IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Vermylen MATI SONNA TSAKOU et Assia BOUKELOUD HAKKOU

Titre : Modélisation de la dérive en santé par séries temporelles : Application à l'Open DATA.

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : Groupama GAN VIE
Nom : Vincent LAUDOU

Signature et Cachet : **Groupama Gan Vie**
Direction Collectives
Entreprise régie par le Code des Assurances
Société Anonyme au capital de 1 371 100 605 euros
RCS Paris 340 427 616 - APE : 6511Z
Siège Social : 8-10 rue d'Astorg - 75383 Paris Cedex 08

Directeur de mémoire en entreprise :

Nom : Vincent LAUDOU
Signature :

Invité :

Nom : _____

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

La protection sociale est une des préoccupations majeures des Français. Elle est souvent au cœur des débats économiques et politiques générant bons nombres d'évolutions législatives depuis plusieurs années impactant à la fois le régime obligatoire et le régime complémentaire.
Les assurances en charge du marché de la complémentaire santé font face à un risque dynamique et sont confrontées à un contexte tendu.
En effet, la consommation de soins et biens médicaux ne fait que croitre. On observe une dérive de sinistralité conséquence de plusieurs facteurs : une réglementation en mouvement et qui encadre le marché, une concurrence importante qui resserre les tarifs, une inflation croissante, un vieillissement de la population, l'augmentation de la survenance des maladies chroniques et des coûts de soins importants.

Ce mémoire a pour vocation d' identifier une modélisation adaptée et robuste pour quantifier la consommation en santé autrement dit la dérive de sinistralité.

A partir des bases de données mises à disposition par l' assureur et l'Open Data DAMIR, nous allons tester et comparer plusieurs modèles mathématiques permettant de modéliser cette dérive.

Abstract

Social protection is one of the major concerns of the French. It is often at the heart of economic and political debates, generating a good number of legislative changes for several years, impacting both the compulsory scheme and the supplementary scheme.
The insurers in charge of the complementary health market face a dynamic risk and are confronted with a tense context.
Indeed, the consumption of medical care and goods is only increasing. We observe a drift in claims which is the occurrence of several factors : changing regulations which frame the market, significant competition which tightens prices, inflation, an aging population, chronic diseases and high healthcare costs. .

This dissertation aims to identify an appropriate and robust model to quantify health consumption, in other words the drift in claims.

Using insurer databases and Open Data DAMIR, we will create a scope of study and apply mathematical models to define the most relevant to model this drift.

Note de Synthèse

Le risque santé représente un marché dynamique en France. La consommation de soins et biens médicaux occupe une place importante, elle représente 8.6% du produit intérieur brut, soit 208 milliards d'euros en 2019.
Aujourd'hui le régime obligatoire absorbe 80% de ces dépenses, ce qui est souvent source de débats économiques et politiques conduisant à de nouvelles réformes et évolutions règlementaires. La maitrise des coûts est un enjeu majeur et laisse sous-entendre une réduction des prises en charges par le régime obligatoire.
Les 20% de reste à charge après remboursement du régime obligatoire sont répartis entre les ménages et le régime complémentaire. Ce dernier représente un marché privé partagé entre assureurs de la place.
Dans ce contexte incertain en perpétuel mouvance avec un risque probant d'augmentation de maladies chroniques et d'épidémies, face à une population vieillissante et un désengagement de la Sécurité Sociales, les assureurs sont dans une posture délicate avec un certain nombre d'inquiétudes.
D'autant que la branche santé en assurance collective présente des résultats déficitaires qui s'observe à travers et un ratio de sinistres sur primes supérieur à 100%.
Ce ratio, indicateur permettant d'apprécier le compte et d'identifier le cas échéant un besoin de revalorisation qui serait la conséquence d'une dérive de sinistralité.

