de cette équation non linéaire en β a été proposée par Nelder et Wedderburn, ils ont montré que cette équation peut être résolue par une méthode itérative du type *Newton-Raphson* qui fait intervenir une matrice hessienne dont l'espérance correspond à la matrice d'information de Fischer lorsque la fonction lien considérée par un modèle linéaire généralisé est sa fonction lien canonique. Ainsi, pour le cadre de modélisation que nous avons retenu, les méthodes itératives du *scoring de Fisher* et de *Newton Raphson* coïncident. La matrice hessienne est donnée par $\partial^2 L = -X'W_{\beta}X$ et ses termes sont définies par

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n x_j x_k w_{ii} = - \sum_{i=1}^n \frac{x_j x_k}{\text{var}(N)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

où les $w_{ii} = \frac{1}{\text{var}(N)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$ sont les termes de la matrice diagonale de pondération W_{β} .

Dans le cadre d'une modélisation de Poisson avec sa fonction lien canonique (fonction $\log(\cdot)$), $\eta = \log(\lambda) = \theta$, $\frac{\partial \mu}{\partial \eta} = \frac{\partial \mu}{\partial \theta} = b''(\theta) = \frac{\text{var}(N)}{a(\phi)}$, c'est-à-dire que $\frac{\partial L}{\partial \beta} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_i}{a(\phi)}$. L'équation de vraisemblance est alors

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_i}{a(\phi)} = 0 \quad \text{ou} \quad \sum_{i=1}^n (y_i - \mu_i)x_i = 0$$

car $a(\phi)$ est une constante ($a(\phi) = 1$).

Nous avons ainsi $\frac{\partial L}{\partial \beta} = \sum_{i=1}^n (y_i - \mu_i)x_i = \sum_{i=1}^n (y_i - g^{-1}(x_i\beta))x_i = \sum_{i=1}^n (y_i - \exp(x_i\beta))x_i$. Les coefficients de la matrice diagonale W_β sont donnés par :

$$\partial^2 L(\beta) = - \sum_{i=1}^2 x_i x_i' \exp(x_i\beta) = -X' W_\beta X$$

avec W_β la matrice diagonale, définie par les termes $w_{ii} = \exp(x_i\beta)$.

Les itérations de *Newton Raphson* qui permettent d'estimer le paramètre β sont résumées de la manière suivante :

- Choisir une valeur initiale β^0 ;
- Estimer β^{k+1} à partir de β^k

$$\beta^{k+1} = \beta^k - (\partial^2 L(\beta))^{-1} \frac{\partial L}{\partial \beta}$$

Le passage de k à $k+1$ est itéré jusqu'à la convergence de cette algorithm. Sous forme matricielle, les itérations des β^k sont données par

$$\beta^{k+1} = \beta^k + (X' W_{\beta^k} X)^{-1} X' (Y - W_{\beta^k} \mathbf{1})$$

où $\mathbf{1}$ est un vecteur dont toutes les composantes valent 1.

Interprétation du $\hat{\beta}$ obtenu

L'estimateur $\hat{\beta}$ s'interprète de la même manière que β . On a :

$$\frac{\partial E(N/X)}{\partial X} = \frac{\partial g^{-1}(X\beta + \epsilon)}{\partial X} = \beta \times (g^{-1})'(X\beta + \epsilon)$$

Plus particulièrement, avec $g(x) = \ln(x)$ on obtient :


$$\frac{\partial E(N/X)}{\partial X} = \frac{\partial \exp(X\beta + \epsilon)}{\partial X} = \beta \times \exp(X\beta + \epsilon)$$

La quantité $\beta \times \exp(X\beta + \epsilon)$ représente donc l'effet d'une augmentation d'une unité X sur $E(N/X)$. Une estimation de cette quantité est $\hat{\beta} \times \exp(X\hat{\beta})$. La fonction exponentielle étant positive, l'effet est positif si $\hat{\beta}$ est positif et négatif si $\hat{\beta}$ est négatif.


Application

La sous-section précédente a mis en évidence l'importance de la région, de la tranche d'âge et du genre vis-à-vis du reste à charge. Afin de rendre compte de l'impact du reste à charge sur la fréquence d'acte de santé, il convient donc d'avoir une approche par profil d'assuré (il s'agit ici des assurés sociaux), un profil est donné par la région, la tranche d'âge et le genre. Sachant que la variable région a 13 modalités, la variable tranche d'âge 8 modalités et la variable genre 2 modalités, nous avons donc $13 \times 8 \times 2 = 208$ profils différents. A dire d'expert, nous avons

réduit ce nombre de profils en considérant uniquement 3 classes d'âges : 0 – 19 *ans* (les enfants), 20 – 59 *ans* ("population active") et ≥ 60 *ans* ("les retraités"). Nous distinguons ainsi 6 profils par région. Dans la suite de cette sous-section **nous nous intéressons à la région Île-de-France**.

Pour cette mise en œuvre, nous considérons les données 2016 (nombre d'actes et restes à charge sécurité sociale) relatives aux soins dentaires. Pour le profil "hommes de 0 à 19 ans qui résident en Île-de-France", l'algorithme converge rapidement ; en 16 itérations nous avons obtenu le résultat fourni par la figure 8.9. Le $\hat{\beta}$ obtenu est identique à celui retourné par la fonction *glm()* de  à 10^{-8} près : la procédure d'estimation de $\hat{\beta}$ que nous avons décrite est donc fiable. Pour les 6 profils de la région Île-de-France, les $\hat{\beta}$ obtenus sont donnés par le tableau 8.7.

```
> Algo.NR(dat=dat0_19ANS_H)
itération = 1 1.344487 0.0001100343
itération = 2 0.5049666 0.000109239
itération = 3 -0.08964669 0.0001072301
itération = 4 -0.174999 0.0001026975
itération = 5 0.4463172 9.466243e-05
itération = 6 1.432175 8.481911e-05
itération = 7 2.217032 7.596853e-05
itération = 8 2.675564 6.873251e-05
itération = 9 2.923784 6.254214e-05
itération = 10 3.060133 5.692915e-05
itération = 11 3.136052 5.169678e-05
itération = 12 3.178009 4.687285e-05
itération = 13 3.199916 4.276825e-05
itération = 14 3.209483 4.000678e-05
itération = 15 3.212106 3.895632e-05
itération = 16 3.212358 3.883486e-05
> glm.comp <- glm(FREQUENCE ~ RAC,data=dat0_19ANS_H, family=poisson)
> coef(glm.comp)
(Intercept)          RAC
3.212361e+00 3.883342e-05
```

FIGURE 8.9 – Estimation de $\hat{\beta}$ par les itérations de Newton Raphson et comparaison avec l'estimation donnée par la fonction *glm()* de .

Tranche d'âge	homme	femme
0 - 19 ans	3.883486e-05	6.190292e-05
20 - 59 ans	5.694503e-05	4.723192e-05
≥ 60 ans	9.494755e-05	7.077351e-05

TABLE 8.7 – Valeurs estimées des $\hat{\beta}$ par profils.

Les estimations données par le tableau 8.7 sont toutes proches de zéro et positives. D'après ce tableau, conditionnellement à une hausse d'un euro de reste à charge, le nombre moyen d'actes de soins augmente faiblement pour tous les profils. Ce résultat obtenu à partir des données DAMIR ne rend pas compte de l'intuition et de la réalité. En effet, de nombreuses analyses économiques de la santé ont montré que la demande de soins diminue avec la part des dépenses supportées par les ménages. Par exemple d'après *l'Enquête santé européenne - Enquête santé et protection sociale* (EHIS-ESPS) de 2014, 25% des personnes interrogées déclarent avoir renoncé à au moins

un soin pour des raisons financières. Dans le dossier de presse *Agir contre le renoncement aux soins, diagnostic, solutions et déploiement* du 28 mars 2017, la sécurité sociale rend compte du fait que dans 3 cas sur 4, c'est le coût trop élevé des restes à charge qui est évoqué pour expliquer le renoncement aux soins. La demande de soins est donc une fonction décroissante du reste à charge : les $\hat{\beta}$ du tableau 8.7 devraient donc être négatifs. Ce tableau met ainsi en évidence une limite du jeu de données principalement exploité dans ce mémoire : étant par construction agrégées, les données DAMIR ne permettent pas de capter le comportement de l'individu. Comme nous l'avons dit dans la première partie de ce mémoire, 95% de français bénéficient d'une complémentaire santé. Il convient donc de considérer les restes à charge après intervention de l'assurance maladie complémentaire. La seconde limite de la base DAMIR est qu'elle ne fournit pas d'information sur les prestations des organismes complémentaires.

Dans la suite de cette sous-section et uniquement pour cette sous-section, nous nous intéressons aux données issues d'un portefeuille santé de CNP Assurances, le but étant de rendre compte de l'impact réel de la part des dépenses de maladie restantes à la charge de l'individu sur sa demande de soins.

Les informations démographiques du portefeuille considéré sont présentées ci-dessous, elles concernent l'exercice 2016 (12% des personnes couvertes résident en Île-de-France).

- Nombre de personnes présentent dans l'année : adhérents 10251 soit 45%, conjoints 4525 soit 20%, enfants 7981 soit 35%.
- Effectif pondéré par la durée de présence : adhérents 8630 soit 45%, conjoints 3815 soit 20%, enfants 6826 soit 35%.
- Répartition par tranche d'âge de l'effectif pondéré : 0 - 19 ans 5 746 soit 30%, 20 - 59 ans 12 412 soit 64%, ≥ 60 ans 1 113 soit 6%.
- Âge moyen par type d'assuré : adhérent 43 ans, conjoint 46 ans, enfant 13 ans (soit 33 ans en moyenne pour l'ensemble des assurés).

Du point de vue de la répartition par nature des frais réels (FR), des remboursements sécurité sociale (RSS), des restes à charge sécurité sociale (RAC SS) et des restes à charge après intervention de l'assurance maladie complémentaire (RAC AMC), les proportions sont données par la figure 8.10.

La figure 8.10 montre clairement qu'il existe une importante disparité de prise en charge par la sécurité sociale des différentes natures d'actes de santé. Les actes de médecine courante sont les plus remboursés par la sécurité sociale (23% du total des prestations), les moins remboursés sont les actes d'optique. En effet, les dépenses d'optique sont très faiblement prises en charge par la sécurité sociale, elles représentent pourtant 14% des frais réels de santé. En terme de reste à charge sécurité sociale, l'optique représente 23% (un point de moins que la pharmacie). Lorsqu'on déduit des frais de réels de santé les remboursements de l'assurance maladie obligatoire et de l'assurance maladie complémentaire, 37% des montants restant à la charge des assurés concernent l'optique et 31% le dentaire. Pour les assurés du portefeuille considéré, les actes d'optique et de dentaire sont ceux qui coûtent les plus chères.

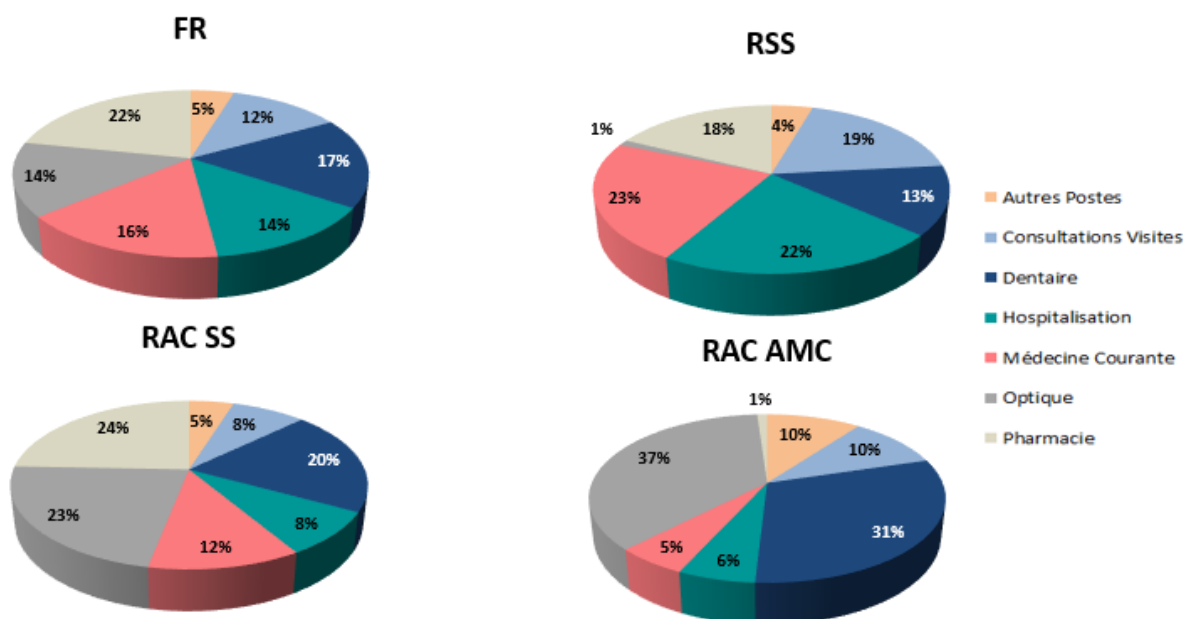


FIGURE 8.10 – Répartition des dépenses de santé par nature (données 2016 d'un portefeuille de CNP Assurances).

Concernant les dépenses de soins dentaires, les frais réels (FR) par nombre d'actes (Nb.Acte), les remboursements sécurité sociale (RSS), les remboursements de l'assurance complémentaire (AMC), les restes à charge sécurité sociale (RAC SS) et les restes à charge après intervention de la complémentaire (RAC AMC) sont donnés par le tableau 8.8.

	Nb.Acte	FR	RSS	AMC	RAC SS	RAC AMC
Femme	10 132	35,78	21,99	11,76	13,78	2,02
0 - 19 ans	1 324	30,59	21,74	8,24	8,85	0,60
20 - 59 ans	7 946	36,83	22,11	12,46	14,72	2,26
>= 60 ans	862	34,03	21,32	10,72	12,70	1,98
Homme	9 496	35,63	22,31	11,27	13,31	2,04
0 - 19 ans	1 497	30,51	21,96	8,32	8,55	0,22
20 - 59 ans	7 098	36,39	22,42	11,77	13,97	2,20
>= 60 ans	901	38,13	22,07	12,24	16,06	3,81
Ensemble	19 628	35,70	22,15	11,53	13,56	2,03

TABLE 8.8 – Dépenses moyennes (en euros) des soins dentaires par nombre d'actes, par genre et tranche d'âge en 2016 (données issues d'un portefeuille de CNP Assurances).

Rapportés au nombre de bénéficiaires (NB.Ben), les frais réels (FR), les remboursements sécurité sociale (RSS), les remboursements complémentaires (AMC), les restes à charge sécurité sociale (RAC SS) et les restes à charge après intervention de l'assurance complémentaire (RAC AMC) des soins dentaires sont donnés par le tableau 8.9.

	Nb.Ben	FR	RSS	AMC	RAC SS	RAC AMC
Femme	3 156	114,85	70,61	37,77	44,24	6,48
>= 60 ans	235	124,77	78,19	39,32	46,58	7,25
0 - 19 ans	554	73,08	51,94	19,70	21,14	1,44
20 - 59 ans	2 367	123,65	74,23	41,85	49,42	7,58
Homme	2 889	117,11	73,35	37,05	43,76	6,71
>= 60 ans	267	128,62	74,45	41,30	54,17	12,87
0 - 19 ans	607	75,23	54,16	20,52	21,07	0,55
20 - 59 ans	2 015	128,21	78,98	41,47	49,22	7,75
Ensemble	6 045	115,93	71,92	37,43	44,01	6,59

TABLE 8.9 – Dépenses moyennes (en euros) par assurés ayant bénéficié d’au moins un acte de soin dentaire en 2016 (données issues d’un portefeuille de CNP Assurances).

Partant des données de soins dentaires, les $\hat{\beta}$ obtenus sont donnés par le tableau 8.10. D’après ce tableau, **en Île-de-France, quels que soient le genre et la tranche d’âge, la part des dépenses restantes à la charge des assurés a un effet négatif sur la demande de soins dentaires** : une hausse du reste à charge des soins dentaires après remboursement de la sécurité sociale et de la complémentaire entraîne une diminution du nombre d’actes. A titre d’illustration, le figure 8.11 permet de visualiser l’impact de ces restes à charge sur la fréquence d’acte de soins dentaires.

Tranche d’âge	homme	femme
0 - 19 ans	-0.194366	-0.145763
20 - 59 ans	-0.067923	-0.045102
>= 60 ans	-0.026671	-0.047338

TABLE 8.10 – Valeurs estimées des $\hat{\beta}$ par profils (données issues d’un portefeuille de CNP Assurances).

Sachant qu’une augmentation de $\Delta X = x$ euros de la part des dépenses restantes à la charge des assurés induit une variation de $\hat{\beta} \times \exp(\hat{\beta}x)$ du nombre moyen d’actes de soins, le tableau 8.11 donne, par profil, une quantification de l’effet d’une hausse de 1 euro de reste à charge après remboursement sécurité sociale et remboursement complémentaire sur la demande d’acte de soins dentaires des assurés d’île-de-France présents dans le portefeuille considéré.

Qualité de l’ajustement

Afin d’évaluer la qualité d’ajustement des différents modèles, nous nous basons sur les différences entre les observations et les estimations issues de chaque modèle. Ces différences sont appréciées au travers de l’analyse de la déviance et des résidus de *Pearson*.

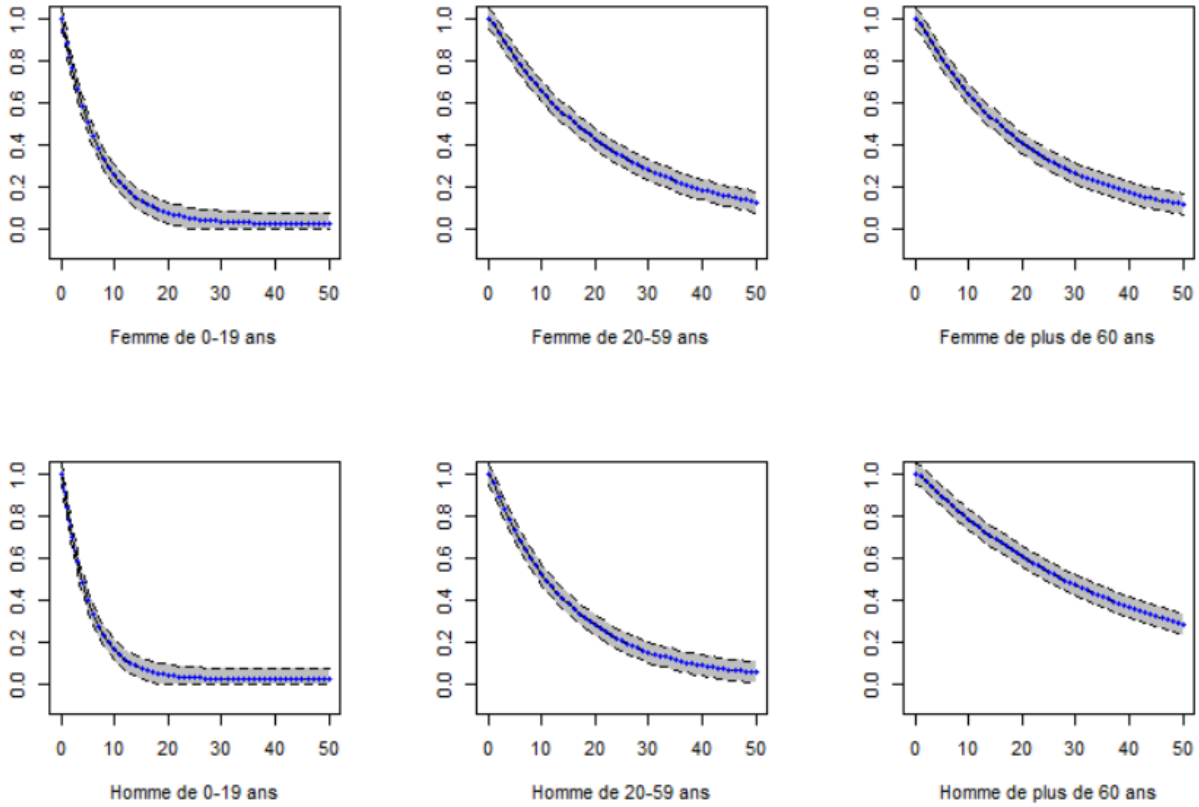


FIGURE 8.11 – Fréquence de soins dentaires en fonction de la part des dépenses de santé restantes à la charge des assurés (données issues d'un portefeuille de CNP Assurances).