Définition de la dérive :

La dérive des soins de santé est l'évolution du coût du risque santé. Elle correspond à l'évolution des prestations payées, toutes choses égales par ailleurs.
Autrement dit, sur un groupe fermé ayant les mêmes caractéristiques d'une année sur l'autre (même moyenne d'âge, même répartition hommes/femmes, même région, mêmes garanties), l'évolution, i.e la dérive est calculée comme suit :

$$\text{Dérive de sinistralité N} = \text{Evolution de la consommation N/N-1} = \frac{\text{Coût moyen N}}{\text{Coût moyen N-1}} - 1$$

Où

$$\text{Coût moyen} = \frac{\text{Consommation}}{\text{nombre de bénéficiaires}}$$

Cette dérive est la survenance de différents facteurs :

- vieillissement de la population,
- développement de maladies chroniques,
- coûts de soins importants,
- l'inflation,

— réformes.

Ce mémoire a pour objectif de trouver une modélisation adéquate afin de mesurer la dérive de soins en santé.

Pour cela, nous disposons de bases de données internes sur 2 exercices de survenance. La profondeur de ces données étant assez faible, elles sont néanmoins très complètes et permettent d'avoir une bonne vision du portefeuille que couvre l'assureur.
C'est donc à partir des statistiques descriptives de ces bases que nous allons définir les caractéristiques de notre portefeuille assureur.

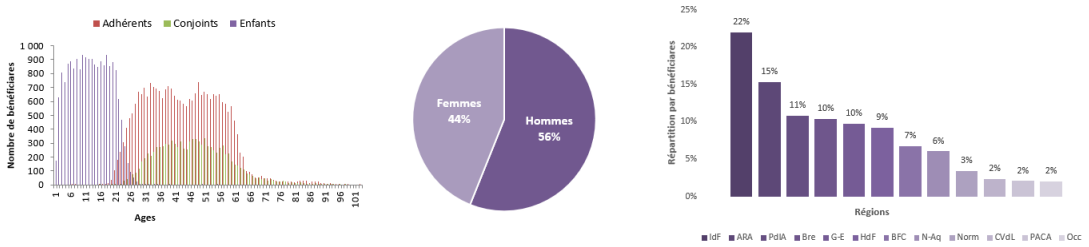

FIGURE 1 – Caractéristiques du portefeuille interne. De gauche à droite : Pyramide des âges, répartition des bénéfiaires par hommes/femmes, par régions

Par la suite et pour créer notre périmètre d'étude, nous allons faire appel à l'Open Data DAMIR. Celle-ci renferme une mine d'informations précieuses et présente une réelle opportunité pour les professionnels de santé notamment les assureurs. En outre, l'un de ses avantages c'est son historique de prestations.

Ainsi, notre périmètre d'étude sera défini sur les prestations de la pharmacie des bases DAMIR de 2014 à 2019. Ce périmètre possèdera les mêmes caractéristiques que notre portefeuille interne assureur.

Les modèles de série temporelles

Notre base de travail sera pour cette partie, une base de données avec comme variables : l'année, le mois et le reste à charge.
Il s'agit d'une base de 72 lignes les montants de reste à charge ont été agrégés par mois entre 2014 et 2019.

Des modèles seront entraînés sur les bases de 2014 à 2018, puis ils seront testés et validés sur les bases de 2019.
La série temporelle de notre base de données se présente comme ci-dessous :

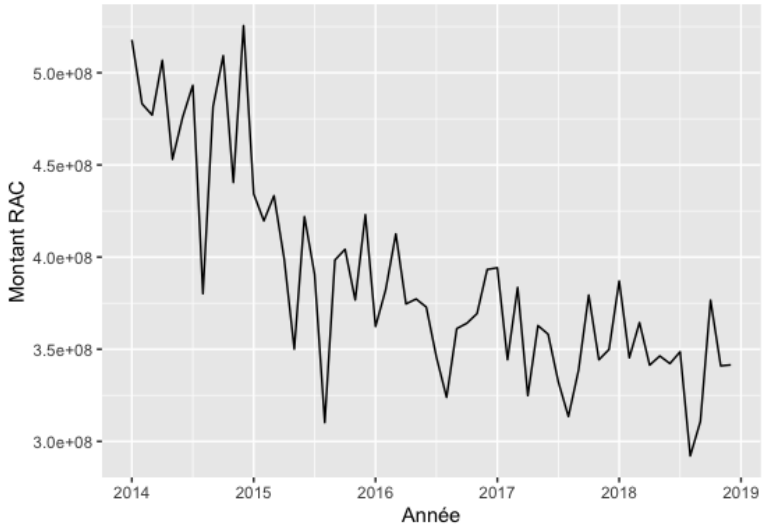

FIGURE 2 – Chronogramme du reste à charge

La phase de modélisation fera intervenir différents modèles. Les candidats sont :
- Les modèles de décomposition simple,
- Les modèles de lissage exponentiel,
- Les modèles de Box-Jenkins,
- Le modèle de régression linéaire.

La modélisation par décomposition simple :

Le postulat de départ est de dire que sachant que sur la série temporelle nous observons une tendance et des éléments faisant penser à de la saisonnalité, si on la décompose en une expression de la tendance et de saisonnalité on aura un reste qui devrait correspondre au bruit blanc.
Le résidu étant la partie aléatoire et négligeable du modèle, on ne la modélise pas.

L'exercice de projection de la série temporelle se résume donc à projeter un modèle à tendance et saisonnalité pour estimer les prédictions futures des restes à charge.
La première étape de cette décomposition simple de la série temporelle a été d'identifier le type de structure de modèle de séries temporelles (additif, multiplicatif ou mixte) afin de modéliser la nature du lien entre la tendance, la saisonnalité et le résidu.
Nous avons utilisé la méthode du profil, la méthode analytique du tableau de Buys-Ballot et la méthode de la bande.
Nous sommes arrivés à la conclusion que nous étions dans le cadre d'un modèle multiplicatif.

Afin de stabiliser la variance de nos données, nous avons utilisé une transformation de Box-Cox avec un choix de λ optimal égal à -0.6548732. Ce qui a eu comme avantage de rendre le modèle additif, de réduire la variance de la série et de se rapprocher de la normalité des résidus.
Nous avons essayé une première décomposition manuelle mais les résidus présentaient de la saisonnalité résiduelle.
Dans un second temps nous avons utilisé des méthodes de décomposition automatiques proposées par le logiciel R notamment la décomposition à l'aide de la fonction *decompose()* et *STL()*.
Le modèle obtenu par la fonction *décompose()* a permis d'obtenir des résidus non corrélés et normaux.

Ce qui n'est pas le cas du modèle *STL()*.
La prédiction sur les données de 2019 avec les modèles *STL()* et *decompose()* ont confirmé que le meilleur modèle était celui réalisé à l'aide de la fonction décompose.

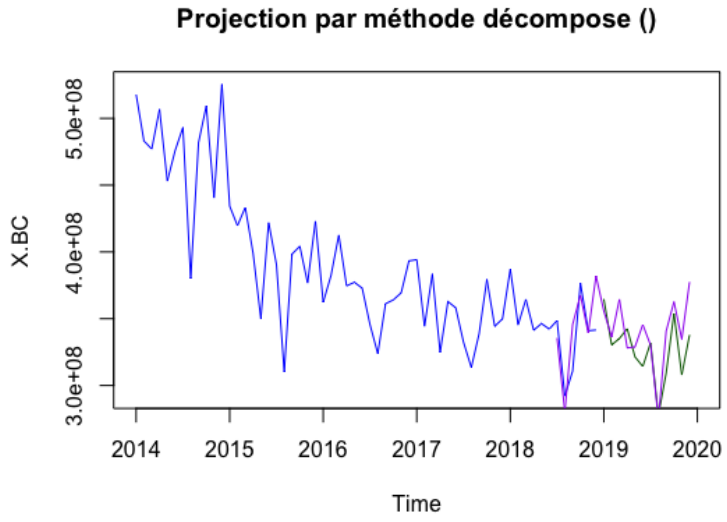

FIGURE 3 – Comparaison des indicateurs de performances entre les méthodes STL() et decompose()

La modélisation par lissage exponentiel

Nous avons testé plusieurs méthodes de modélisation de séries temporelles en fonction des différentes composantes de type de tendance, type de saisonnalité et type d'erreur.

D'un côté sur les bases de données brutes et de l'autre, des données transformées par Box-Cox.
Nous sommes arrivés à la conclusion que les modèles de type Holt-Winters étaient les meilleurs modèles adaptés à la présence de saisonnalité et de tendance.
Sur les données brutes, le meilleur modèle de Holt-Winters obtenu est de type multiplicatif.
Pour les modèles transformés par Box-Cox, c'est le modèle additif avec tendance additive qui a été retenu.
Entre ces 2 modèles celui qui s'est ajusté le mieux à nos données de test est le modèle additif sur des données transformées par Box-Cox.