Genre	Tranche d'âge	$\Delta X = x$	$\hat{\beta}$	$\hat{\beta} \times \exp(\hat{\beta}x)$	$N \{X = x_0 + x\}$
femme	0 - 19 ans	$x = 1$	-0.145763	-0.125992	0.874008
femme	20 - 59 ans	$x = 1$	-0.045102	-0.043113	0.956887
femme	≥ 60 ans	$x = 1$	-0.047338	-0.045149	0.954851
homme	0 - 19 ans	$x = 1$	-1.194366	-0.160033	0.839967
homme	20 - 59 ans	$x = 1$	-0.067923	-0.063463	0.936537
homme	≥ 60 ans	$x = 1$	-0.026671	-0.025969	0.974031

TABLE 8.11 – Partant de $N|\{X = x_0\} = 1$ acte de soin dentaire, ce tableau illustre l'impact d'une hausse de $\Delta X = x$ ($x = 1$ euro) de reste à charge sur ce nombre d'acte qui va varier de $\hat{\beta} \times \exp(\hat{\beta}x) : N|\{X = x_0 + x\} = 1 + \hat{\beta} \times \exp(\hat{\beta}x)$.

La déviance

La déviance est un outil de base dans l'analyse de la qualité d'une régression. Elle est définie par la quantité suivante :

$$D_{N=(y_i)_i} = -2[L(\hat{\beta}, N) - L_{sat}(N)]$$

où $L(\hat{\beta}, N)$ est la log-vraisemblance maximale du modèle considéré et $L_{sat}(N)$ est la log-vraisemblance saturée qui s'obtient lorsque le modèle ajuste parfaitement les données. Dans un cadre de modélisation du type Poisson, la déviance standardisée est donnée par :

$$D = 2 \sum_{i=1}^n \{y_i \times \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\}$$

La statistique D ainsi définie suit asymptotiquement une loi de χ^2 à $n-p-1$ degrés de liberté, ce qui permet de construire un test d'acceptation. Pour chacun des 6 modèles que nous avons ajusté, le tableau 8.12 donne les résultats obtenus : statistique D et P -valeur du test de χ^2 au seuil $\alpha = 5\%$. Les P -valeur observées sont toutes largement inférieures à 5%, elles représentent les probabilités que les statistiques D soient supérieures aux statistiques χ^2_{n-p-1} des tests. Ainsi, avec une erreur de 5%, nous acceptons le fait que les différents modèles ajustent bien les données.

Profil	statistique D	P -valeur
femme de 0 à 19 ans	226.10	1.1899e-23
femme de 20 à 59 ans	3469.05	1.0215e-24
femme d'au moins 60 ans	188.86	1.3083e-25
homme de 0 à 19 ans	209.27	1.4537e-16
homme de 20 à 59 ans	4227.43	1.0215e-21
homme d'au moins 60 ans	171.79	6.7924e-23

TABLE 8.12 – Validation des modèles : analyse de la déviance.

Les résidus de *Pearson*

Formellement, avec une estimation $\hat{N} = (\hat{\mu}_i)_{i \in [1, n]}$ d'une variable $N = (y_i)_{i \in [1, n]}$, les résidus de *Pearson* sont définies par $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(\hat{\mu}_i)}}$ et la somme de leurs carrés est asymptotiquement un χ^2 à $n - p - 1$ degrés de liberté, ce qui permet de construire un test d'ajustement. Les résultats obtenus sont donnés par le tableau 8.13. Ce tableau conduit lui aussi à accepter la bonne qualité d'ajustement des différents modèles avec une erreur de 5%.

Profil	$\sum_{i=1}^n r_i^2$	P -valeur
femme de 0 à 19 ans	12.454	1.1313e-4
femme de 20 à 59 ans	852.375	6.7108e-9
femme d'au moins 60 ans	10.298	1.8544e-4
homme de 0 à 19 ans	9.737	3.8567e-3
homme de 20 à 59 ans	917.056	2.4725e-8
homme d'au moins 60 ans	17.887	8.4967e-9

TABLE 8.13 – Validation des modèles : analyse des résidus de *Pearson*.

8.3 Le reste à charge zéro

8.3.1 De quoi est-il question ?

Le projet "reste à charge zéro" concerne la part des dépenses de santé qui restent à la charge des ménages après remboursement de l'assurance maladie obligatoire et de l'assurance complémentaire. Depuis le 23 janvier 2018, le Ministère de la Santé a lancé des concertations sur ce projet qui s'inscrit dans une logique d'amélioration de l'accès aux soins de l'ensemble de la population et en particulier des personnes aux revenus les plus modestes.

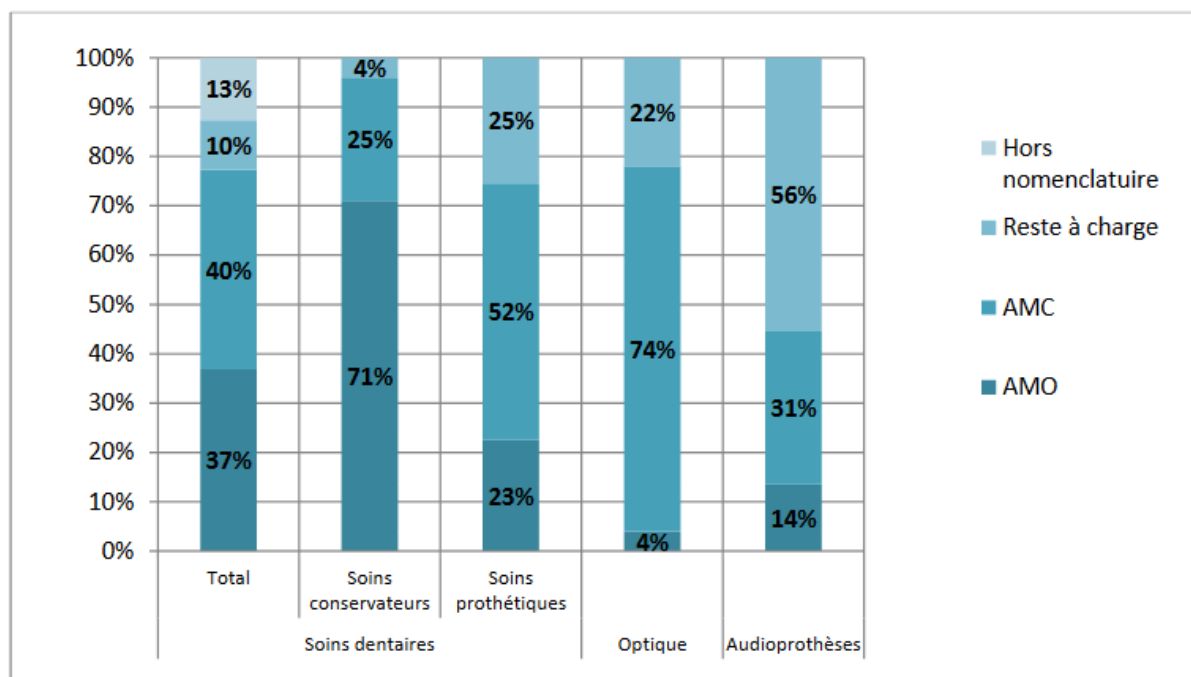


FIGURE 8.12 – Parts de l'assurance maladie obligatoire, de l'assurance maladie complémentaire et reste à charge pour les secteurs optiques, audioprothèses et dentaires. Cette illustration est issue du rapport de la commission des comptes de la santé de 2017.

Les volumes importants de reste à charge sont principalement dus à 3 secteurs : optique, audioprothèses, dentaires. La figure 8.12, issue du rapport de 2017 de la commission des comptes de la santé met en évidence le fait que la part des dépenses de santé couvertes par l'assurance maladie obligatoire pour ces trois secteurs est faible, ce qui laisse des montants importants de restes à charge aux assureurs complémentaires et aux ménages.

Le souhait du Ministère de la Santé est que ce projet se réalise à coût d'assurance constant pour les français et sans impact sur les comptes de l'assurance maladie obligatoire. Ce souhait signifie que des efforts doivent être faits par tous les intervenants des trois secteurs principalement concernés.

8.3.2 Les prothèses dentaires, les verres optiques et les audioprothèses

Comme le montre le tableau 8.14, la part des dépenses remboursées par l'assurance maladie obligatoire pour les verres optiques, les prothèses dentaires et les audioprothèses est faible. Avec un coût moyen de 674 euros, une prothèse dentaire a un reste à charge de 564 euros (83,67% du coût réel). Un pareil niveau de reste à charge peut expliquer le renoncement aux soins des personnes à bas revenu. Le tableau 8.14 montre également qu'en moyenne, le coût réel d'un verre

	Nombre d'actes	Frais réels moyens	AMO moyen	RAC SS moyen
Prothèse dentaire	6 427 816	673,96	110,06	563,90
verre optique	23 205 574	175,92	5,86	170,07
Audioprothèse	1 864 147	584,72	79,18	505,54

TABLE 8.14 – Frais réels, remboursements sécurité sociale et restes à charge moyens en euros des prothèses dentaires, verres optiques et audioprothèses (données DAMIR 2016).

optique en 2016 était de 175,92 euros (soit 351,84 euros pour la paire de verres) avec un reste à charge sécurité sociale de 170,07 euros (97,67% du coût réel). Concernant les audioprothèses, le coût réel moyen donné par la base DAMIR correspond à un coût d'entrée de gamme : 585 euros pour une prothèse, soit 1170 euros pour une paire d'audioprothèses. En terme de reste à charge sécurité sociale, l'assurance maladie complémentaire et les ménages supportent 86,45% des frais réels d'audioprothèses.

Qu'il s'agisse de l'optique, du dentaire ou des audioprothèses, des volumes importants de restes à charge sécurité sociale poussent à se demander ce qu'il en sera de l'avenir. Concernant les verres optiques, le chapitre suivant répond à cette question par une analyse chronologique.

Chapitre 9

Analyse temporelle du reste à charge sécurité sociale : cas des verres optiques

9.1 Chronique des données

Dans ce chapitre nous proposons une analyse chronologique du reste charge sécurité sociale des paires de verres optiques. Il est concrètement question de donner une idée des niveaux futurs de reste à charge après remboursement de la sécurité sociale. Le tableau 9.1 donne, par mois et par année le niveau moyen (par acte) de reste à charge sécurité sociale des paires de verres optiques.

Année	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2009	316	314	313	313	311	314	317	313	306	314	311	316
2010	326	324	323	324	322	324	327	324	316	325	322	327
2011	327	325	324	324	322	325	328	325	317	325	322	327
2012	334	331	331	331	329	331	335	331	323	332	329	334
2013	331	329	328	328	326	329	332	329	321	330	326	331
2014	329	326	326	326	326	328	335	334	326	335	334	342
2015	337	335	336	336	331	336	341	337	328	336	336	345
2016	338	340	335	339	337	342	343	338	328	344	342	349

TABLE 9.1 – Chronique des restes à charges moyen des paires de verres optiques.

Cette chronique est graphiquement représentée par la figure 9.1. Un regard sur cette illustration nous permet d’observer une tendance qui est globalement haussière et un effet de saisonnalité suggéré par des pics et des creux régulièrement espacés.

Entre deux dates $t - 1$ et t , le taux d’accroissement est donné par $\Delta Y_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$ avec Y_t est la valeur en t de la chronique des restes à charges.

La série $(\Delta Y_t)_t$ (cf. figure 9.3) ne présente globalement pas de tendance sur la période d’étude. Elle présente des accroissements brusques et régulièrement espacés, souvent suivis par des variations plus faibles les mois d’après.

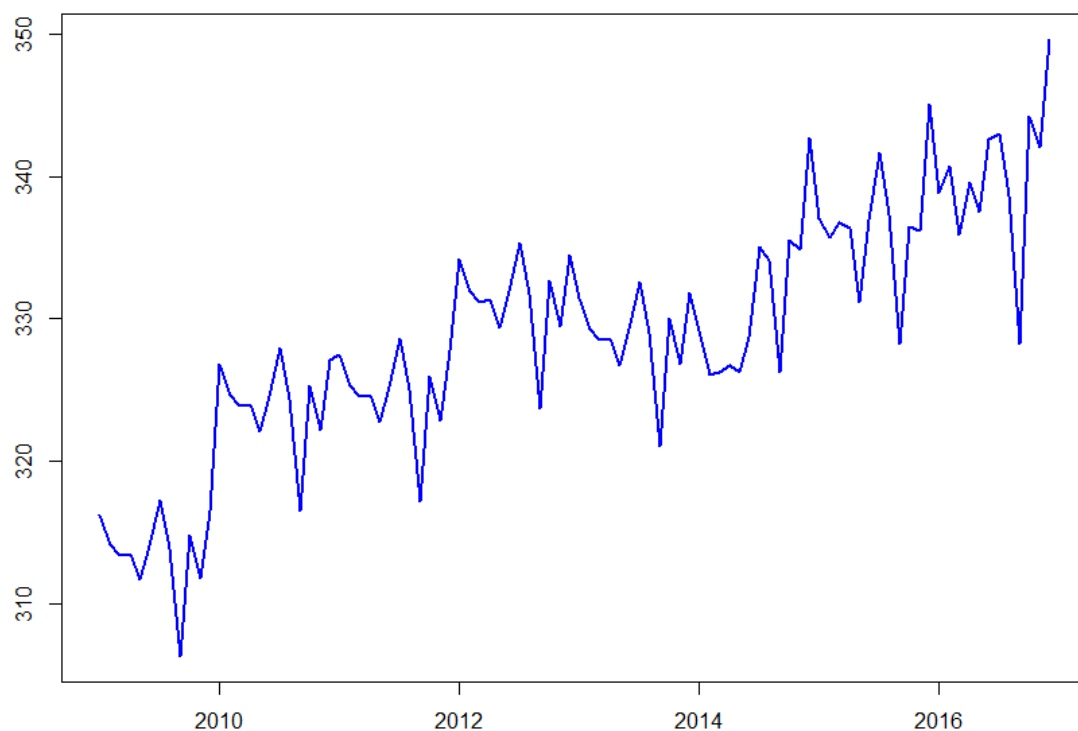


FIGURE 9.1 – Evolution mensuelle des restes à charge moyen des paires de verres optiques par mois et année, de 2009 à 2016

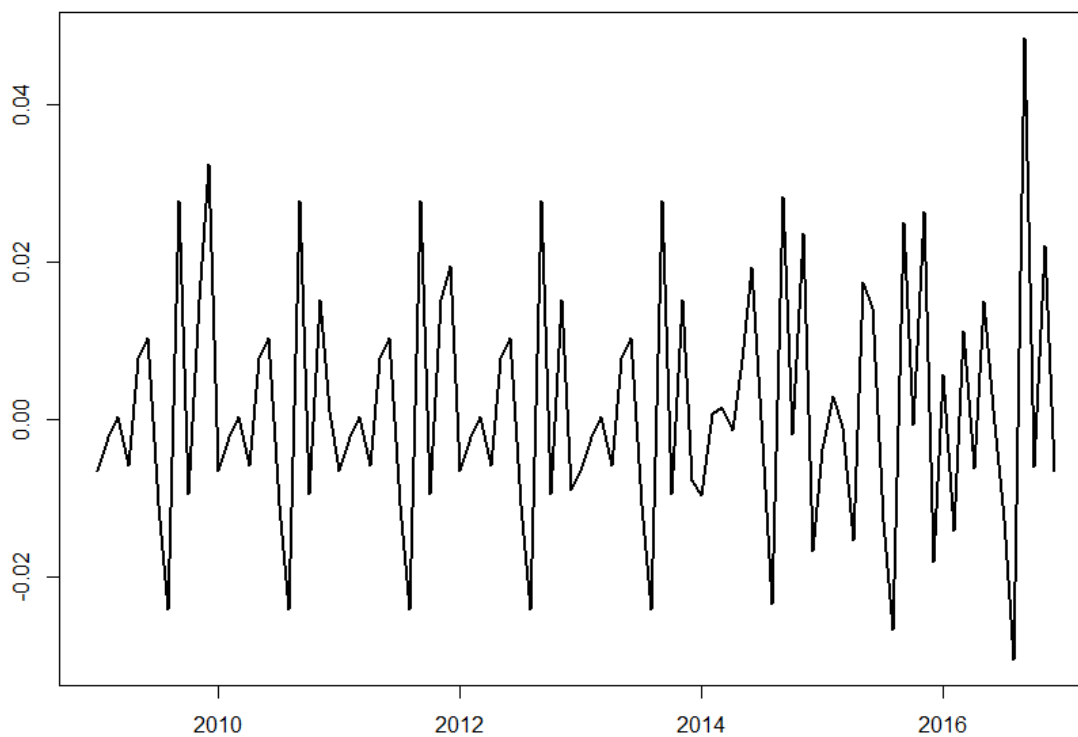


FIGURE 9.2 – Représentation des accroissements ΔY_t des restes à charge.

Afin de rendre compte l'effet du mois sur la valeur de la chronique, nous comparons les données mois à mois des 8 années considérées.

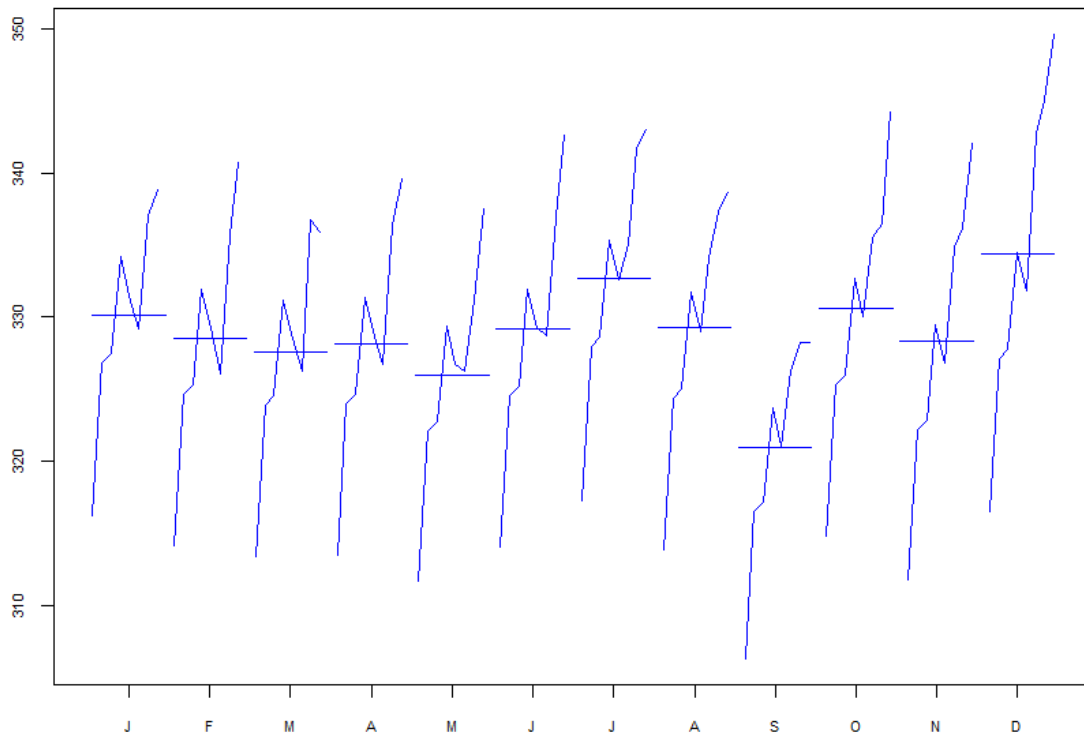


FIGURE 9.3 – *Month plot* des restes à charges moyen des paires de verres optiques.

La figure 9.3 permet de remarquer que de 2009 à 2016, les données mois à mois des restes à charge moyens des paires de verres optiques évoluent de manières quasi-identique, elles sont globalement croissantes. Elles présentent un décrochage entre 2012 et 2013. En moyenne, les restes à charge des verres optiques sont décroissants de janvier à mai et croissants de mai à juillet. Cette croissante est suivie d’une importante dépression. Chaque année, les plus bas niveaux de reste à charge sont observés en septembre. De septembre à décembre, les niveaux de reste à charge remontent, le plus haut annuel étant enregistré en décembre.

L’un des préalables à la prévision des niveaux futurs d’une série temporelle est sa décomposition en trois grandes composantes : tendance, saisonnalité et résidus. Le but de la section suivante est de déterminer la décomposition la plus adaptée à notre série chronologique.

9.2 Type de décomposition : additive ou multiplicative

9.2.1 Choix du type de décomposition par la méthode de la bande

Cette méthode s’appuie sur une représentation graphique des maxima et des minima de chaque saison. On trace une droite passant par les minima et une autre passant par les maxima, si les deux droites sont parallèles alors le modèle est additif. Dans le cas contraire, c’est un modèle multiplicatif. La figure 9.4 représente les résultats obtenus pour notre série de données. Les droites des minima (en rouge) et maxima (maxima) ne sont pas parfaitement parallèles. Elles nous suggèrent toutefois une **décomposition additive**.

Les droites des minima et maxima n’étant pas rigoureusement parallèles, il existe une incertitude sur le modèle de décomposition suggéré par la méthode de la bande. La sous-section suivante propose une deuxième méthode de choix du type de décomposition : la méthode des courbes superposées ou des profils.

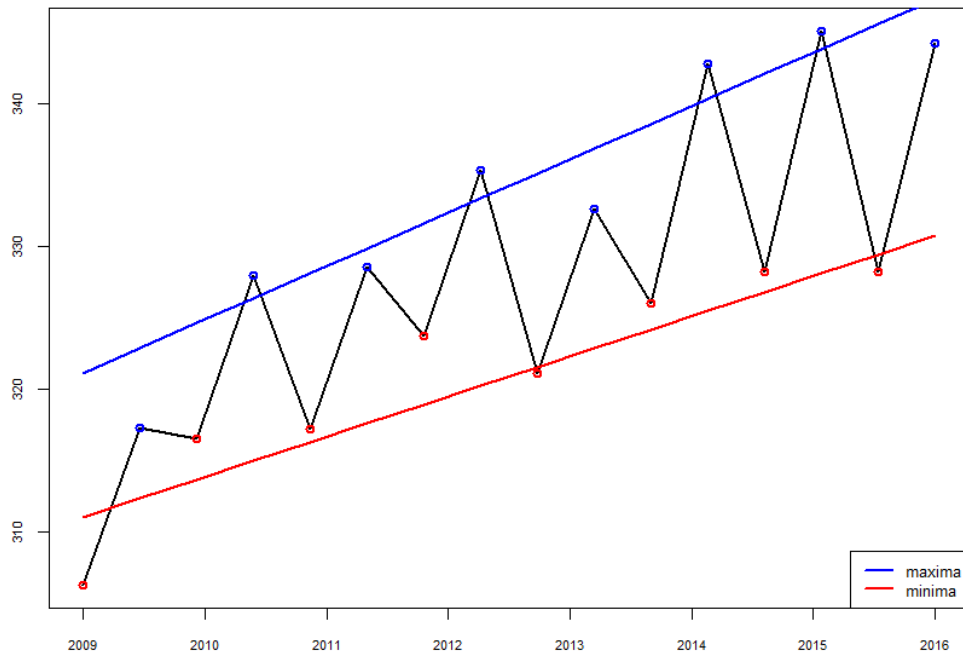


FIGURE 9.4 – *Méthode de la bande* pour le choix de la décomposition adaptée à la série des restes à charge moyens des paires de verres optiques.

9.2.2 Choix du type de décomposition par la méthode des profils

Cette méthode consiste à représenter sur un même graphique les courbes d'évolutions annuelles de la série chronologique. Si les courbes sont à peu près parallèles, le modèle est alors additif sinon multiplicatif. La figure 9.5 fait apparaître de grandes similitudes de forme entre les courbes d'évolution annuelle de la chronique des restes à charge des paires de verres optiques. Les courbes annuelles ont pratiquement la même forme, elles ne se coupent qu'en quelques points bien identifiables. Cette méthode suggère donc également une **décomposition additive**.

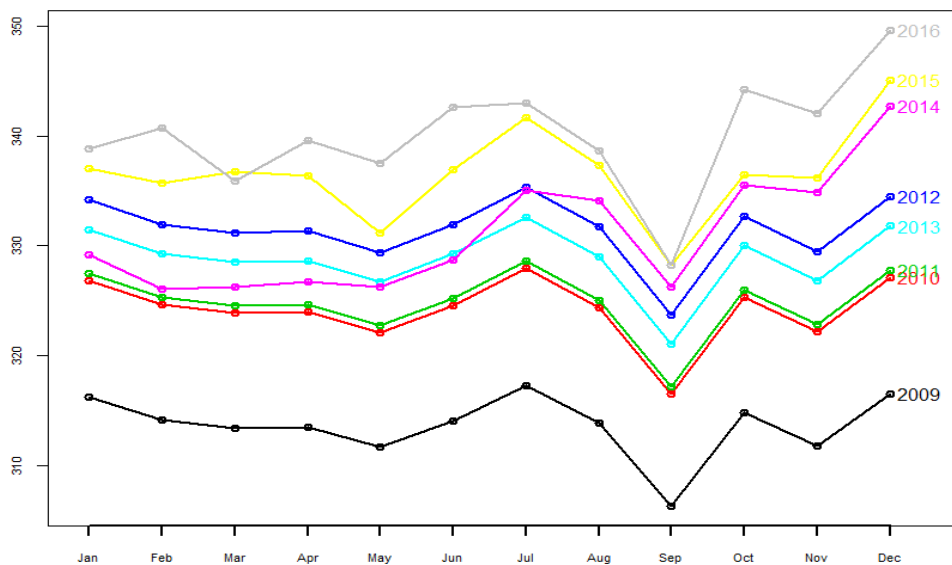


FIGURE 9.5 – *Méthode des profils* pour le choix de la décomposition adaptée à la chronique des restes à charge moyens des paires de verres optiques.

Bien que cette deuxième méthode confirme le caractère additif de notre série de données, l'existence de quelques points de coupures entre les courbes nous impose de valider le type de décomposition suggéré par une méthode analytique.

9.2.3 Méthode analytique du tableau de Buys-Ballot

La méthode du tableau de Buys-Ballot consiste à calculer pour chaque année, la moyenne et l'écart-type de la série de données (cf. tableau 9.2), puis à tracer la droite des moindres carrés correspondante au nuage de points obtenu : $\sigma = a\bar{x} + b$. Lorsque a est presque nul, la décomposition est additive, dans le cas contraire elle est multiplicative.

Année	2009	2010	2011	2012	2013	2014	2015	2016
Moyenne	313.62	324.13	324.78	331.46	328.78	330.98	336.58	339.20
Écart-type	2.88	2.98	2.98	3.05	3.02	5.37	4.25	4.22

TABLE 9.2 – Moyennes et écarts-types annuels de la série chronologique.

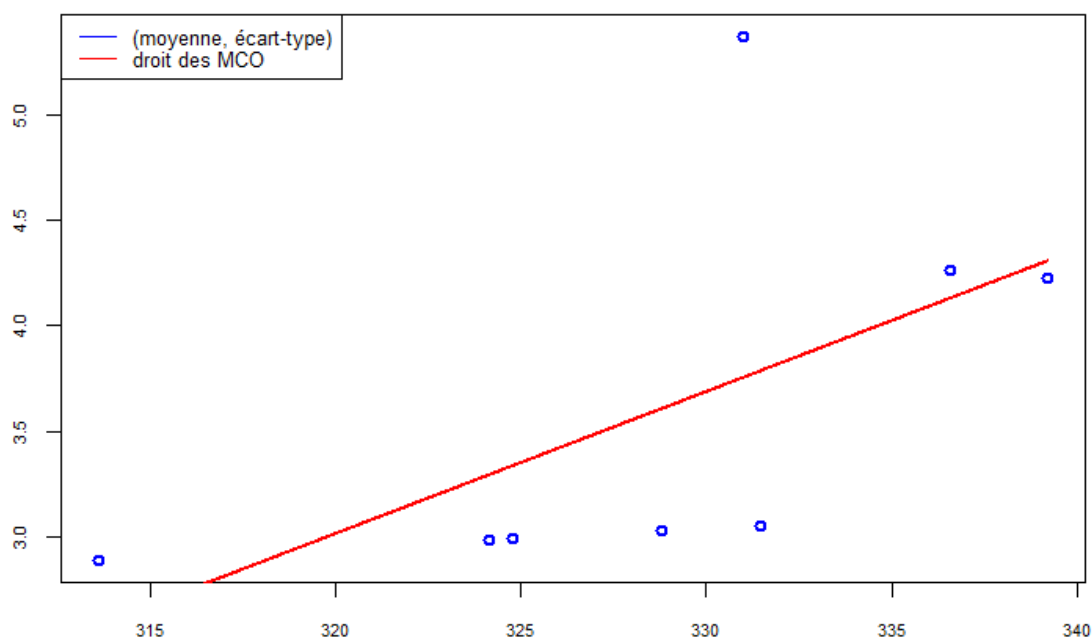


FIGURE 9.6 – Méthode du tableau de Buys-Ballot pour le choix de la décomposition adaptée à la chronique des restes à charge des paires de verres optiques.

Les résultats obtenus par cette méthode sont donnés par le tableau 9.3 et la représentation de la droite d'ajustement est donnée par la figure 9.6. La valeur estimée du coefficient a est $\hat{a} = 0,06723$ avec une erreur de $0,03766$. Le critère $Estimate \pm 2 \times Std.Error$ nous conduit à accepter la nullité du coefficient a car $[\hat{a} \pm 2 \times Std.Error(\hat{a})] = [-0.00809, 0.14255]$. Nous retenons donc la décomposition suggérée par les deux méthodes précédentes : **une décomposition additive**.

	Estimate	Std.Error	t value	Pr(> t)
\hat{b}	-18.49974	12.38153	-1.494	0.186
\hat{a}	0.06723	0.03766	1.785	0.124

TABLE 9.3 – Estimation des coefficients a et b de la droite des moindres carrés.

9.3 Décomposition de la série

L'hypothèse à la base des méthodes de décomposition est qu'il est possible d'écrire une série chronologique additive $Y = (Y_t)_t$ comme somme de trois composantes qui sont : la tendance Z_t , la saisonnalité S_t et les résidus ϵ .

$$Y = Z_t + S_t + \epsilon_t, \quad t = 1 \cdots T,$$

avec $t = 1$ qui correspond à janvier 2009 et $t = T$ à décembre 2016. La composante qui représente la tendance exprime le mouvement à long terme de la série Y . La composante S exprime un phénomène périodique, il se reproduit de manière identique sur des intervalles de temps successifs. Dans notre cas S est périodique de période 12 c'est-à-dire que $\forall t; S_{t+12} = S_t$. Concernant les erreurs ϵ_t , nous supposons qu'elles sont centrées et homoscédastiques. En général le processus $(\epsilon_t)_t$ est supposé être un bruit blanc, c'est-à-dire $E(\epsilon_t) = 0$ et $E(\epsilon_{t_1}, \epsilon_{t_2}) = \sigma^2 \delta_{t_1 t_2}$. De plus ce bruit blanc est gaussien si $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

9.3.1 Méthode de Buys-Ballot généralisée (BB)

Afin de modéliser l'évolution globale de la série des restes à charge des paires de verres optiques, nous avons testé plusieurs tendances : linéaire, quadratique et polynomiale. La méthode consiste à effectuer une régression linéaire sur l'ensemble des données. Pour choisir la tendance la plus adaptée et la saisonnalité la plus adéquate, nous étudions la normalité des résidus par le test de *Shapiro-Wilk*.

Test de Shapiro-Wilk

Soit un échantillon $X_n = (x_1, \dots, x_n)$ de variables aléatoires. L'hypothèse nulle H_0 du test de *Shapiro-Wilk* est : X^n est issu d'une distribution gaussienne, contre H_1 : X^n n'est pas issu d'une distribution gaussienne. Ce test est basé sur une statistique W_n définie par :

$$W_n = \frac{\{\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_i \times (x_{(n-i+1)} - x_{(i)})\}^2}{\sum_i (x_i - \bar{x})^2}$$

où les $x_{(i)}$ correspondent aux valeurs de la série X^n triée, $\lfloor \frac{n}{2} \rfloor$ est la partie entière de $\frac{n}{2}$ et les a_i sont des coefficients calculés à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille n de loi normale.

La statistique W_n peut donc être interprétée comme le carré du coefficient de corrélation entre la série des quantiles théoriques générés à partir de la loi normale et l'échantillon empirique X^n . Ainsi, plus W_n est élevé, plus la compatibilité avec la loi normale est forte.

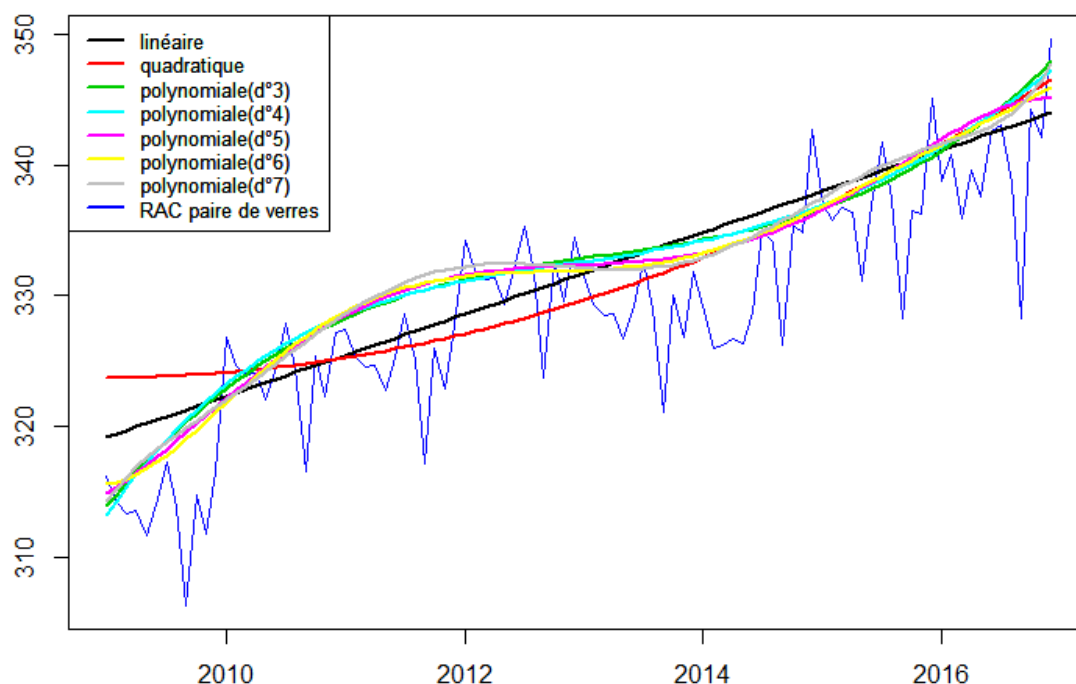


FIGURE 9.7 – La chronique des restes à charge moyens des paires de verres optiques et les différentes tendances testées.

Les tendances estimées par les différents modèles de régression que nous avons ajustés à notre chronique sont représentées par la figure 9.7. Le tableau 9.4 donne un résumé des résultats du test de *Shapiro-Wilk* que nous avons effectué sur les résidus des différents ajustements. À l'exception du modèle associé au polynôme de degré 2, toutes les *P-valeurs* sont supérieures à 0,05, elles conduisent donc à accepter l'hypothèse nulle du test de *Shapiro-Wilk* qui est la normalité des résidus des différents modèles de régression. Un regard sur les statistiques W_n montre que les résidus associés au polynôme de degré 6 sont les plus normaux. Toutefois, comme on peut le visualiser sur figure 9.7, à partir du degré 3 les différents polynômes offrent quasiment le même ajustement. C'est le modèle de régression associé au polynôme de degré 3 que nous retenons pour modéliser la tendance et la saisonnalité de la chronique des restes à charges moyens des paires de verres optiques.

tendance	W_n	<i>P-valeur</i> du modèle
linéaire	0.9740506	0.0535629745
quadratique	0.9336148	0.0001119708
polynôme(d°3)	0.9741239	0.0542490086
polynôme(d°4)	0.9736468	0.0519410386
polynôme(d°5)	0.9852735	0.3604133845
polynôme(d°6)	0.9862315	0.4174321197
polynôme(d°7)	0.9845732	0.3226964217

TABLE 9.4 – Statistiques du test de Shapiro-Wilk et *P-valeur*.

Afin d'assurer l'unicité de la décomposition, nous imposons aux coefficients saisonniers (qui sont associés à ce polynôme de degré 3) de se compenser sur une année, ce qui nous conduit à imposer la contrainte :

$$\sum_{t=1}^{12} c_t = 0$$

où les c_i sont les coefficients saisonniers issus de la régression. La composante saisonnière S_t est telle que :

$$S_t = \sum_{i=1}^{12} c_i S_t^i \text{ avec } S_t^i = \mathbf{1}_{t=i \bmod 12}$$

Pour respecter la contrainte d'identifiabilité, nous remplaçons c_1 par $-\sum_{i=2}^{12} c_i$, on obtient ainsi :


$$S_t = \sum_{i=2}^{12} c_i S_t^{*i} \text{ avec } S_t^{*i} = S_t^i - S_t^1$$

Significativité du modèle retenu

Afin d'apprécier la significativité du modèle que nous avons retenu pour modéliser la tendance et la saisonnalité de notre série, nous utilisons la statistique du test de significativité globale du modèle : la statistique de *Fisher* issue de la régression. Les résultats de la régression sont données par le tableau 9.5. Les coefficients estimés sont tous significatifs avec des *p-valeurs* presque toutes inférieures à 2e-16. La bonne qualité du modèle retenu est garantie par le R^2 qui est de 0.9386, le R^2 ajusté qui est de 0.9233 et la *p-valeur* globale du modèle qui est inférieure à 2.2e-16.


9.3.2 Méthode *seasonal decomposition of times series by loess(STL)*

La méthode STL proposée par (Cleveland R.B et al [2008]) consiste à estimer les composantes déterministes par des régressions locales. Cette méthode consiste en une procédure itérative qui utilise des fonctions *LOESS* (*LOcally wEighted regreSsion Smoother*) qui ont été initialement proposées par (Cleveland S.W. [1979]). Chaque point de la série est remplacé par une valeur issue d'une régression sur les points de son voisinage, affectés d'une pondération décroissante avec l'éloignement par rapport au point d'estimation. Pour plus de détails sur cette méthode, nous renvoyons le lecteur à la thèse d'Anne JACQUIN [23].

Sous  la mise en œuvre de cette méthode se fait par la fonction *stl()* du *package stats*. Les résultats obtenus par cette méthode pour notre série chronologique sont donnés par la figure 9.8.

9.3.3 Méthode moyenne mobile (MM)

Dans cette partie, nous considérons toujours la série des restes à charge moyens des paires de verres optiques. Le principe de la présente méthode est d'appliquer à la série une transformation (généralement simple) qui permet de conserver uniquement la composante que l'on souhaite estimer. Dans *Série Temporelle Et Modèles Dynamiques*[1995], Christian GOURIEROUX ET Alain MONFORT [13] donnent une présente théorique très détaillée de cette méthode.