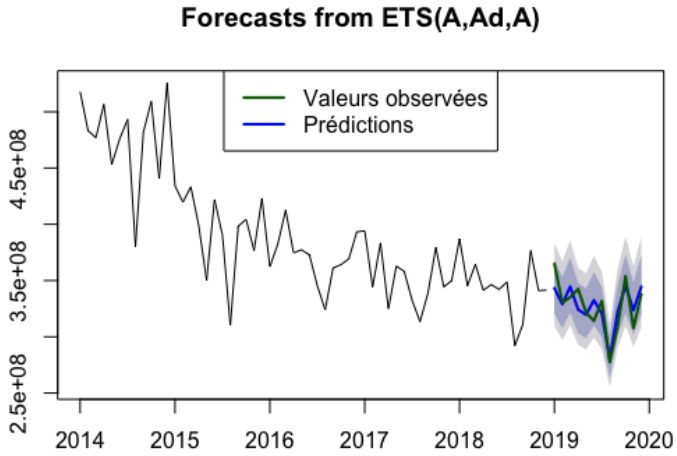

FIGURE 4 – Ajustement du modèle de lissage exponentiel de type (A,Ad,A) sur l'année 2019

La modélisation par Box-Jenkins

Après avoir étudié la stationnarité et rendu la série stationnaire nous l'avons transformée grâce au logarithme afin de tester une autre méthode de stabilisation de la variance.

La série stationnaire a été obtenue par différentiation première afin d'éliminer la tendance et ensuite par différenciation saisonnière d'ordre 12.

Nous avons effectué des modélisations pas à pas à partir d'un modèle SARIMA(1,1,1)(1,1,1)[12] car les autocorrélogrammes empiriques mettaient en évidence la présence d'une structure autorégressive.

Parmi les modèles candidats qui ont été testés, le meilleur modèle d'un point de vue significativité des paramètres et optimisation des critères d'AICc est le modèle SARIMA(0,1,1)(0,1,0)[12]. Ce modèle présente des caractéristiques attendues à savoir des résidus qui sont non corrélées et normaux.
Cependant l'exercice de prévision sur les modèles SARIMA testés ont mis en évidence le fait que c'est le modèle SARIMA(1,1,1)(1,1,1)[12] avec des coefficients non significatifs mais avec le plus faible AICc qui a eu le meilleur score de prédiction.

Les modèles de séries temporelles ont permis avec peu de données (uniquement la chronique des restes à charge) d'estimer et de prévoir le reste à charge future. Cependant les modèles testés ne permettent pas de prendre en compte les variables exogènes et ont donc un faible pouvoir explicatif.

La modélisation par GLM

Nous avons décidé de tester un modèle GLM avec une fonction de lien Gamma.
Le résultat de cette modélisation est :
$Der2019_{estime} = -0.015349 \times Masculin - 0.043572 \times Ag20-59 - 0.060416 \times Ag60-79 - 0.068407 \times Ag80 + 0.331791 \times Der2017 + 0.675721 \times Der2018$.
Ce qui nous a donné comme résultat une équation de la dérive avec des coefficients qui sont tous significatifs.

Les différents modèles retenus, ont eu de bonnes qualités prédictives : le taux de reste à charge "RA-Créel" de 2019 a toujours été dans l'intervalle de confiance à 95% donné par ces modèles.

En conclusion :
L'exploration de ces différentes méthodes d'estimation montre qu'il existe plusieurs outils différents à disposition de l'assureur afin d'estimer la dérive.

L'utilisation de l'Open Data DAMIR, nous a permis d'avoir une profondeur d'historique plus importante afin d'entrainer nos modèles.

Les modèles de séries temporelles et GLM modélisent uniquement un lien linéaire entre les différentes covariables, il pourrait être intéressant d'explorer la piste des modèles de type réseaux de neurones traitant de série temporelles, qui ont de bonnes performances d'après la littérature.