Sous  la mise en œuvre de cette méthode se fait par la fonction *décompose()* du *package stats*. Les résultats obtenus par cette méthode sont présentés par la figure 9.9.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.130e+02	1.464e+00	213.762	< 2e-16 ***
<i>trend</i>	9.617e-01	1.072e-01	8.972	9.03e-14 ***
<i>trend2</i>	-1.667e-02	2.561e-03	-6.508	< 2e-16 ***
<i>trend3</i>	1.088e-04	1.737e-05	6.263	< 2e-16 ***
<i>season2</i>	-2.046e+00	1.396e+00	-1.465	< 2e-16 ***
<i>season3</i>	-3.343e+00	1.397e+00	-2.394	< 2e-16 ***
<i>season4</i>	-3.147e+00	1.397e+00	-2.252	< 2e-16 ***
<i>season5</i>	-5.681e+00	1.398e+00	-4.063	0.000111 ***
<i>season6</i>	-2.818e+00	1.399e+00	-2.014	< 2e-16 ***
<i>season7</i>	3.585e-01	1.401e+00	0.256	< 2e-16 ***
<i>season8</i>	-3.453e+00	1.402e+00	-2.462	< 2e-16 ***
<i>season9</i>	-1.212e+01	1.404e+00	-8.631	4.27e-13 ***
<i>season10</i>	-2.804e+00	1.407e+00	-1.993	< 2e-16 ***
<i>season11</i>	-5.482e+00	1.409e+00	-3.890	0.000204 ***
<i>season12</i>	2.718e-01	1.412e+00	0.193	< 2e-16 ***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error : 2.792 on 81 degrees of freedom
Multiple R-squared : 0.9386, Adjusted R-squared : 0.9233
F-statistic : 57.64 on 14 and 81 DF, p-value : < 2.2e-16

TABLE 9.5 – Significativité du modèle retenu.

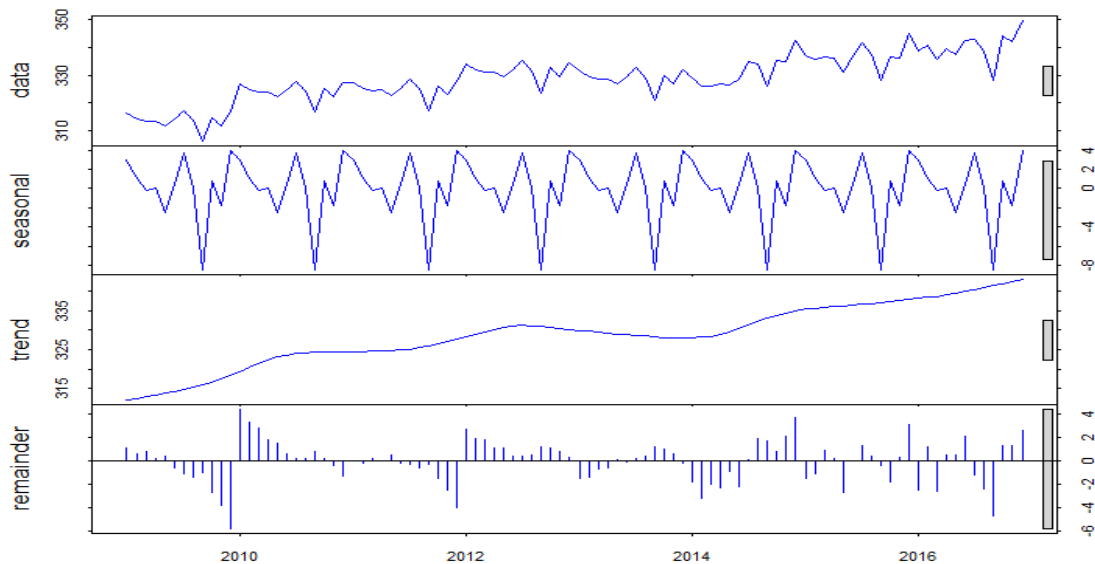


FIGURE 9.8 – Décomposition *stl* de la série des restes à charge moyens des paires de verres optiques.

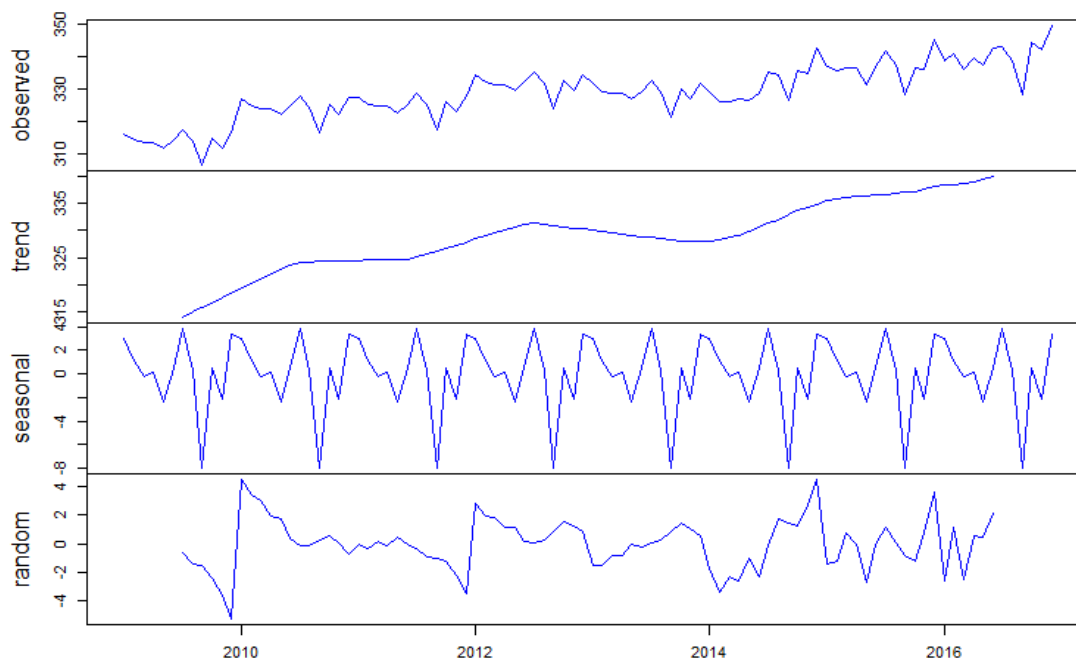


FIGURE 9.9 – Décomposition *MM* de la série des restes à charges moyens des paires de verres optiques.

9.4 Comparaison des modèles de décomposition et prévision

Le but de cette section est de choisir le meilleur modèle parmi les trois modèles de décomposition présentés dans la section précédente et de l'utiliser pour faire des prévisions des valeurs 2017 de la chronique des restes à charge moyens des paires de verres optiques. Nous allons réaliser quelques analyses graphiques et des tests sur les résidus. Pour choisir la décomposition la plus adéquate, nous allons comparer les graphes des résidus, les *qq-plots*, les autocorrélogrammes et les résultats de quelques tests statistiques.

9.4.1 Les autocorrélogrammes

L'examen des autocorrélogrammes (cf. figure 9.10) met en évidence l'existence de corrélations non nulles entre les résidus de Buys-Ballot. En effet, les pics observés sont majoritairement à l'extérieur de l'intervalle de nullité (intervalle à 95% autour de zéro) des coefficients d'autocorrélation, ce qui conduit à rejeter l'hypothèse de blancheur des résidus de Buys-Ballot. Les autocorrélogrammes des résidus STL et moyenne mobile ne présentent pas de tendance et leurs autocorrélogrammes tendent rapidement vers zéro, ce qui nous amène à considérer comme blanc les résidus issus des décompositions STL et moyenne mobile.

Remarque : L'autocorrélogramme d'un processus $(X_t)_t$ est une représentation de sa fonction d'autocorrélation. Cette fonction est définie de \mathcal{Z} vers \mathcal{R} et est donnée par : $\forall h \in \mathcal{Z}, \rho(h) = \frac{\tau(h)}{\tau(0)}$ avec $\tau(h) = Cov(X_t, X_{t-h})$.

Afin de faire un choix entre les méthodes STL et moyenne mobile, nous réalisons dans la sous-section suivante quelques tests sur les résidus.

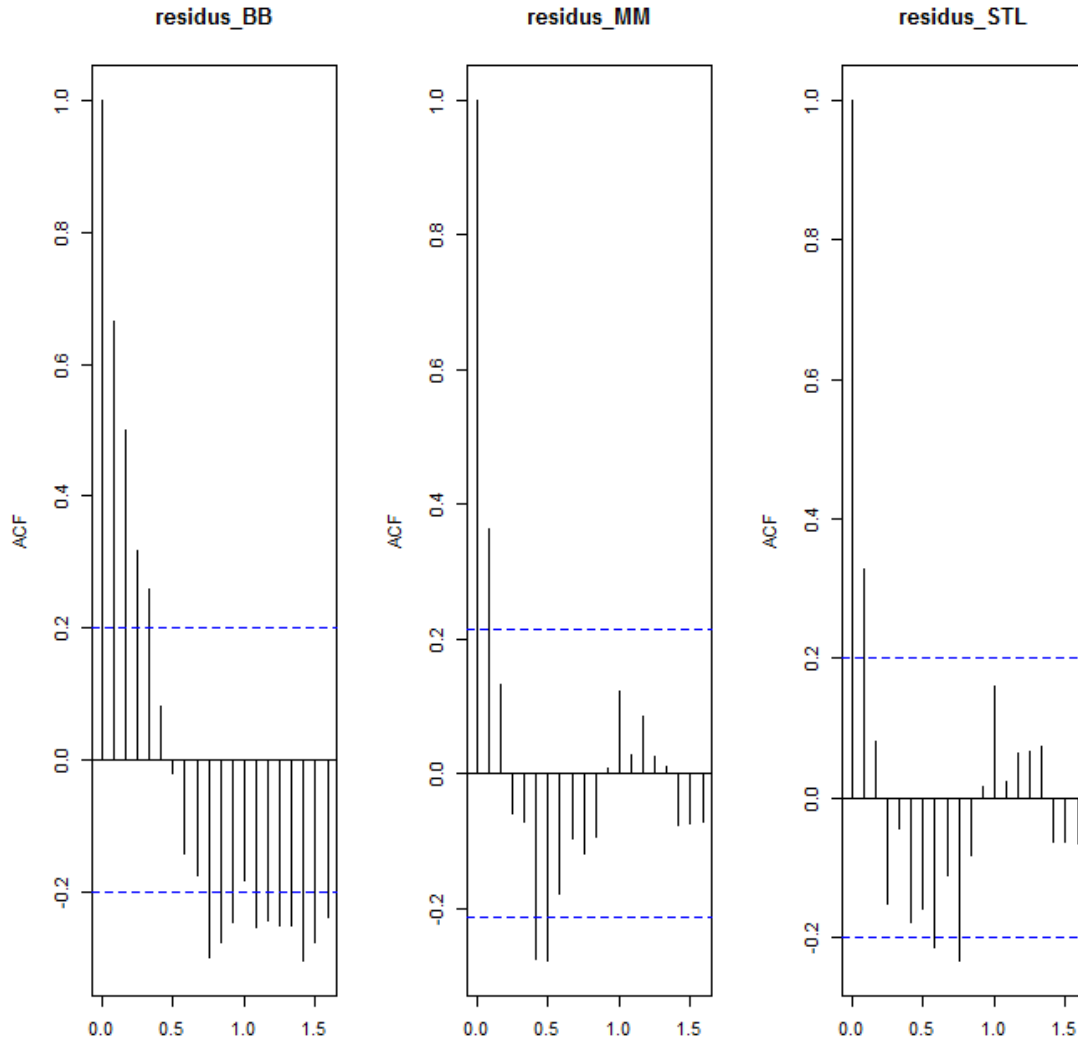


FIGURE 9.10 – Autocorrélogrammes des résidus issus des trois méthodes de décomposition.

9.4.2 Tests sur les résidus

Les tests que nous allons réaliser sur les résidus issus des différentes méthodes de décomposition sont les suivants :

— **Test de Shapiro-Wilk :**

Une présentation de ce test de normalité est proposé par la sous section 9.3.1 ; nous y renvoyons le lecteur.

— **Ljung-Box test :**

L'hypothèse nulle de ce test est l'indépendance des résidus et l'alternative est la non-indépendance des résidus. En cas d'indépendance, la statistique $Q_{LB} = T(T+2) \sum_{h=1}^k \frac{\hat{\rho}(h)^2}{T-h}$ du test tend vers zéro et la *p-valeur* vers 1. Cette statistique est calculée sur les k premières estimations des autocorrélations.

— **Test de Kwiatowski-Phillips-Schidt-Shin (KPSS) :**

L'hypothèse nulle de ce test est celle de la stationnarité autour d'une constante ou d'une

tendance déterministe linéaire et l'hypothèse alternative est l'absence de stationnarité. Le calcul de la statistique η_{kpss} est basé sur l'estimateur non centré du moment d'ordre 2 des résidus et sur la fonction d'autocorrélation partielle de ces derniers. Les résidus les plus stationnaires sont ceux pour lesquelles le *trend* (tendance) estimé est le plus bas et la *p-valeur* supérieure à 0,1. Pour plus de détails sur ce test nous renvoyons le lecteur à l'article *Testing the null hypothesis of stationarity against the alternative of a unit root* [29].

— **Test de Durbin Watson :**

L'hypothèse nulle de ce test est l'absence d'autocorrélation des résidus et l'hypothèse l'alternative est la corrélation des résidus. Ce test permet de détecter une autocorrélation d'ordre un des résidus. La statistique de Durbin et Watson est donnée par : $DW = \frac{\sum_{t=2}^T (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\epsilon}_t^2}$. Cette statistique est comprise entre zéro et quatre. L'hypothèse nulle d'absence d'autocorrélation des résidus est admise lorsque la valeur de cette statistique est proche de 2 avec une *p-valeur* proche de 1 [11].

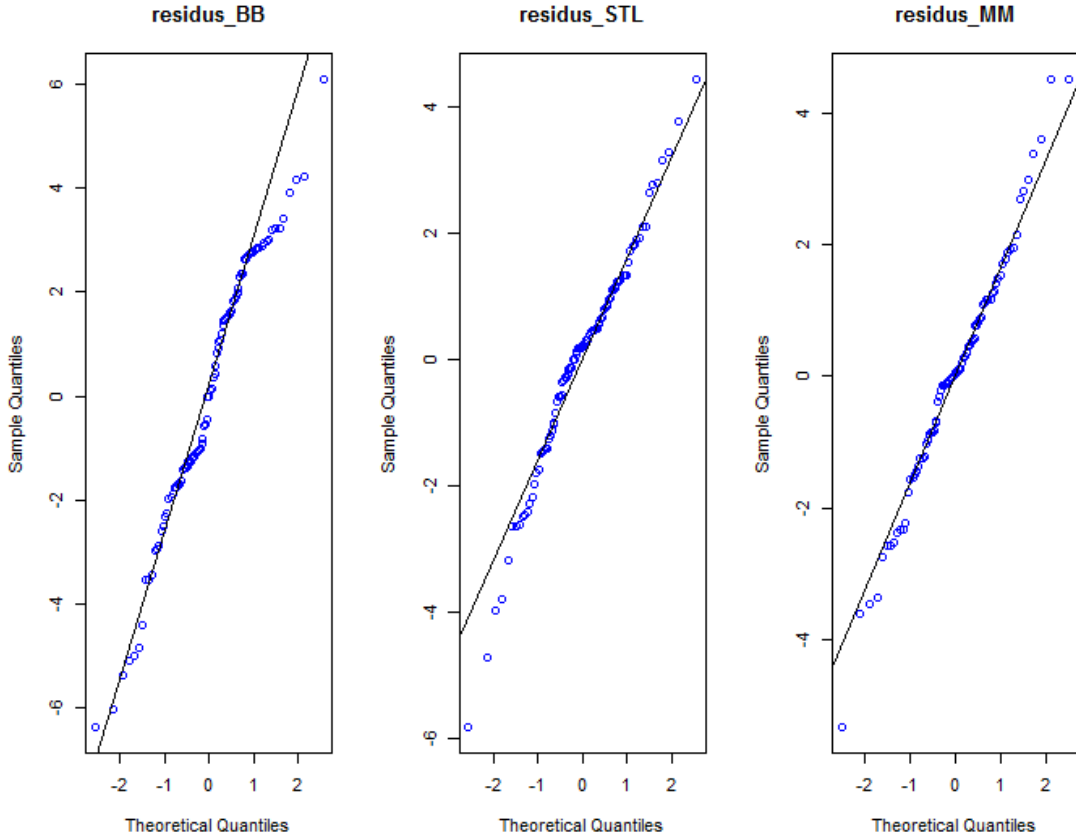


FIGURE 9.11 – *qqplot* des résidus issus des trois méthodes de décomposition.


Les résultats obtenus par ces différents tests sont consignés dans le tableau 9.6. Le test de *Shapiro-Wilk* fournit des statistiques W_n proches de 1 pour les 3 méthodes, on accepte donc la normalité des trois séries de résidus. La normalité de ces résidus est aussi confirmée par le tracé de leurs *qqplots* (cf. figure 9.11) : les quantiles des résidus suivent bien la droite de normalité de Henry. Les résidus obtenus par la méthode STL sont ceux qui minimisent la statistique Q_{LB} et fournissent la *p-valeur* la plus grande. Concernant le *trend* estimé par le test de KPSS, les résidus issus de la méthode STL sont également ceux qui ont le *trend* minimal et la *p-valeur* la plus

élevée. Ce sont donc ces résidus qui rejettent le plus l'hypothèse nulle d'absence de stationnarité de ce test. Comme le montre le tableau 9.6 la statistique DW de Durbin et Watson la plus proche de 2 est celle correspondant aux résidus issus de la décomposition STL, elle est associée à la p -valeur la plus grande parmi les trois estimées pour ce test. D'après cette analyse statistique, les résidus issus de la décomposition par méthode STL sont plus blancs que ceux associés aux méthodes de *Buys-Ballot* (BB) et moyenne mobile. C'est donc cette méthode (la méthode STL) que nous retenons pour les prévisions des valeurs 2017 des restes à charge moyens des paires de verres optiques.

	W_n	P -valeur	Q_{LB}	P -valeur	η_{kpss}	P -valeur	DW	P -valeur
BB	0.97412	0.05425	43.679	3.869e-11	0.067625	0.1	0.6599	1.646e-14
MM	0.97646	0.08136	11.564	6.723e-02	0.032978	0.1	1.2521	0.0001793
STL	0.98772	0.6132	10.699	1.072e-01	0.027265	0.1	1.3151	0.0002602

TABLE 9.6 – Statistiques et p -valeurs des tests sur les résidus des trois méthodes de décomposition.

9.4.3 Prédiction par la méthode STL

Le but de cette sous-section est de proposer une prévision des niveaux de restes à charge moyens des paires de verres optiques pour l'année 2017. A ce jour, *l'open DAMIR* n'a pas encore été complété par l'ajout des dépenses 2017. Les prévisions que propose cette sous-section sont donc des anticipations des niveaux de restes à charge que nous connaissons dans quelques mois. Afin de prédire les restes à charge moyens des paires de verres optiques des 12 mois de l'année 2017, nous nous basons sur la décomposition proposée par la méthode STL. Nous utilisons la fonction `forecast()` du logiciel  pour cette prévision. Les valeurs prédites avec un intervalle de confiance gaussien à 95% sont présentées par le tableau 9.7 et une représentation graphique des prévisions est donnée par la figure 9.12.