Synthesis Note

Health risk represents a dynamic market in France. With consumption of medical care and goods occupying an important place, it represents 8.6% of the gross domestic product, or 208 billion euros in 2019.
Today, the mandatory scheme absorbs 80% of this expenditure, which is often a source of economic and political debate leading to new reforms and regulatory developments. Controlling costs is a major issue and implies a reduction in coverage by the compulsory scheme.
The remaining 20% payable after reimbursement from the compulsory scheme is divided between the households and the supplementary scheme. The latter represents a private market shared between local insurers.
In this uncertain context in perpetual motion with a convincing risk of an increase in chronic diseases and epidemics, faced with an aging population and a withdrawal from Social Security, insurers are in a delicate position with a certain number of concerns.
Especially since the health branch in collective insurance presents loss-making results which can be seen through and a ratio of claims to premiums greater than 100% ;
This ratio, an indicator for assessing the account and identifying, if necessary, a need for revaluation which would be the consequence of a drift in claims.

Definition of Drift :

Health care drift is the evolution of the cost of health risk. It corresponds to the change in benefits paid, all other things being equal.
In other words, on a closed group with the same characteristics from one year to the next (same average age, same male/female distribution, same region, same guarantees), the change, i.e. the drift is calculated as follows :

$$\text{Loss rate drift N} = Evolution N/N - 1 = \frac{\text{Average cost N}}{\text{Average cost N-1}} - 1$$

Où

$$\text{Average cost} = \frac{\text{Consumption}}{\text{number of beneficiaries}}$$

This drift is the occurrence of different factors :

— Aging of the population,
— development of chronic diseases,
— significant care costs,
— inflation,
— reform.

This thesis aims to find an adequate model to measure the drift of health care.

We have at our disposal internal databases on 2 occurrence exercises. The depth of these data being quite low, they are nevertheless very comprehensive and provide a good view of the portfolio covered by the insurer.
It is therefore from the descriptive statistics of these databases that we are going to define the characteristics of our insurer portfolio.

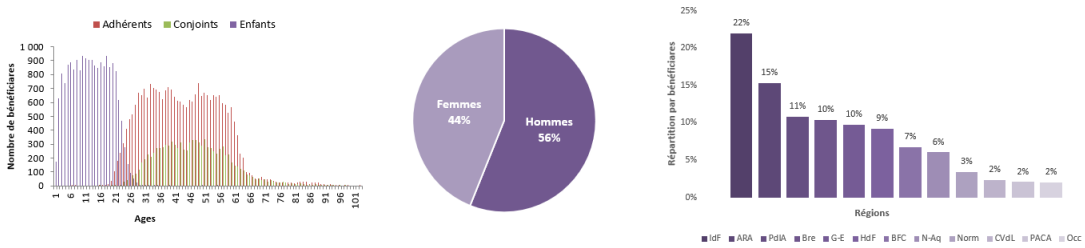

FIGURE 5 – Features of the internal wallet. From left to right : Age structure, distribution of beneficiaries by sex, by region

Subsequently and to create our scope of study, we will use the Open data DAMIR. It contains a wealth of valuable information and presents a real opportunity for health professionals, especially insurers. In addition, one of its advantages is its performance history.

Thus, our scope of study will be defined on the pharmacy services of the DAMIR bases from 2014 to 2019. This scope will have the same characteristics as our internal insurer portfolio.
To do this, we will test and then compare different models.
Thus, our scope of study will be defined on the pharmacy services of the DAMIR bases from 2014 to 2019. This scope will have the same characteristics as our internal insurer portfolio.

Time series models

Our working base will be for this part is a database with as variables : the year, the month and the rest to be paid.
This is a base of 72 lines the amounts of remaining dependents have been aggregated by month between 2014 and 2019.