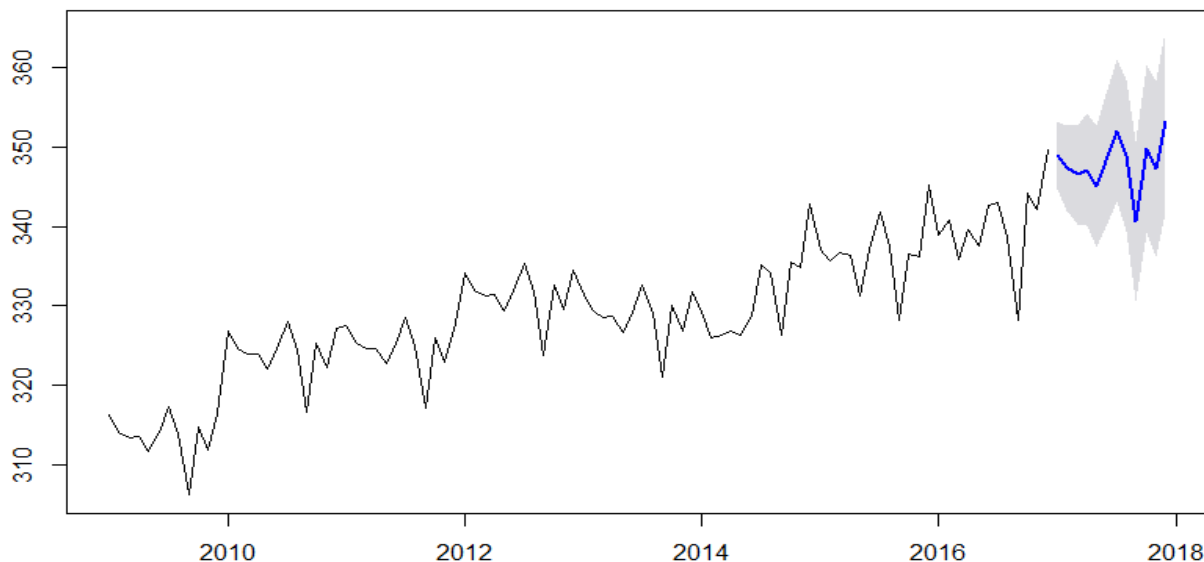


FIGURE 9.12 – Prédiction par la méthode STL des restes à charge moyens des paires de verres optiques.

	Prévision-STL	Lo 95	Hi 95
Jan 2017	348.9210	344.6502	353.1917
Feb 2017	347.3703	342.0374	352.7032
Mar 2017	346.5163	340.3015	352.7312
Apr 2017	347.1351	340.1412	354.1290
May 2017	345.0204	337.3418	352.6990
Jun 2017	348.3140	339.9805	356.6474
Jul 2017	352.0690	343.1145	361.0234
Aug 2017	348.8890	339.3803	358.3977
Sep 2017	340.5725	330.5756	350.5694
Oct 2017	349.7570	339.2316	360.2824
Nov 2017	347.2237	336.2255	358.2218
Dec 2017	353.2663	341.7822	364.7504

TABLE 9.7 – Prévisions par la méthode STL et intervalle de confiance gaussien à 95%.

9.5 Autres méthodes de prévision

L'objectif de cette section est de proposer deux méthodes de prévision qui serviront à challenger la méthode STL. Dans cette section nous intéressons à la prévision par la méthode de lissage de *Hold-Winters* et à la prévision par l'algorithme de *Box-Jenkins*.

9.5.1 Lissage exponentiel de Hold-Winters


Parmi les méthodes de lissage exponentiel, les méthodes de Hold-Winters sont les plus adaptées pour la prévision des valeurs futurs d'une série chronologique qui présente une tendance et une saisonnalité. Elles opèrent des estimations locales de la série désaisonnalisée, de la tendance et de la saisonnalité. Dans la littérature, on retrouve principalement deux versions de méthodes de Hold-Winters : l'une adaptée aux séries multiplicatives et l'autre adaptée aux séries additives. La chronique des restes à charge moyens des paires de verres optiques étant additive, nous nous intéressons au modèle additif de Holt-Winters.

Soit $L = (L_t)_t$ la série désaisonnalisée des restes à charge moyens des paires de verres optiques. Comme dans les précédentes sections notre série de données est notée $Y = (Y_t)_t$, $S = (S_t)_t$ représente sa saisonnalité et $Z = (Z_t)_t$ sa tendance et $\epsilon = (\epsilon_t)_t$ sa composante d'erreur. La prévision à l'horizon h est donnée par :

$$\hat{Y}_{t+h} = (L_t + h \times Z_t)S_{t-s+h}$$

où s est l'ancienneté retenue pour l'estimation du terme en $t+h$, $L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + Z_{t-1})$, $b_t = \beta \times (L_t - L_{t-1}) + (1-\beta) \times Z_{t-1}$ et $S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$. Le choix des paramètres α , β et γ peut être fait en minimisant un critère des moindres carrés des erreurs de prévision.

Une initialisation possible de cet algorithme est donnée par : $\forall t = 1, \dots, r$ $L_t = \frac{Y_1 + \dots + Y_r}{r}$, $Z_t = \frac{1}{r}(\frac{Y_{1+r}-Y_1}{r} + \dots + \frac{Y_{2r}-Y_r}{r})$ et $S_t = \frac{Y_t}{L_t}$.

Sous  la mise en œuvre du lissage additif de Holt-Winter peut être réalisée à l'aide de la fonction *holtwinters()*. Les résultats de la prévision (avec un intervalle de confiance gaussien à 95%) sont donnés par le tableau 9.8 et une représentation graphique est donnée par la figure 9.13.

	Prévision-HW	I.inf 95	I.sup 95
Jan 2017	343.4290	337.6972	349.1608
Feb 2017	345.1754	338.3991	351.9518
Mar 2017	344.8401	337.1360	352.5442
Apr 2017	348.8030	340.2496	357.3563
May 2017	346.8852	337.5394	356.2310
Jun 2017	351.8748	341.7797	361.9700
Jul 2017	354.5313	343.7207	365.3419
Aug 2017	351.1121	339.6135	362.6108
Sep 2017	341.9825	329.8184	354.1467
Oct 2017	353.6618	340.8509	366.4727
Nov 2017	350.1403	336.6986	363.5820
Dec 2017	356.1529	342.0940	370.2119

TABLE 9.8 – Prévision des restes à charge moyens des paires de verres optiques par le méthode additif de Holt-Winter et intervalle de confiance gaussien à 95%.

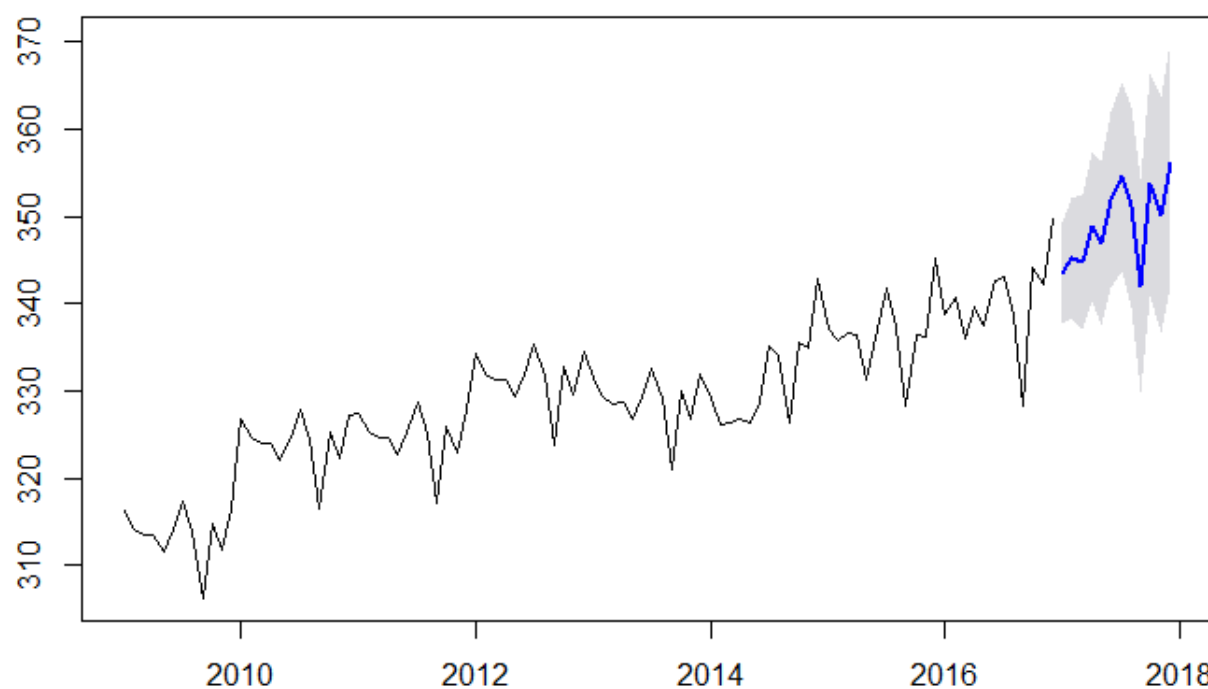


FIGURE 9.13 – Prévision des restes à charge moyens des paires de verres optiques par le modèle additif de Holt-Winter.

9.5.2 Prévision par l'ajustement d'un modèle SARIMA

Une série chronologique Y_t suit un processus SARIMA (*Seasonnal AutoRegressive Integrated Moving Average*) d'ordre $(p, d, q) * (P, D, Q)_s$ si elle est saisonnière de période s et qu'elle peut s'écrire sous la forme suivante :

$$\Phi_1(B)\Phi_2(B^2)(1-B)^d(1-B^s)^DY_t = \Theta_1(B)\Theta_2(B^2)\epsilon_t$$

où Φ_1 est un polynôme de degré p , Φ_2 est un polynôme de degré P , Θ_1 est un polynôme de degré q , Θ_2 est un polynôme de degré Q , ϵ_t est un bruit blanc et B est un opérateur de recul tel que $\forall t \quad BY_t = Y_{t-1}$.

Dans cette sous-section, la détermination du modèle SARIMA qui ajuste le mieux la série des restes à charge moyens des paires de verres optiques est réalisée par l'algorithme de *Box-Jenkins*. Cet algorithme à quatre grandes étapes :

- La stationnarisation de la série,
- L'identification et l'estimation des paramètres adéquats,
- La validation du modèle retenu,
- La prévision des valeurs futures.

Stationnarité de la chronique

L'analyse de la fonction d'autocorrélation empirique $\hat{\rho}(h)$ révèle la non stationnarité de la série des restes à charge moyens des paires de verres optiques. Un regard sur la figure 9.14 montre l'absence d'indépendance entre les termes de notre série car $\forall h \leq 0$ les corrélations sont toutes à l'extérieur de l'intervalle de nullité. Cet autocorrélogramme présente également une tendance et l'autocorrélogramme partiel révèle l'existence d'une saisonnalité : la série des restes à charge moyens des paires de verres optiques n'est donc pas stationnaire.

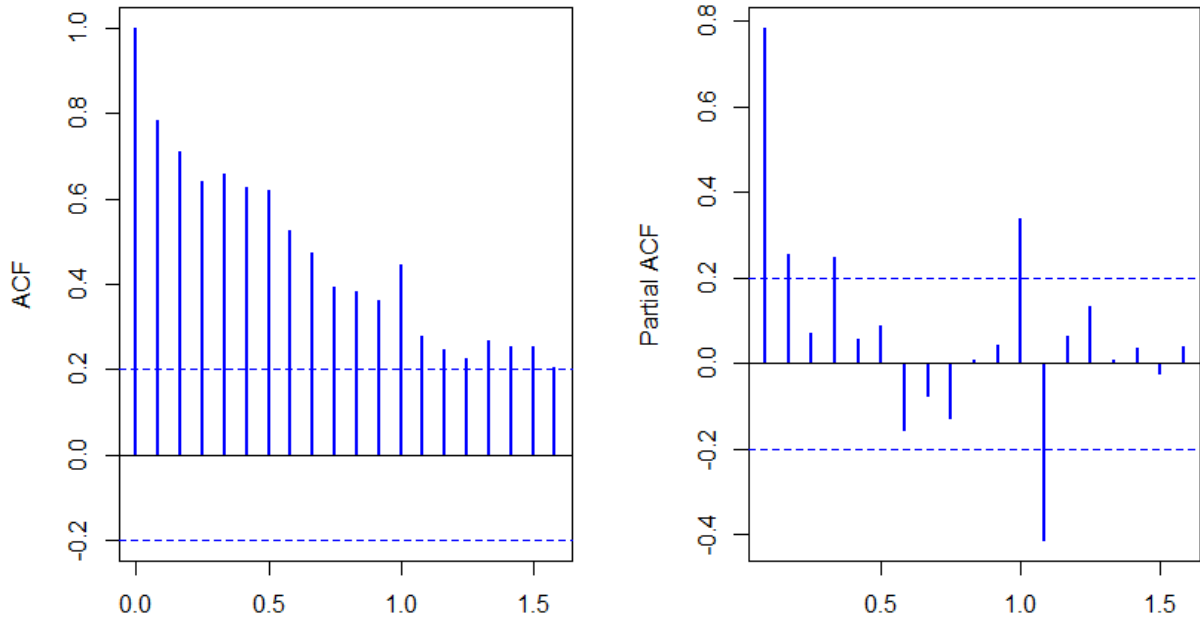


FIGURE 9.14 – Fonctions d'autocorrélations.

L'analyse des résultats du *Ljung-Box test* d'indépendance confirme la présence d'une forte dépendance entre les termes de la série car la statistique du test est très élevée et la p-valeur est quasiment nulle (cf. tableau 9.9). Concernant la stationnarité de la série, le test de KPSS fournit

une *p-valeur* faible (inférieure à 0,1) et une estimation élevée du *trend*. Nous rappelons que, pour ce test, en cas de stationnarité la statistique qui représente la tendance de la série est proche de zéro et sa *p-valeur* supérieure à 0,1. Cet examen analytique confirme la non stationnarité de la série des restes à charge moyens des paires de verres optiques.

	statistique	<i>p-valeur</i>
Ljung-Box (indépendance)	86.711	6.661e-17
KPSS (stationnarité)	2.6293	0.01

TABLE 9.9 – Tests d'indépendance et de stationnarité de la série des restes à charge moyens des paires de verres optiques.

Considérons à présent la série $X = (X_t)_t$ définie par $X_t = Y_t - Y_{t-1}$ où $Y = (Y_t)_t$ est notre chronique d'intérêt. La figure 9.15 donne une représentation de X , cette série ne présente apparemment aucune tendance. Cette figure montre également que les fonctions d'autocorrélations tendent rapidement vers zéro, ce qui suggère que la série X est stationnaire. Cette stationnarité semble être confirmée par le test KPSS dont la statistique est 0,02688 et *p-valeur* 0,1 (cf. tableau 9.10). Nous considérons donc la série X comme stationnaire. Le *Ljung-Box test* donne une statistique faible ($Q_{LB} = 7,57$) et une *p-valeur* supérieur à 0,05. Ces résultats nous conduisent à admettre l'indépendance des termes de la série X .

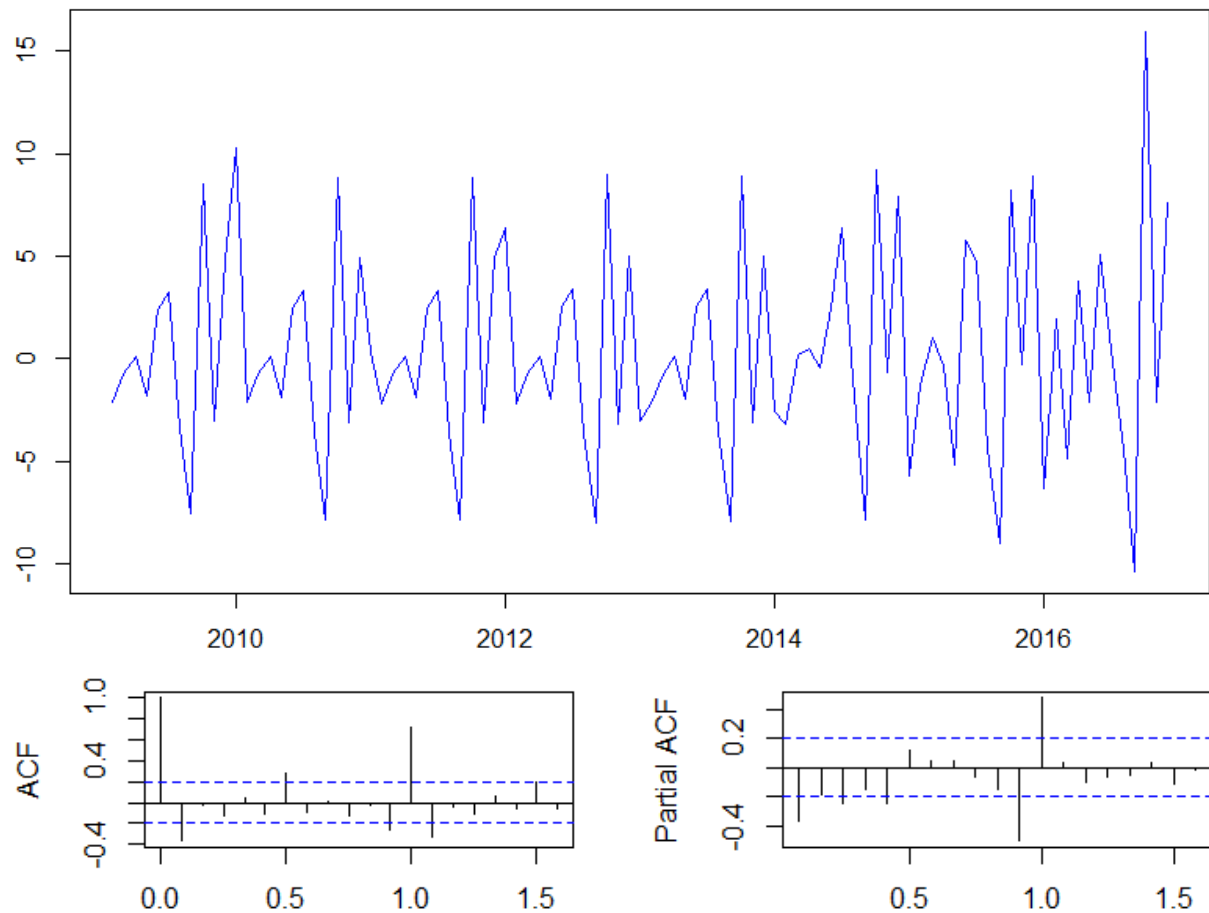



FIGURE 9.15 – Série des restes à charge moyens des paires de verres optiques différenciée à l'ordre 1.

	statistique	<i>p-valeur</i>
Ljung-Box (indépendance)	7.5754	0.07917
KPSS (stationnarité)	0.02688	0.1

TABLE 9.10 – Tests d’indépendance et de stationnarité de la série différenciée.

Identification et estimation des paramètres

Nous nous intéressons ici à la détermination des paramètres du processus SARIMA qui ajuste le mieux la série des restes à charge moyens des paires de verres optiques. Pour cela nous utilisons l’algorithme de Hyndman et Khandakar qui combine des tests de racine unitaires, de minimisation du critère d’information d’Akaike et du maximum de vraisemblance. Sous  cet algorithme est implémenté par la fonction `auto.arima()`. Cet algorithme détermine les paramètres de différenciation d et D par une série de tests de stationnarités du type KPSS et les valeurs des paramètres p , q , P et Q sont déterminés par une analyse basée sur le critère d’information d’Akaike. Parmi tous les modèles possibles, cette analyse conduit à retenir les paramètres du modèle qui minimise ce critère.

Remarque : Les critères d’information d’Akaike associé à un modèle à k paramètres et de vraisemblance maximisée \tilde{L} est défini par : $AIC = 2k - 2\ln(\tilde{L})$. Avec ce critère, la déviance $-2\ln(\tilde{L})$ est pénalisée par 2 fois le nombre de paramètres. Ce critère représente donc un compromis entre le **biais** d’estimation qui diminue avec le nombre de paramètres et la **parcimonie** qui est la volonté de décrire les données avec le plus petit nombre de paramètres possibles.

La mise en œuvre de l’algorithme de Hyndman et Khandakar a fourni le résultat suivant : $SARIMA(1, 0, 0)(1, 1, 0)[12]$. Ainsi, $p = 1$, $d = 0$, $q = 0$, $P = 1$, $D = 1$, $Q = 0$ et $s = 12$.

Validation du modèle

Afin de valider le modèle que nous avons obtenu, nous vérifions la blancheur des résidus. Pour cela nous observons les résidus, leur autocorrélogramme, leur *qqplot* et les *p-valeurs* du test de *Ljung-Box*. La figure 9.16 prouve la blancheur des résidus estimés par l’ajustement du modèle $SARIMA(1, 0, 0)(1, 1, 0)[12]$ à la série des restes à charge moyens des paires de verres optiques. En effet un regard sur cette figure confirme la normalité des résidus car leurs quantiles sont alignés majoritairement avec la droite de Henry. Aussi, l’autocorrélogramme met en évidence l’absence de corrélation entre les résidus car $\forall h \hat{\rho}(h)$ est dans l’intervalle de nullité. De plus, les *p-valeurs* du *Ljung-Box test* sont quasiment toutes au dessus de la valeur seuil 0,05. Il y a donc une bonne adéquation entre le modèle $SARIMA(1, 0, 0)(1, 1, 0)[12]$ et la série des restes à charge moyens des paires de verres optiques.