Models will be trained on the bases from 2014 to 2018, then they will be tested and validated on the bases of 2019.
The time series of our database is presented as below :

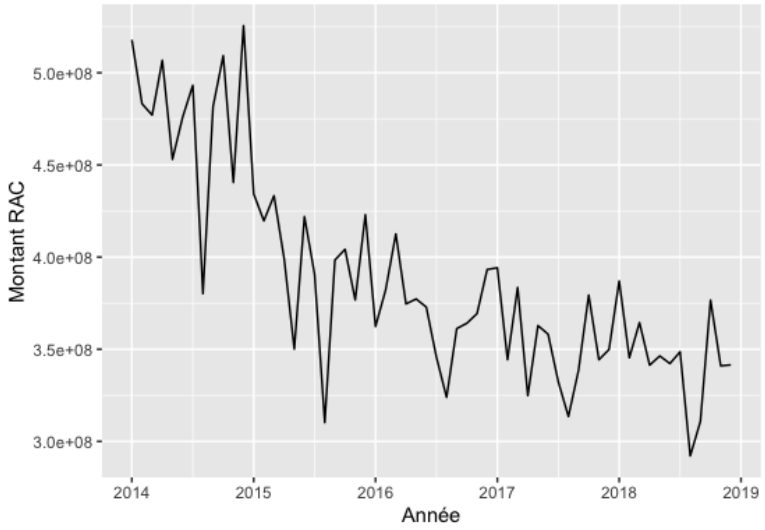

FIGURE 6 – Chronogram of the remaining charge

The modeling phase will involve different models. The candidates are :

- Simple decomposition models
- Exponential smoothing models
- Box-Jenkins Models
- The linear regression model

The simple decomposition method :

The starting postulate is to say that knowing that on the time series we observe a trend and elements reminiscent of seasonality, if we break it down into an expression of the trend and seasonality we will have a remainder which should correspond to the noise White.
The residues being the random and negligible part of the model, one does not model it.

The exercise of projecting the time series therefore boils down to projecting a trend and seasonality model to estimate future predictions of drift.
The first step in this simple time series decomposition was to identify the type of time series model in order to model the nature of the link between the trend and seasonality St and the epsilon residual.
We used the profile method, the analytical method of the Buys-Ballot table and the band method.
We came to the conclusion that we were in the context of a multiplicative model.

In order to stabilize it we used a Box-Cox transformation with a choice of λ = -0.6548732. This had the advantage of making the model additive, reducing the variance of the series and getting closer to the normality of the residuals.
We tried a first manual decomposition but the residuals showed residual seasonality.
Secondly, we used automatic decomposition methods offered by the R software, in particular decomposition with the help of the function *decompose()* and *STL()*.
The model obtained by the function *Decompose()* allowed to obtain uncorrelated and normal residuals.
Which is not the step of the STL model.
The prediction on the 2019 data with the *STL* and *decompose* models confirmed that the best model was the one made using the decompose function.

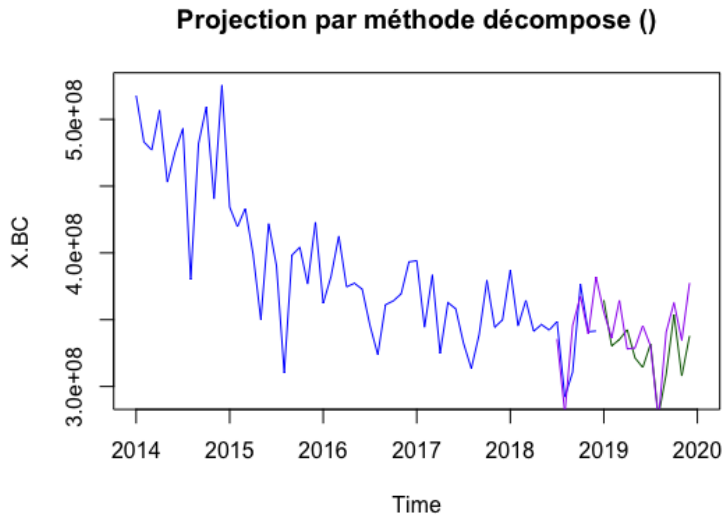

FIGURE 7 – Comparison of performance indicators between *STL* and *decompose* methods

Exponential smoothing

We tested several time series modeling methods according to the different components of type of trend, type of seasonality and type of error.

On one side the raw databases and on the other data transformed by Box-Cox.
We came to the conclusion that the Holt-Winters type models were the best models adapted to the presence of seasonality and trend.
On the raw data, the best Holt-Winters model obtained is of the multiplicative type.
For the models transformed by Box-Cox, it is the additive model with additive tendency additives and additives which was retained.
Between these 2 models, the one that fitted best to our test data is the additive model on transformed data.