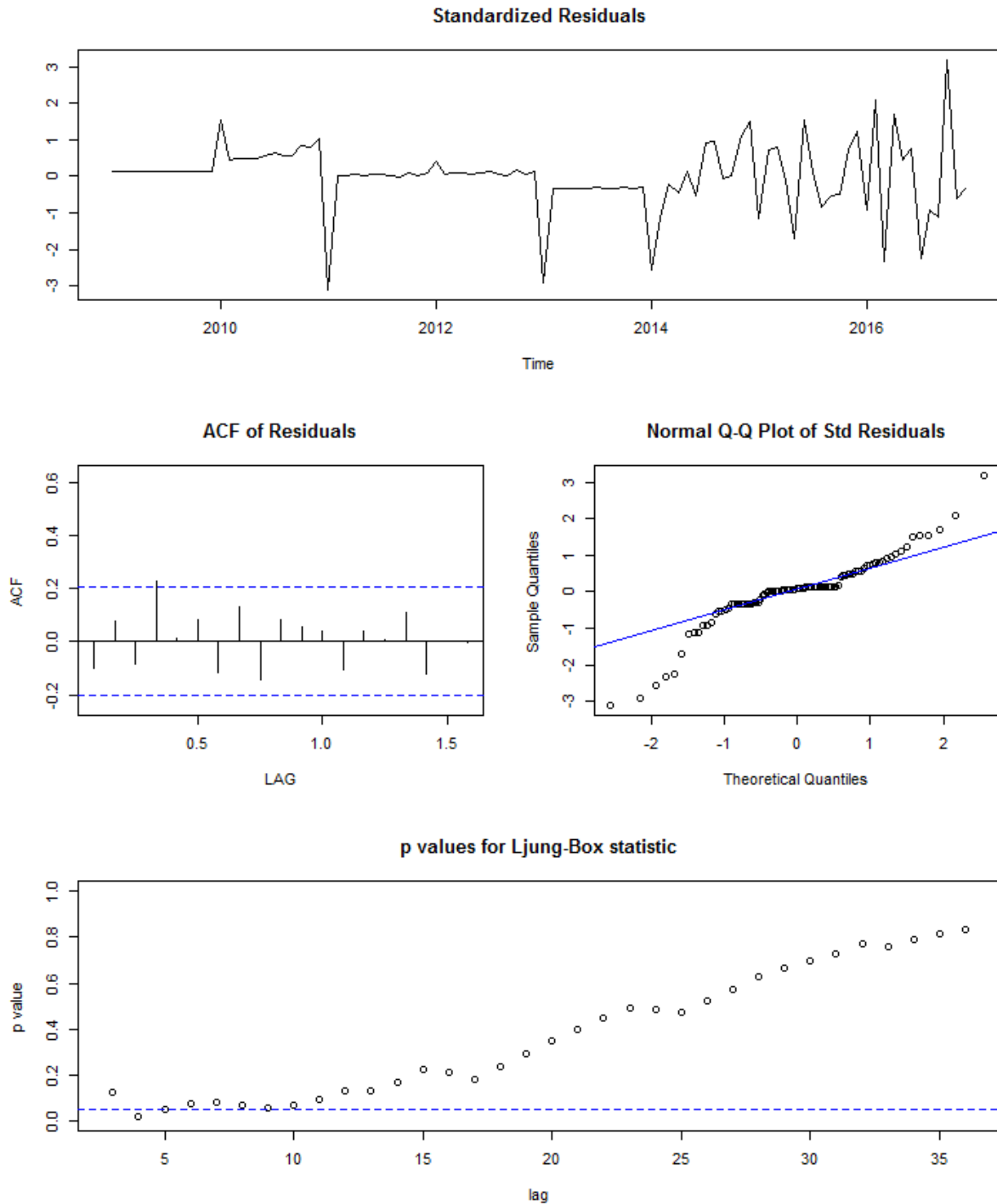


FIGURE 9.16 – Analyse des résidus issus de l’ajustement du modèles $SARIMA(1, 0, 0)(1, 1, 0)[12]$ à la série des restes à charge moyens des paires de verres optiques.

Prévision

A partir du modèle $SARIMA(1, 0, 0)(1, 1, 0)[12]$, nous avons estimé les 12 prochaines valeurs de notre série chronologique. Les résultats de cette estimation sont donnés dans le tableau 9.11 avec un intervalle de confiance gaussien à 95% et sont graphiquement représentés par la figure 9.17.

	Prévision-BJ	I.inf 95	I.sup 95
Jan 2017	343.6604	339.5531	339.5531
Feb 2017	343.8410	338.4154	338.4154
Mar 2017	342.1104	335.8816	335.8816
Apr 2017	343.6480	336.8826	336.8826
May 2017	339.8790	332.7400	332.7400
Jun 2017	345.3269	337.9219	337.9219
Jul 2017	348.0720	340.4749	340.4749
Aug 2017	343.6506	335.9135	335.9135
Sep 2017	333.9941	326.1543	326.1543
Oct 2017	345.7819	337.8665	337.8665
Nov 2017	344.6653	336.6941	336.6941
Dec 2017	352.9452	344.9326	344.9326

TABLE 9.11 – Prévisions par le modèle $SARIMA(1,0,0)(1,1,0)[12]$ issu de l'algorithme de *Box – Jenkins* et intervalle de confiance gaussien à 95%.

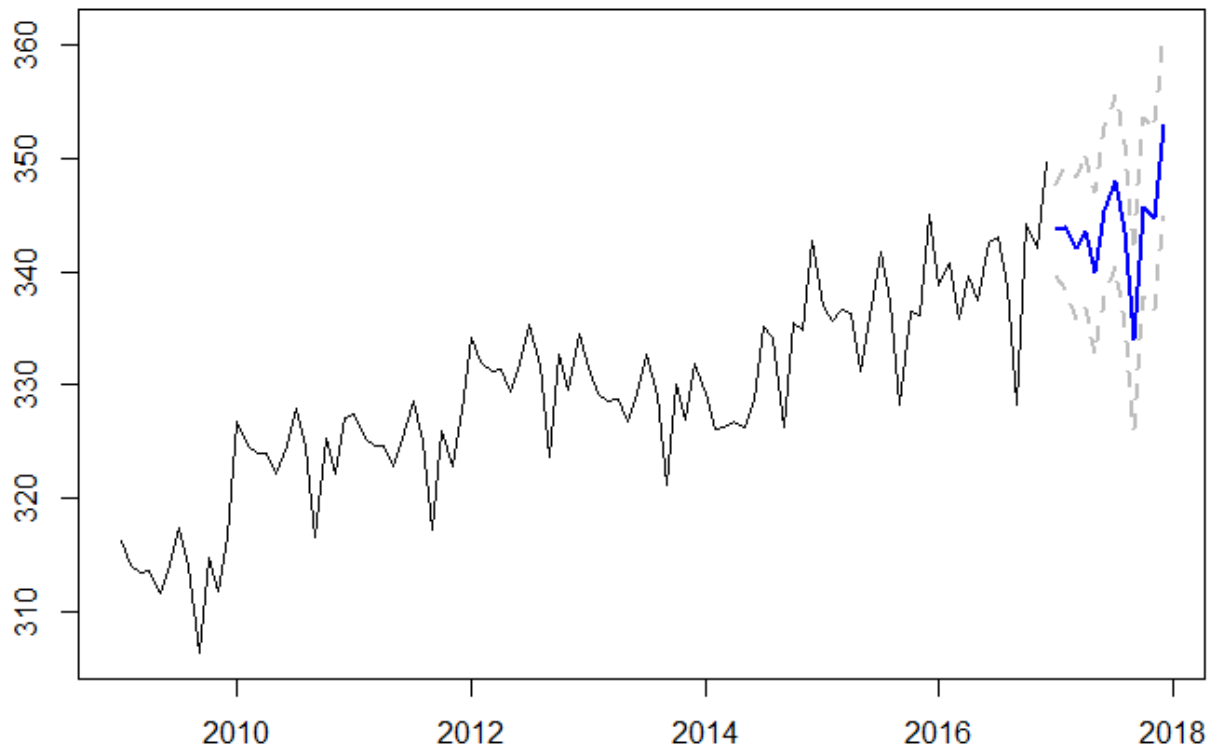


FIGURE 9.17 – Prévision des restes à charge moyens par acte des paires de verres optiques par le modèle $SARIMA(1,0,0)(1,1,0)[12]$.

9.6 Choix du meilleur modèle de prévision

Dans les sections et sous-sections précédentes, nous avons présenté trois modèles pouvant être utilisés pour la prévision des valeurs futures de notre série chronologique des restes à charge moyens des paires de verres optiques : le modèle STL, le modèle de Hold-Winters(HW) et le modèle $SARIMA(1, 0, 0)(1, 1, 0)[12]$. Le but de cette partie est de déterminer lequel de ces trois modèles fournit la prévision la plus juste. Pour cela, nous ajustons ces modèles aux valeurs 2009 à 2015 de notre chronique et nous estimons les 12 termes associés aux mois de l'année 2016. La meilleure prévision est celle qui minimise la racine carrée de l'erreur quadratique moyenne (RMSE ou *Root Mean Square Error*), l'erreur quadratique moyenne étant la moyenne arithmétique des carrés des écarts entre les prévisions et les observations 2016.

Les résultats obtenus sont donnés par le tableau 9.12 et graphiquement présentés par la figure 9.18. La prévision donnée par le modèle STL a un RMSE de 5,16. Celle donnée par le modèle de Holt-Winters à un RMSE de 7,13 et celle obtenue par le modèle $SARIMA(1, 0, 0)(1, 1, 0)[12]$ issu de l'algorithme de *Box-Jenkins* à un RMSE de 3,80. **Le modèle $SARIMA(1,0,0)(1,1,0)[12]$ est donc le plus adapté pour la prévision des valeurs futures de la série chronologique des restes à charge moyens des paires de verres optiques.** Par le critère de la *Root Mean Square Error*, les prévisions données par le tableau 9.11 sont les meilleures anticipations des restes à charge moyens des paires de verres optiques pour l'année 2017.

	prévision-STL	prévision-HW	prévision-BJ	Vraie valeur
Jan 2016	345.9316	341.4093	338.5271	338.82
Feb 2016	343.7652	342.8231	335.6670	340.75
Mar 2016	343.4654	345.3775	336.5955	335.89
Apr 2016	343.6168	346.5732	336.7942	339.64
May 2016	341.2569	344.5673	334.8034	337.53
Jun 2016	344.4271	349.0527	338.5462	342.61
Jul 2016	348.4389	352.9346	345.1635	343.03
Aug 2016	345.4004	348.6822	343.7993	338.63
Sep 2016	337.6074	339.7414	335.7815	328.26
Oct 2016	346.3849	348.1366	344.1564	344.20
Nov 2016	343.5957	346.2996	344.8170	342.09
Dec 2016	349.1591	352.0978	354.1271	349.66
RMSE	5.162511	7.131183	3.804691	/

TABLE 9.12 – Choix du meilleur modèle de prévision.

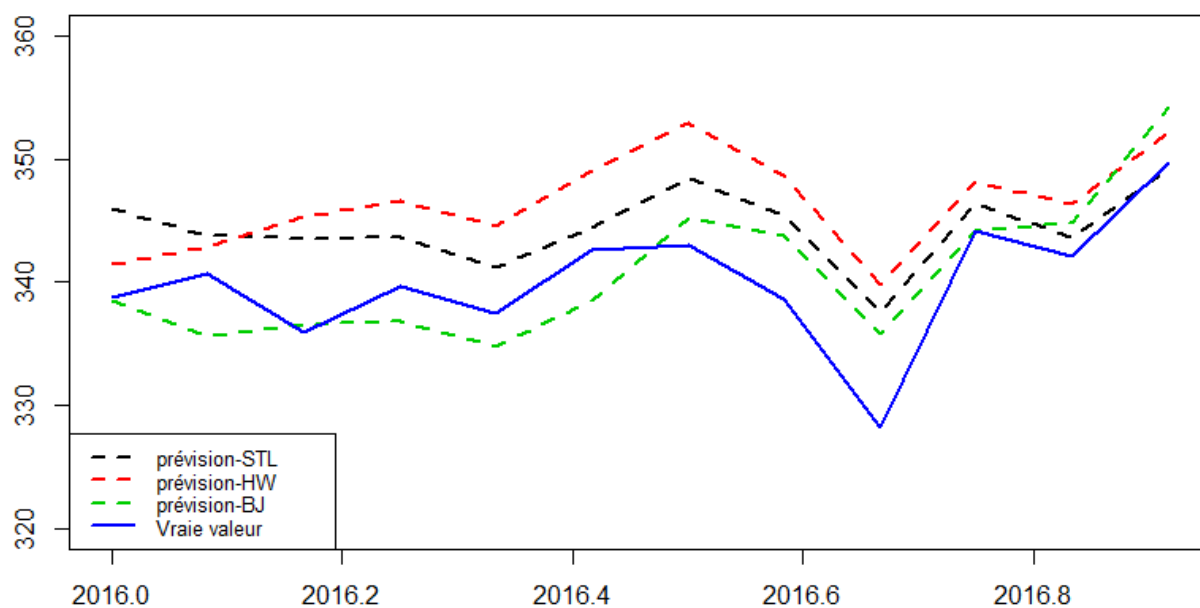


FIGURE 9.18 – Prévisions 2016 et valeurs réelles des restes à charge moyens des paires de verres optiques.

Partant de l'historique de données (2009 à 2016) et des estimations présentées par le tableau 9.11, nous estimons une augmentation de 1,2% par an entre 2009 et 2017 des montants de restes à charge sécurité sociale des paires de verres optiques. Cette hausse s'explique principalement par l'inflation (1,03% par an en moyenne de 2009 à 2017) et par la sophistication des verres. Qu'ils s'agissent des données de 2009 à 2016 ou des prévisions 2017, toutes les valeurs de la chroniques des restes à charges moyen par paires de verres optiques sont comprises dans la fourchette de remboursement imposée par la réforme des contrats responsables. La figure 9.19 nous rappelle cette fourchette. Sachant que cette réforme limite la prise en charge à un équipement d'optique tous les 2ans, le "reste à charges zéro" pour les verres optiques est donc quasiment réalisé à ce jour par les dispositions contrat responsable.

TYPE D'ÉQUIPEMENT	CORRESPONDANCE	FORFAIT MINIMUM ANI (Contrat collectif obligatoire)	FORFAIT MINIMUM Contrat responsable (Contrat individuel)	FORFAIT MAXIMUM
(a) Équipement à verres simple foyer avec : sphère comprise entre - 6 et + 6 et cylindre ≤ 4	Simple - Simple	100 €	50 €	470 €
(b) Équipement à verres simple foyer avec : sphère > - 6 ou > + 6 ou cylindre > 4 (verres multifocaux ou progressifs)	Complexe - Complexe	200 €	200 €	750 €
(c) Équipement comportant un verre mentionné au (a) et un verre mentionné au (b)	Simple - Complexe	150 €	125 €	610 €
(d) Équipement pour adulte à verres multifocaux ou progressifs sphéro-cylindrique avec : sphère hors zone - 8 ou > + 8, verres multifocaux ou progressifs sphériques avec : sphère hors zone - 4 ou > + 4	Très complexe - Très complexe	200 €	200 €	850 €
(e) Équipement comportant un verre mentionné au (a) et un verre mentionné au (d)	Simple - Très complexe	150 €	125 €	660 €
(f) Équipement comportant un verre mentionné au (b) et un verre mentionné au (d)	Complexe - Très complexe	200 €	200 €	850 €

FIGURE 9.19 – Fourchette des remboursements par types de verres fixée par la réforme des contrats responsables (source : mcommemutuelle.com)

Conclusion de la troisième partie

Les efforts de financement des dépenses de santé faits par la sécurité sociale ne parviennent pas à endiguer la hausse du volume globale des montants restes à charge supportés par les organismes d'assurance maladie complémentaire et des ménages. L'optique, le dentaire et les audio-prothèses sont les trois secteurs principalement responsables des niveaux importants de restes à charge en santé. Ces restes à charge sont fortement liés à la région de résidence, à la tranche d'âge et au genre du bénéficiaire de soins. Afin d'améliorer l'accès aux soins de l'ensemble de la population et en particulier des personnes à faibles revenus, des concertations ont été lancées depuis du 23 janvier 2018 par le ministère de la santé dans le cadre du projet "reste à charge zéro", un projet dont la réalisation d'ici 2022 se fera probablement par la création d'un panier de soins standard intégralement couvert par l'assurance maladie obligatoire et les complémentaires santé. L'objectif de ce projet n'est donc pas la suppression du reste à charge pour toutes les lunettes, prothèses dentaires et audio-prothèses. L'idée d'un panier de soins "reste à charge zéro" semble être partagée par la majorité des acteurs de l'assurance maladie. Toutefois, les concertations étant à leurs débuts, de nombreuses questions sont actuellement sans réponse, notamment celles relatives à l'articulation de ce panier par régimes (régime général et régimes spéciaux) et par rapport aux dispositifs qui existent actuellement : contrat responsable, contrat collectif, couverture maladie universelle complémentaire (CMU-C) et de l'aide à la complémentaire santé (ACS).

Conclusion générale

Ce mémoire s'intéresse particulièrement à l'Open DAMIR (base complète sur les Dépenses d'Assurance Maladie Inter Régimes). Cette base de données brute est une extraction du Système National d'Information Inter Régime de l'Assurance Maladie (SNIIRAM) qui rend compte des dépenses nationale de santé. Elle est constituée de 55 variables (41 qualitatives et 14 quantitatives), de 2,15 Md de lignes et la profondeur de ses données remonte au 1^{er} Janvier 2009. Au format sas7bdat, les données DAMIR disponibles à ce jour font 900 GB.

Les travaux d'exploration et de traitement de ce volumineux jeu de données brut ont été réalisés dans un premier temps à l'aide d'une machine virtuelle sous AWS (Amazon Web Service). Ces travaux se sont poursuivis et terminés sur un serveur local de CNP Assurances. Les principaux langages de programmation utilisés pour ce mémoire sont R et SAS. Afin de traiter les valeurs manquantes, de nombreuses méthodes de complétion ont été mises en œuvre, la méthode LOCF (Last Observation Carried Forward), l'imputation par la moyenne, l'imputation par la médiane, la méthode LOESS (Local regrESSion), la méthode K-NN (K-Nearest Neighbors) et la méthode MiseForest. La qualité de ces modèles de complétion a été analysée à partir d'un échantillon de données DAMIR. Le modèle MissForest s'est avéré meilleur que les autres modèles à chaque pas de complétion. C'est cette méthode basée sur une imputation par les forêts aléatoires que nous avons choisi pour la complétion des données manquantes à la base DAMIR.

A partir des données DAMIR traitées, ce mémoire propose des statistiques descriptives des dépenses réelles de santé, des prestations de la sécurité sociale et des restes à charge de la sécurité sociale par année (de 2009 à 2016), par régions, par tranches d'âge et par genre. Cette description statistique a mise en évidence le fait que sur les 862 actes de soins que compte la base DAMIR, seules 15 actes représentent 67,6% du reste à charge après remboursement de la sécurité sociale en 2016. Elle a également montré que l'Île-de-France a la proportion la plus importante de reste à charge (environ 20% du volume global de reste à charge sécurité sociale en 2015 et 2016). Ce mémoire montre également que les restes à charge sécurité sociale sont plus importants pour les femmes que pour les hommes. En termes de croissance cumulée, ces restes à charge haussent plus vite chez les hommes et chez les femmes.