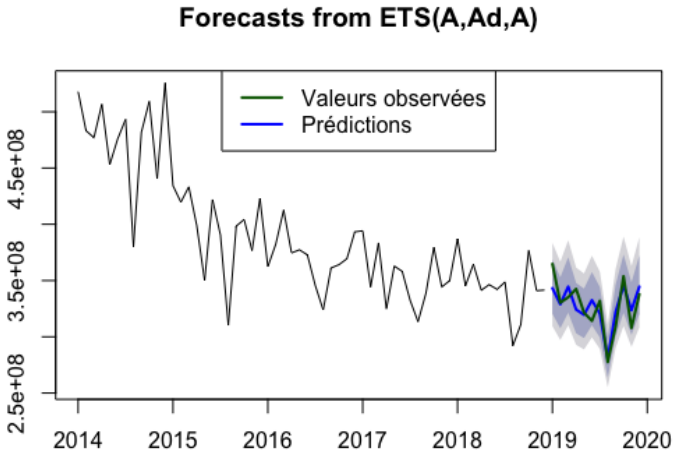

FIGURE 8 – Adjustment of the exponential smoothing model of type (A,Ad,A) on the year 2019

Modeling by Box-Jenkins

After studying stationarity and making the series stationary, we transformed it using the logarithm in order to test another method of stabilizing the variance.

The stationary series was obtained by first differentiation in order to eliminate the trend and then by seasonal differentiation of order 12.

We performed step-by-step modeling from a SARIMA(1,1,1)(1,1,1)[12] model because the empirical autocorrelograms highlighted the presence of an autoregressive structure.

Of the candidate models that have been tested, the best model from a parameter significance and AICc criteria optimization point of view is the SARIMA(0,1,1)(0,1,0)[12] model presents the expected characteristics, namely residuals that are uncorrelated and normal.

However, the forecasting exercise on the SARIMA models tested highlighted the fact that it is the SARIMA(1,1,1)(1,1,1)[12] model with insignificant coefficients but the lowest AICc which had the best prediction score.

The time series models allowed with little data only the chronicle of the remainders to estimate and to forecast the future remainder to the burden. However, the models tested do not take into account exogenous variables.

Modeling by MLG

We decided to test a MLG model with a gamma link function.
The result of this modeling is :
$\text{Last2019}_{estimated} = -0.015349 \times Masculin - 0.043572 \times Ag20-59 - 0.060416 \times Ag60-79 - 0.068407 \times Ag80 + 0.331791 \times Der2017 + 0.675721 \times Der2018.$
Gave us as a result a drift equation with coefficients that are all significant.
The different models retained had good predictive qualities : the 2019 Real drift rate was always within

the 95% confidence interval given by these models.

In conclusion :
Exploring these different estimation methods, there are several different tools available to the insurer to estimate the drift.

The use of open data allowed us to have a greater depth of history in order to train our models.

Time series and GLM models only model a linear link between the different covariates, it could be interesting to explore the path of neural network type models dealing with time series, which have good performance according to the literature.

Remerciements

Nous souhaiterions remercier toutes les personnes qui ont contribué à l'élaboration de ce mémoire, par leur soutien leur aide ou leur présence.
Nous remercions plus particulièrement Vincent LAUDOU, Manuel PARMENTIER et Guillaume PLEYNET-JESUS.