Les études réalisées dans ce mémoire à partir des données DAMIR ont permis de rendre compte :

- De la forte liaison qui existe entre le reste à charge sécurité sociale, la région de résidence, la tranche d'âge et le genre du bénéficiaire des soins. Ces liaisons sont quantifiées dans ce mémoire par le V de Cramer.
- De l'importance de la région de résidence, de la tranche d'âge et du genre du bénéficiaire des soins vis-à-vis du reste à charge à charge sécurité sociale. Les mesures d'importance ont été réalisées par un arbre de régression du type CART.
- De l'impact du reste à charge sur la demande de soins. La méthode d'ajustement intro-

duite par Nelder and Wedderburn [1972] dans le cadre des modèles linéaires généralisés (GLM) a rendu compte du fait que, étant par construction agrégée, les données DAMIR ne rendent pas compte du comportement de l'individu. Les résultats de modélisation à partir des données DAMIR ont montré qu'une hausse d'un euro de reste à charge sécurité sociale entraîne une faible augmentation du nombre moyen d'actes de soins dentaire. Il s'agit ici d'un résultat qui ne rend pas compte de l'intuition et qui est contraire à ceux donnés par de nombreuses analyses économiques de la santé. La seconde limite des données DAMIR est le fait qu'elles ne fournissent pas d'informations sur les prestations de l'assurance maladie complémentaire. Ces prestations auraient permis de déterminer les restes à charge réels des assurés sociaux car la majorité des français bénéficient d'une complémentaire santé.

Afin de rendre compte de l'effet de la part des dépenses de santé supportée par les ménages sur la demande de soins, nous avons utilisé les données d'un portefeuille de CNP Assurances. Nous nous sommes intéressés aux assurés vivants en Île-de-France. Les résultats des modèles GLM montrent qu'une hausse d'un euro du reste à charge après remboursement sécurité sociale et remboursement de l'assurance complémentaire entraîne une diminution de la demande de soins quel que soit la tranche d'âge et le genre.

- Et du fait que les restes à charge des prothèses dentaires, des verres optiques et des audioprothèses représentent des quantités importantes qui seraient à l'origine du renoncement aux soins des personnes à faibles revenus. Contrairement à ce que pourrais penser certains, le projet reste à charge zéro n'ambitionne pas de supprimer les restes à charge pour tous les types de verres, de monture, de prothèses dentaires et d'audioprothèses. La réalisation de ce projet d'ici 2022 se fera probablement par la définition du panier de soins standard. Dans le cas particulier des verres optiques, une analyse chronologique réalisée dans ce mémoire montre que depuis 2009, les restes à charge moyen par paire de verres optiques sont à la hausse, une hausse qui s'explique par l'inflation et la sophistication des verres. Cette hausse n'entrave pas la réalisation du projet « reste à charge zéro » pour les verres optiques car toutes les valeurs de la chronique des restes à charge moyen par paires de verres optiques sont comprises dans la fourchette de remboursement imposée par la réforme des contrats responsables. Sachant que ce dispositif limite la prise en charge à 1 équipement d'optique tous les 2 ans, le "reste à charge zéro" pour les verres optiques est donc à ce jour quasiment réalisé par les dispositions contrat responsable.

Par ce mémoire, nous avons levé un pan de voile sur le potentiel réel de la base DAMIR. Nous pensons que l'enrichissement progressif des données DAMIR et leur réutilisation donnera lieu à de nombreux projets et études. Cette base pourrait être un vecteur d'innovation en matière de stratégie de gestion du risque santé.

Bibliographie

- [1] Agir contre le renoncement aux soins, diagnostic, solutions et déploiement. https://www.ameli.fr/fileadmin/user_upload/documents/DP_Lutte_contre_le_renoncement_aux_soins_-_VDEF.pdf, 2017.
- [2] G. TROUESSI et E. CORDONNIER A. EL KALAM, Y. DESWARTE. Une démarche méthodologique pour l'anonymisation de données personnelles sensibles. <http://sondage.sstic.org/SSTIC04/Anonymisation/SSTIC04-article-Abou-Anonymisation.pdf>.
- [3] Nicolas BARADEL. *LANGAGE R : APPLICATION A LA STATISTIQUE, A L'ACTUARAT ET A LA FINANCE*. 2015.
- [4] B.DELYON. Estimation parametrique. <https://perso.univ-rennes1.fr/bernard.delyon/param.pdf>, 2017.
- [5] P-L. BRAS. Rapport sur la gouvernance et l'utilisation des donnees de sante. http://solidarites-sante.gouv.fr/IMG/pdf/Gouvernance_et_utilisation_des_donnees_de_sante_septembre_2013.pdf, 2013.
- [6] K. Tribouley C. Genest, E.Masiello. Estimating copula densities through wavelets. <https://hal.archives-ouvertes.fr/hal-00257425v1/document>, 2008.
- [7] A. CHARPENTIER. Beta kernels and transformed kernels applications to copulas and quantiles. <https://fr.slideshare.net/charthur/slides-laval-stat-avril-2011>, 2011.
- [8] W.A. Fuller D.A. Dickey. Distribution of the estimators for autoregressive time series with a unit root. <http://debis.deu.edu.tr/userweb//onder.hanedar/dosyalar/kpss.pdf>, 1979.
- [9] data.gouv.fr. Descriptif open damir. https://www.data.gouv.fr/fr/datasets/open-damir-base-complete-sur-les-depenses-dassurance-maladie-inter-regimes/#_.
- [10] Université de Toulouse. Imputation de données manquantes. <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-idm.pdf>.
- [11] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. https://www.researchgate.net/publication/243644934_Distribution_of_the_Estimators_for_Autoregressive_Time_Series_With_a_Unit_Root, 1979.
- [12] des Études et des Statistiques Direction de la Stratégie. Le système national d'information interrégimes de l'assurance maladie - sniiram. https://www.ameli.fr/fileadmin/user_upload/documents/Presentation_du_Sniiram.pdf, 2015.
- [13] Christian GOURIEROUX et Alain MONFORT. *SERIES TEMPORELLES ET MODELES DYNAMIQUES*. 1995.
- [14] A. CHARPENTIER et C. DUTANG. L'actuariat avec \mathbb{R} . https://cran.r-project.org/doc/contrib/Charpentier_Dutang_actuariat_avec_R.pdf, 2012.
- [15] Michael J. Kane et John W. Emerson. The bigmemory project. <https://cran.r-project.org/web/packages/bigmemory/vignettes/Overview.pdf>, 2010.

- [16] Junqiang Lui et Ke Wang. On optimal anonymisation for l-diversity, 2010.
- [17] L. BELLANGER et R. TOMASSONE. *Exploration de données et méthodes statistiques*. ellipses.
- [18] Little R.J et Rubin D.B. Statistical analysis with missing data, wiley serie in probability and statistics, 1987.
- [19] Nicolas Célant et Thierry Rochereau. L'enquête santé européenne - enquête santé et protection sociale (ehis-esps) 2014. <http://www.irdes.fr/recherche/rapports/566-enquete-sante-europeenne-ehis-enquete-sante-et-protection-sociale-esps-2014.pdf>, 2017.
- [20] Norbert Fouemkeu. Modélisation de l'incertitude sur les trajectoires d'avions. <https://tel.archives-ouvertes.fr/tel-00710595/document>, 2012.
- [21] B. GHATTAS. Importance des variables dans les méthodes cart. <http://lumimath.univ-mrs.fr/~ghattas/mypapers/importance.pdf>.
- [22] John Hull. *Options, futures et autres actifs dérivés*. Pearson, 8ème édition.
- [23] A. JACQUIN. Dynamique de la végétation des savanes en lien avec l'usage des feux à madagascar. analyse par série temporelle d'image de télédétection. <http://oatao.univ-toulouse.fr/7223/1/jacquin.pdf>, 2010.
- [24] R. Olshen et C. Stone L. Breiman, J. Friedman. Classification and regression trees, 1984.
- [25] R. LASSAIGNE. Sécurité différentielle dans les bases de données et complexité. http://iml.univ-mrs.fr/ati/crypto_puces/2013/slide/lassaigne.pdf, 2015.
- [26] Vincent Couallier Léo Gerville-Réache. Échantillon reprÉsentatif (d'une population finie) : DÉfinition statistique et propriÉtÉs. <https://hal.archives-ouvertes.fr/hal-00655566/document>, 2011.
- [27] Xavier Milhau. Segmentation et modélisation des comportements de rachat en assurance vie. <http://www.xaviermilhau.fr/public/MemoireActuariat-XMilhau.pdf>, 2011.
- [28] B. NGUYEN. Techniques d'anonymisation. <https://hal.inria.fr/hal-01113412/document>, 2015.
- [29] P. Phillips P. Schmidt Y. Shin D. Kwiatkowski. Testing the null hypothesis of stationarity against the alternative of a unit root. <http://debis.deu.edu.tr/userweb//onder.hanedar/dosyalar/kpss.pdf>, 1991.
- [30] Jean-Michel Loubes Philippe Besse, Brendan Guillouet. Apprentissage sur donnees massives trois cas d'usage avec r, python et spark. <https://hal.archives-ouvertes.fr/hal-01350099v1/document>, 2016.
- [31] F. Planchet. Modèles financiers et analyses des risques dynamiques en assurance. <http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/0B9DF464E9543283C1256F130067B2F9/FILE/Copules.pdf?OpenElement>, 2010.
- [32] A. POPIER. Copules, support de cours université du maine, le mans. http://perso.univ-lemans.fr/~apopier/enseignement/M2_risque_credit/slides_copule.pdf, 2010.
- [33] Jean-Michel Poggi Robin Genuer. Arbre cart et forêt aléatoire, importance et sélection des variables. <https://hal.archives-ouvertes.fr/hal-01387654v2/document>, 2017.
- [34] T. RONCALLI. Gestion des risques multiples ou copules et aspect multidimensionnels du risque. <http://thierry-roncalli.com/download/Lecture-Notes-Copula-Ensai.pdf>, 2002.
- [35] Frédéric Santos. L'algorithme em : une courte présentation. <http://www.pacea.u-bordeaux1.fr/IMG/pdf/algo-em.pdf>, 2015.
- [36] L. SWEENEY. K-anonymity : A model for protecting privacy. https://epic.org/privacy/reidentification/Sweeney_Article.pdf, 2002.

- [37] Jean VAILLANT. Initiation à la théorie de l'échantillonnage. http://econometrie.ish-lyon.cnrs.fr/IMG/pdf/Initiation_theo_echantillonnage_-_J_Vaillant.pdf, 2005.

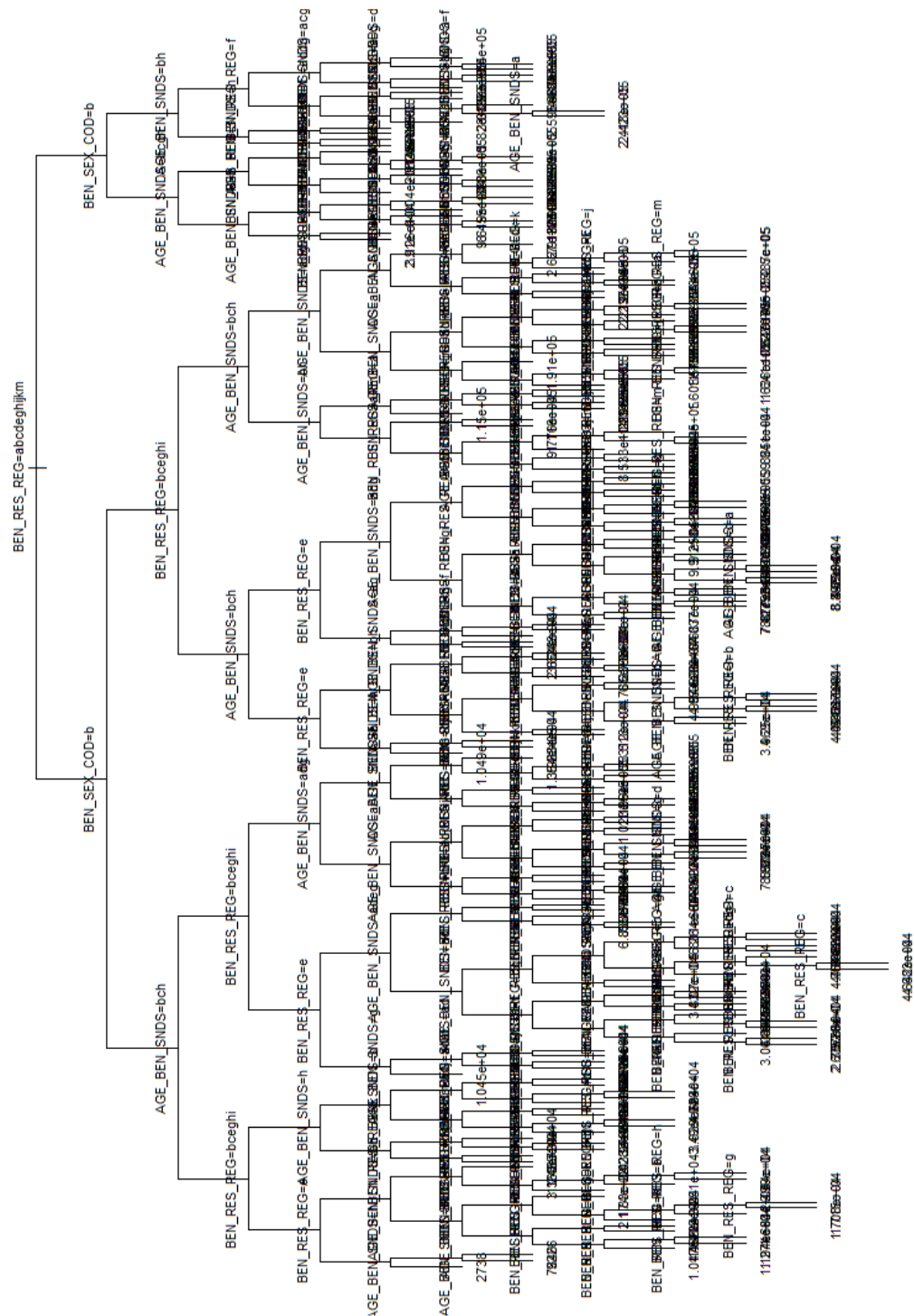
Liste des abréviations

Abréviation	Libellé
CNAMTS	Caisse Nationale d'Assurance Maladie des Travailleurs Salariés
CCMSA	Caisse Centrale de la Mutualité Sociale Agricole
RSI	Régime Social des Indépendants
UNCAM	Union Nationale des Caisses d'Assurance Maladie
PUMa	Protection Universelle Maladie
AT/MP	Accidents du Travail et Maladies Professionnelles
CAT/MP	Commission des accidents du travail et des maladies professionnelles
CRAM	Caisse Régionale d'Assurance Maladie
CGSS	Caisse Générale de Sécurité Sociale
CTR	Comités Techniques Régionaux
CTN	Comités Techniques Nationaux
AMO	Assurance Maladie Obligatoire
CSG	Contribution Sociale Généralisée
OPTAM	Option de Pratique TARifaire Maîtrisée
AMC	Assurance Maladie Complémentaire
ACS	Aide au paiement d'une Complémentaire Santé
RSS	Remboursement de la Sécurité Sociale
BR	Base de Remboursement
TR	Tarif de Responsabilité
TM	Ticket Modérateur
FR	Frais Réels
Dep	Dépassement d'honoraire
BCAB	Bureau Commun des Assurances Collective
TSCA	Taxe Spéciale sur les Conventions d'Assurances
PSAP	Provision pour Sinistres A Payer
PRC	Provision pour Risques Croissants
VAP	Valeur Actuelle Probable
DAMIR	Dépenses d'Assurance Maladie Inter Régimes
SNIIRAM	Système National d'Information Inter-Régime de l'Assurance Maladie
ATIH	Agence Technique d'Information sur l'Hospitalisation
Cnil	Commission nationale de l'information et des libertés

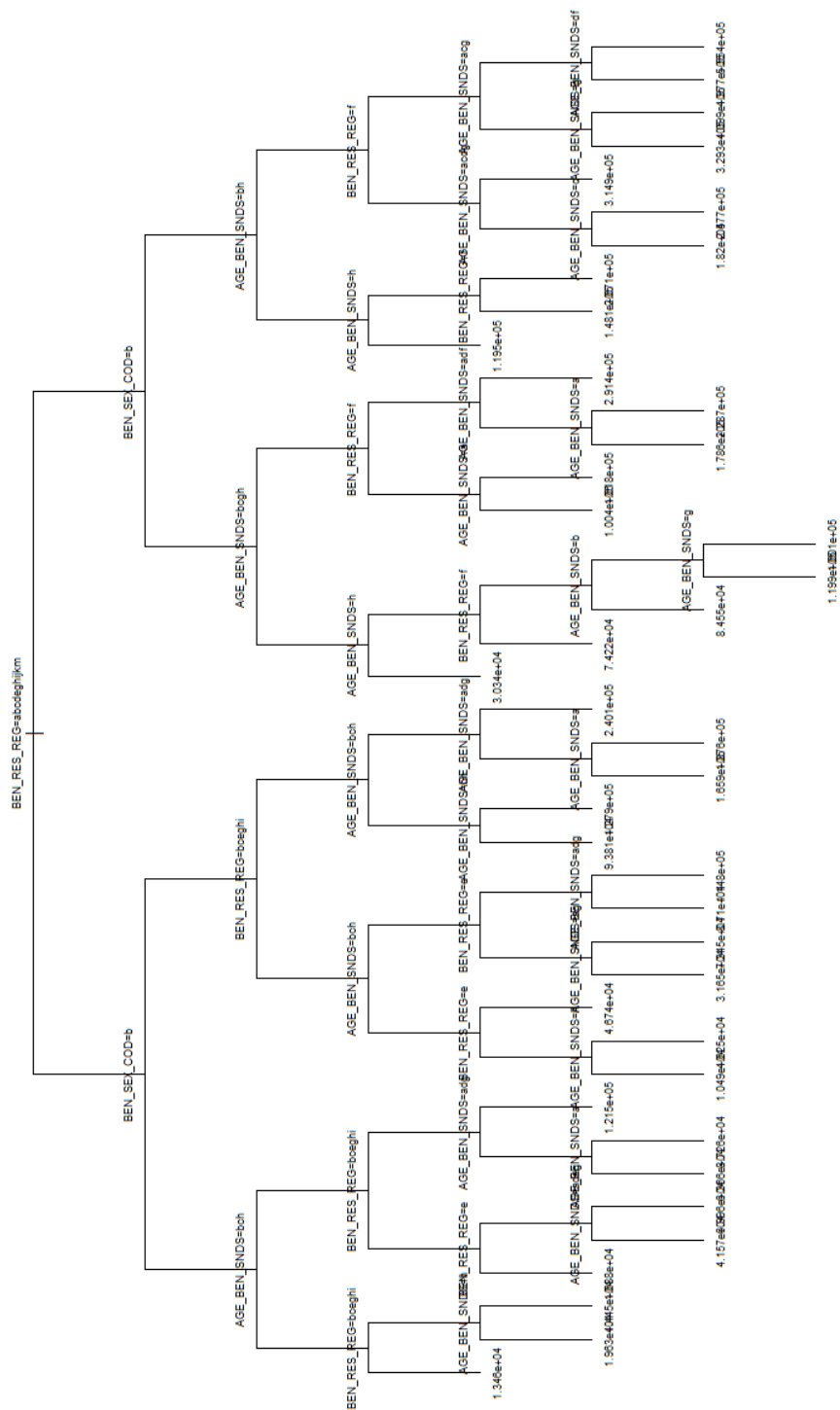
Annexe 1 : Lexique des variables de la base DAMIR

Variable	Libellé	Type	
AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins	qualitative	ordinaire
ASU_NAT	Nature d'Assurance	qualitative	nominale
ATT_NAT	Nature de l'Accident du Travail	qualitative	nominale
BEN_CMU_TOP	Top Bénéficiaire CMU-C	qualitative	nominale
BEN_QLT_COD	Qualité du Bénéficiaire	qualitative	nominale
BEN_RES_REG	Région de Résidence du Bénéficiaire	qualitative	nominale
BEN_SEX_COD	Sexe du Bénéficiaire	qualitative	nominale
CPL_COD	Complément d'Acte	qualitative	nominale
CPT_ENV_TYP	Type d'Enveloppe	qualitative	nominale
DDP_SPE_COD	Discipline de Prestation Etb Exécutant	qualitative	nominale
DRG_AFF_NAT	Nature du Destinataire de Règlement affiné	qualitative	nominale
ETE_CAT_SNDS	Catégorie Etb Exécutant	qualitative	nominale
ETE_IND_TAA	Indicateur TAA Privé/Public	qualitative	nominale
ETE_REG_COD	Région d'Implantation Etb Exécutant	qualitative	nominale
ETE_TYP_SNDS	Type Etb Exécutant	qualitative	nominale
ETP_CAT_SNDS	Catégorie Etb Prescripteur	qualitative	nominale
ETP_REG_COD	Région d'Implantation Etb Prescripteur	qualitative	nominale
EXE_INS_REG	Région du PS Exécutant	qualitative	nominale
EXO_MTF	Motif d'Exonération du Ticket Modérateur	qualitative	nominale
FLT_ACT_COG	Coefficient Global de la Prestation Préfiltré	quantitative	continue
FLT_ACT_NBR	Dénombrement de la Prestation Préfiltré	quantitative	discrète
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée	quantitative	discrète
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltré	quantitative	continue
FLT_PAI_MNT	Montant de la Dépense de la Prestation Préfiltrée	quantitative	continue
FLT_REM_MNT	Montant Versé/Remboursé Préfiltré	quantitative	continue
FLX_ANN_MOI	Année et Mois de Traitement	qualitative	ordinaire
MDT_TYP_COD	Mode de Traitement Etb Exécutant	qualitative	nominale
MFT_COD	Mode de Fixation des Tarifs Etb Exécutant	qualitative	nominale
MTM_NAT	Modulation du Ticket Modérateur	qualitative	nominale
ORG_CLE_REG	Région de l'Organisme de Liquidation	qualitative	nominale
PRE_INS_REG	Région du PS Prescripteur	qualitative	nominale
PRS_ACT_COG	Coefficient Global	quantitative	continue
PRS_ACT_NBR	Dénombrement	quantitative	discrète
PRS_ACT_QTE	Quantité	quantitative	discrète
PRS_DEP_MNT	Montant du Dépassement	quantitative	continue
PRS_FJH_TYP	Type de Prise en Charge Forfait Journalier	qualitative	nominale
PRS_NAT	Nature de Prestation	qualitative	nominale
PRS_PAI_MNT	Montant de la Dépense	quantitative	continue
PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)	qualitative	nominale
PRS_PPU_SEC	Code Secteur Privé/Public	qualitative	nominale
PRS_REM_BSE	Base de Remboursement	quantitative	continue
PRS_REM_MNT	Montant Versé/Remboursé	quantitative	continue
PRS_REM_TAU	Taux de Remboursement	quantitative	continue
PRS_REM_TYP	Type de Remboursement	qualitative	nominale
PSE_ACT_CAT	Catégorie de l'Exécutant	qualitative	nominale
PSE_ACT_SNDS	Nature d'Activité PS Exécutant	qualitative	nominale
PSE_SPE_SNDS	Spécialité Médicale PS Exécutant	qualitative	nominale
PSE_STJ_SNDS	Statut Juridique PS Exécutant	qualitative	nominale
PSP_ACT_CAT	Catégorie du Prescripteur	qualitative	nominale
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur	qualitative	nominale
PSP_SPE_SNDS	Spécialité Médicale PS Prescripteur	qualitative	nominale
PSP_STJ_SNDS	Statut Juridique PS Prescripteur	qualitative	nominale
SOI_ANN	Année de Soins	qualitative	ordinaire
SOI_MOI	Mois de Soins	qualitative	ordinaire
TOP_PSS_TRG	Top Périmètre hors CMU C et prestations	qualitative	nominale

Annexe 2 : Arbre maximal pour la modélisation du reste à charge



Annexe 3 : Arbre élaguée pour la modélisation du reste à charge



Annexe 4 : Imputation par l'algorithme *Expectation Maximization* (EM)

Présentation du modèle

Avec les mêmes notations que précédemment, nous disposons d'un jeu de données $Y = (Y^{obs}, Y^{miss})$ présentant des valeurs manquantes. Si le jeu de données Y était complète, on serait dans le cadre d'estimation classique d'un paramètre θ par maximum de vraisemblance :

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log f(Y, \theta)$$

où $\hat{\theta}$ est la valeur du paramètre θ qui rend $\log f(Y, \theta)$ maximal. En présence de données manquantes, la quantité $l^*(Y, \theta) = \log f(Y, \theta)$ est latente.

L'observation de Y^{obs} donne une information sur la loi de Y . En effet

$$f(Y, \theta) = f(Y^{obs}, Y^{miss}, \theta) = f(Y^{miss} | Y^{obs}, \theta) f(Y^{obs}, \theta)$$

d'où

$$\log f(Y, \theta) = \log f(Y^{miss} | Y^{obs}, \theta) + \log f(Y^{obs}, \theta)$$

En dérivant par rapport à θ , et en intégrant par la suite par rapport à la loi conditionnelle de $Y^{miss} | Y^{obs}$ on a :

$$E\left(\frac{\partial \log f(Y, \theta)}{\partial \theta} | Y^{obs}\right) = E\left(\frac{\partial \log f(Y^{miss} | Y^{obs}, \theta)}{\partial \theta} | Y^{obs}\right) + \frac{\partial \log f(Y^{obs}, \theta)}{\partial \theta}$$

or

$$E\left(\frac{\partial \log f(Y^{miss} | Y^{obs}, \theta)}{\partial \theta} | Y^{obs}\right) = \int \frac{\partial \log f(Y^{miss}=y | Y^{obs}, \theta)}{\partial \theta} f(Y^{miss}=y | Y^{obs}, \theta) dy = \int \frac{\partial f(Y^{miss}=y | Y^{obs}, \theta)}{\partial \theta} dy$$

En permutant l'intégrale et la dérivé, comme l'intégrale de la densité est égale à un, il vient que

$$\int \frac{\partial f(Y^{miss}=y | Y^{obs}, \theta)}{\partial \theta} dy = 0.$$

D'où

$$E\left(\frac{\partial \log f(Y, \theta)}{\partial \theta} | Y^{obs}\right) = \frac{\partial \log f(Y^{obs}, \theta)}{\partial \theta}$$

Le score observables est donc la meilleure estimation du score latent conditionnellement aux observations.

Une étape préalable à l'imputation par l'algorithme EM consiste à déterminer la valeur $\hat{\theta}$ du paramètre θ de sorte que $\frac{\partial \log f(Y^{obs}, \theta)}{\partial \theta} |_{\theta=\hat{\theta}} = 0$.

De ce fait, on a $E\left(\frac{\partial \log f(Y, \hat{\theta})}{\partial \theta} | Y^{obs}\right) = 0$, c'est-à-dire que $E\left(\frac{\partial \log f(Y^{obs}, Y^{miss}, \hat{\theta})}{\partial \theta} | Y^{obs}\right) = 0$. L'idée de l'algorithme EM est de déterminer la valeur du paramètre θ^{miss} telle que

$$E\left(\frac{\partial \log f(Y^{obs}, \theta^{miss}, \hat{\theta})}{\partial \theta} | Y^{obs}\right) = 0$$

Ce qui revient à déterminer θ^{miss} telle que

$$\hat{\theta}^{miss} = \arg \max_{\theta \in \Theta} E(\log f(Y^{obs}, \theta^{miss}, \theta | Y^{obs}))$$

Ici les valeurs manquantes sont vues comme un paramètre à déterminer. Si la distribution conditionnelles $E(\log f(Y^{obs}, \theta^{miss}, \theta | Y^{obs}))$ est assez régulière (ce qui est très rare en pratique) il serait alors judicieux d'utiliser un procédé de maximum de vraisemblance pour déterminer le paramètre θ^{miss} en une itération.

Partant d'une valeur arbitraire θ_0^{miss} l'algorithme itératif à mettre en œuvre est :

— Étape E : Calcule de l'espérance conditionnelle

$$E_{\theta_k^{miss}} = E(\log f(Y^{obs}, \theta_k^{miss}, \theta | Y^{obs}))$$

— Étape M : Calcule du maximum en θ

$$\theta_{k+1}^{miss} = \arg \max_{\theta \in \Theta} E_{\theta_k^{miss}}$$

Point sur la convergence du modèle

La suite (θ_k^{miss}) construite par l'algorithme EM converge idéalement vers $\hat{\theta}$. Théoriquement il n'existe pas (pour le moment) d'argument qui prouve la convergence de cette algorithme vers un maximum global. Toutefois les travaux de Wu[1983] ont prouvé la convergence de ce modèle vers un maximum local de la log-vraisemblance sous l'hypothèse d'une régularité globale de la distribution f .

Le principale reproche de cette algorithme est qu'il est très chronophage et très sensible au point de départ θ_0^{miss} .

Annexe 5 : Imputation par un algorithme de Monte Carlo par Chaîne de Markov (MCMC)

Le principe de base des méthodes de MCMC est l'utilisation d'une chaîne de Markov ergodique de loi stationnaire f . L'idée est de partir d'une valeur $x^{(0)}$ pour générer une chaîne $(x^{(t)})$ markovienne qui converge en loi vers f .

Définition

On appelle méthode de Monte Carlos par chaîne de Markov, tout procédé donnant une chaîne de Markov ergodique dont la loi stationnaire est celle de la distribution d'intérêt.

Une chaîne de Markov est dite ergodique ou irréductible si chaque état est atteignable depuis un autre état. Elle est dite régulière s'il existe une puissance de P^k de sa matrice de transition P dont tous les éléments sont tristement positifs. Une chaîne régulière est donc ergodique. Une chaîne de Markov $(x^{(t)})$ produite par un procédé de MCMC s'utilise de manière analogue à un échantillon indépendant et identiquement distribué suivant une loi f car le théorème d'ergodicité garantit la convergence de la moyenne vers l'espérance.

$$\frac{1}{T} \sum_{t=1}^T h(x^{(t)}) \xrightarrow{T \rightarrow +\infty} E(h(x))$$

Une telle chaîne de Markov peut être produite à partir de l'algorithme de *Hastings-Metropolis*.

Algorithme de *Hastings-Metropolis*

L'algorithme de *Hastings-Metropolis* associée à une fonction objective f a pour but de simuler une chaîne de Markov $(x^{(t)})$ en utilisant une densité conditionnelle $q(y|x)$ facilement simulable, disponible analytiquement et symétrique ($q(y|x) = q(x|y)$). L'algorithme consiste à :

- Générer $y_t \stackrel{L}{\sim} q(y|x^{(t)})$.
- Déterminer la probabilité

$$\rho(x^{(t)}, y_t) = \min\left\{\frac{f(y_t)}{f(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}, 1\right\}.$$

- Déterminer le terme $x^{(t+1)}$ de la chaîne

$$x^{(t+1)} = \begin{cases} y_t & \text{avec proba } \rho(x^{(t)}, y_t) \\ x^{(t)} & \text{avec proba } 1 - \rho(x^{(t)}, y_t) \end{cases}$$

La probabilité $\rho(x^{(t)}, y_t)$ n'est définie que si $f(x^{(t)}) \neq 0$. En initialisant convenablement la chaîne de Markov $(x^{(t)})$ par une valeur appropriée $X^{(0)}$ pour laquelle $f(x^{(0)}) > 0$, on a $f(x^{(t)}) > 0, \forall t \in N$ car les valeurs de y_t telles que $y_t = 0 \Rightarrow \rho(x^{(t)}, y_t) = 0$, elles sont donc

rejetées par l'algorithme.

Afin d'assurer la convergence de la chaîne $(x^{(t)})$ vers la loi limite f , il est nécessaire d'imposer la condition du théorème suivant :

Théorème : Quelque soit la loi conditionnelle q , f est une loi stationnaire de la chaîne $(x^{(t)})$ construite par l'algorithme de *Hastings-Metropolis*.

Preuve : Le noyau de transition associé à l'algorithme de *Hastings-Metropolis* s'écrit

$$K(x, y) = \rho(x, y)q(y|x) + (1 - \rho(x, y))\delta_x(y)$$

où $\delta_x(\cdot)$ est une masse de Dirac en x .

Cette définition du noyau de transition signifie que si $y \neq x$, alors

$$K(x, y) = \rho(x, y)q(y|x) = \frac{f(y) q(x|y)}{f(x) q(y|x)}q(y|x) = \frac{f(y)q(x|y)}{f(x)} = q(y|x)$$

or d'après l'algorithme, $y \stackrel{L}{=} q(y|x)$. Donc si y est différent de x , la passage de la chaîne $(X_{(t)})$ de l'état x à l'état y se réalise. Dans le cas contraire, la chaîne de Markov $(X_{(t)})$ reste en x .

Considérons un quelconque ensemble mesurable A au sens de la tribu des boréliens de R , A est donc Lebesgue-mesurable.

$$\int K(x, A)f(x)dx = \int \int 1_A(y)K(x, y)f(x)dxdy = \int \int 1_A(y)[\rho(x, y)q(y|x) + (1 - \rho(x, y))\delta_x(y)]f(x)dxdy$$

En posant $D = \{(x, y); f(y)q(x|y) \leq f(x)q(y|x)\}$, on a :

$$\begin{aligned} \int K(x, A)f(x)dx &= \int \int_D 1_A(y) \frac{f(y) q(x|y)}{f(x) q(y|x)} q(y|x) f(x) dxdy \\ &\quad + \int \int_{\bar{D}} 1_A(y) q(y|x) f(x) dxdy \\ &\quad + \int \int_D 1_A(x) \left(1 - \frac{f(y) q(x|y)}{f(x) q(y|x)}\right) q(y|x) f(x) dxdy \\ \int K(x, A)f(x)dx &= \int \int_D 1_A(y) f(y) q(x|y) dxdy \\ &\quad + \int \int_{\bar{D}} 1_A(y) q(y|x) f(x) dxdy \\ &\quad + \int \int_D 1_A(x) f(x) q(y|x) dxdy \end{aligned}$$

$$\begin{aligned}
& - \int \int_D 1_A(x) f(y) q(x|y) dx dy \\
& = \int \int 1_A(y) f(y) q(x|y) dx dy \\
& = \int_A f(y) dy
\end{aligned}$$

En démontrant que $\int K(x, A) f(x) dx = \int_A f(y) dy$, nous avons prouvé que f est une loi stationnaire de la chaîne de Markov $(x^{(t)})$ pour toute loi conditionnelle q .

Il existe d'autres variantes à cette, c'est pourquoi il est courant d'entendre parlé *des algorithmes de Hastings-Metropolis*. Les lecteurs intéressés par le sujet peuvent se référer à *METHODS DE MONTE CARLO PAR CHAÎNES DE MARKOV* (Christian Robert [1996]). Tous ces algorithmes proposent des techniques de simulation qui ne nécessite qu'une connaissance limitée de la loi f à simuler. Il sont donc très adaptés pour le traitement des données manquantes. Un cas particulier de l'algorithme générale de *Hastings-Metropolis* que nous avons présenté est celui de *Gibbs*.

Échantillonnage de Gibbs

La méthode d'échantillonnage de *Gibbs* utilise les propriétés de la loi f à simuler à un degré plus avancé que l'algorithme de *Hastings-Metropolis*.

Structure de l'algorithme.

S'il existe un entier n tel que l'état x se décompose en $x = (x_1, \dots, x_n)$ et si les lois conditionnelles correspondantes sont simulable, alors l'algorithme de *Gibbs* associée à cette décomposition à pour transition $x^{(t)}$ à $x^{(t+1)}$:

Il faut simuler

$$\begin{aligned}
& \text{— 1.} \quad x_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \dots, x_n^{(t)}) \\
& \text{— 2.} \quad x_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_n^{(t)}) \\
& \text{— 3.} \quad x_3^{(t+1)} \sim f_3(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \dots, x_n^{(t)}) \\
& \quad \vdots \\
& \text{— } n. \quad x_n^{(t+1)} \sim f_n(x_n | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{n-1}^{(t+1)})
\end{aligned}$$

Il est facile de généraliser cette algorithme car, pour toute fonction de densité $g(y) = g(y_1, \dots, y_n)$, avec $n > 1$ telle que

$$\int g(x, y_1, y_2, \dots, y_n) dy_1, y_2, \dots, y_n = f(x)$$

et telle que les densités conditionnelles $g(y_1 | y_2, \dots, y_n), \dots, g(y_n | y_1, \dots, y_{n-1})$ soient simulables, l'algorithme de *Gibbs* associé à cette décomposition est fourni par la transition $y^{(t)}$ à $y^{(t+1)}$:

Il faut simuler

$$\text{— 1.} \quad y_1^{(t+1)} \sim g_1(y_1 | y_2^{(t)}, \dots, y_n^{(t)})$$

$$\begin{aligned}
- 2. \quad & y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)}, y_3^{(t)}, \dots, y_n^{(t)}) \\
- 3. \quad & y_3^{(t+1)} \sim g_3(y_3|y_1^{(t+1)}, y_2^{(t+1)}, y_4^{(t)}, \dots, y_n^{(t)}) \\
& \vdots \\
- n. \quad & y_n^{(t+1)} \sim g_n(y_n|y_1^{(t+1)}, y_2^{(t+1)}, \dots, y_{n-1}^{(t+1)})
\end{aligned}$$

Lorsque la complétion de f en g n'est pas obligatoire, il reste juste à choisir judicieusement le nombre n de composantes d'un état x de $(x^{(t)})$.

Comme dans le cas générale de l'algorithme de *Hastings-Metropolis*, il est facile de montrer que l'algorithme de *Gibbs* admet g comme loi invariante de la chaîne de Markov $(y^{(t)})$ et donc f comme loi limite de la sous-chaîne $(x^{(t)})$ lorsque $(y^{(t)})$ vérifie le théorème ergodique.

L'échantillonnage de *Gibbs* à comme avantage le faite que l'algorithme auquel il se rapporte correspond à la composition de p algorithmes de *Hastings-Metropolis*, de probabilités d'acceptation uniformément égales à 1. C'est-à-dire toutes les valeurs simulées sont acceptées.

L'une des premières apparitions historique de l'algorithme de *Gibbs* est **la méthode d'augmentation des données** (*Data Augmentation*).

Méthode de Data Augmentation

Cette méthode à été proposée par Tanner & Wong [1987] et mise en œuvre pour la première fois par Geman & Geman [1984].

Afin d'exploiter l'échantillonnage de *Gibbs*, le principe de la *data augmentation* est de considérer les valeurs manquantes comme de nouveaux paramètres à estimer, ce qui permet de mettre en œuvre l'algorithme comme si nous étions sur une jeu de données complet.

Le principe de cette méthode est la suivant : étant donnée une distribution prédictive $P(Y_{miss}|Y_{obs}, \theta)$ [Dempster et al., (1977)]. On simule alternativement les données manquantes Y_{miss} et paramètres θ . À chaque pas de temps, deux étapes sont réalisées :

- 1.) Estimer $Y_{miss}^{(t+1)}$ par la distribution de probabilité $P(Y_{miss}|Y_{obs}, \theta^{(t)})$.

Dans notre cas, nous avons fait l'hypothèse que les données suivent une distribution gaussien multivariée. A partir d'une estimation du vecteur moyenne et de la matrice des variance-covariance, les données manquantes sont ainsi simulées conditionnellement aux valeurs observées.

- 2.) Estimer $\theta^{(t+1)}$ par la distribution à posteriori $P(\theta|Y_{obs}, Y_{miss}^{(t+1)})$.

Cette étape consiste à simuler les vecteurs paramètres en partant des données augmentées. A chaque itération, cette étape simule le vecteur moyenne a posteriori ainsi que la matrice de covariance a posteriori

Une fois cette deuxième étape terminée, l'algorithme passe à l'itération suivante (retour à l'étape 1, on estime de nouvelle valeur manquante à partir des données observées préalablement augmentées).

On obtient ainsi une chaîne de Markov $((Y_{miss}^{(1)}, \theta^{(1)}), (Y_{miss}^{(2)}, \theta^{(2)}), \dots)$ qui converge en probabilité vers $P(Y_{miss}, \theta | Y_{obs})$.