Table des matières

Table des matières

Introduction

La France est l'un des pays qui consacre le plus de richesse à la dépense courante de santé : si l'on se compare avec les autres pays développés, seuls les États-Unis, l'Allemagne et la Suisse dépensent davantage. La santé est, avec les retraites, le poste de dépense le plus important de la protection sociale en France. Cette dernière se plaçant en second rang dans les préoccupations des Français.
C'est d'ailleurs l'une des raisons qui pousse les politiques à de nombreuses évolutions législatives dont l'objectif est de trouver le bon équilibre entre protection des Français et maîtrise des coûts.
L'une des réformes a consisté au renforcement du système en 2011 par l'obligation à chaque employeur de proposer une assurance complémentaire à leurs salariés. C'est ainsi que le marché de la santé des collectifs a connu une dynamique accélérée et s'est vu gagner des parts de marché au détriment du marché de l'individuel. Mais ce gain en part de marché ne va pas toujours de pair avec le gain de technique. Le déficit s'inscrit dans le temps, les dépenses qui ne cessent d'augmenter du fait de nombreuses réformes, de l'élévation générale du niveau de vie, du développement de maladies chroniques, de l'accroissement démographique et du vieillissement de la population. A cela s'ajoute une concurrence accrue qui casse les prix et une réglementation qui cadre le marché.
Outre la tarification, acteur direct de la rentabilité, la maitrise du portefeuille d'assurés et leurs comportements sont un atout majeur pour le pilotage et les prévisions, un levier de performance intéressant et recherché par chaque assureur.
L'enjeu d'une surveillance est d'identifier la dérive de sinistralité de son portefeuille.
L'objectif de ce mémoire sera l'étude de cette dérive de sinistralité.
Dans le premier chapitre, nous ferons un tour d'horizon sur la santé telle qu'elle existe aujourd'hui avec son mécanisme, la réglementation en vigueur, les acteurs et leurs poids sur le marché. Nous présenterons le contexte de l'étude et les données dont nous disposons.
Le chapitre suivant sera consacré à la théorie des modèles candidats pour la modélisation de la consommation.
Enfin, le dernier chapitre sera la mise en œuvre de ces modèles et l'identification du plus performant.

Chapitre 1

Le domaine de la santé en France

Ce chapitre s'organise autour de trois parties.
La première partie présente le fonctionnement de l'assurance santé. Nous y exposerons les mécanismes de prise en charge, les acteurs du marché ainsi que les dernières évolutions législatives et réglementaires.
En seconde partie dédiée à la présentation de l'étude, c'est à l'aide des statistiques publiées par la Direction de la recherche, des études de l'évaluation et des statistiques DRESS, que nous présenterons d'un point de vue macro, la place de la complémentaire collective sur le marché de la santé en France. La tarification et les facteurs discriminants de la branche y seront abordés.
La partie qui suivra sera consacrée à la nécessité pour un assureur de surveiller et piloter son portefeuille dans le temps et des méthodes dont il dispose.
Enfin, la dernière partie sera consacrée à la présentation des données pour réaliser notre étude.

1.1 Éléments contextuels

Introduction

Cette première partie vise à présenter le système de santé en France.
Les acteurs du marché, le principe de remboursement ainsi que les dernières évolutions législatives qui l'ont impacté.

La couverture du risque santé se décompose en deux parties principales :
- le régime de base obligatoire, c'est l'Assurance Maladie,
- le régime complémentaire.

1.1.1 Le régime obligatoire

Ce n'est qu'au lendemain de la deuxième guerre mondiale, en 1945, que la sécurité sociale verra le jour avec comme but de protéger socialement l'ensemble des Français.
Elle est organisée autour de cinq branches :
- la branche maladie
- la branche accidents du travail
- la branche retraites
- la branche famille
- la branche recouvrements

C'est dans la branche maladie que l'on retrouve l'assurance maladie obligatoire qui intervient notamment dans le remboursement de tout ou partie de frais médicaux.
Cette prise en charge du régime obligatoire (RO) est définie sur une base de tarif sécurité sociale, fixée par convention ou d'autorité.
Le financement se fait au travers des cotisations des employeurs et des salariés mais aussi au travers d'autres sources comme la CSG, la CRDS, les taxes sur le tabac et l'alcool...

1.1.2 Le régime complémentaire

Rôle :
La prise en charge des frais de santé par le RO n'est que partielle. Il demeure un reste un reste à charge pour l'assuré. C'est dans ce contexte que l'assurance complémentaire entre en jeu. Elle prend en charge une part des dépenses de santé laissées à la charge des assurés. Elle permet donc de compléter les remboursements de la sécurité sociale mais également de prendre en charge certaines dépenses de santé non couvertes par la sécurité sociale (exemple : les frais de chambre particulière).

Acteurs :
Le marché de l'assurance complémentaire se répartis en trois catégories :
- Les sociétés d'assurances, régies par le code des assurances et représentées par la Fédération Française de l'Assurance (FFA)
- Les mutuelles, régies par le code de la mutualité et représentées notamment par la Fédération Nationale de la Mutualité Française (FNMF)
- Les institutions de prévoyance, régies par le code de la sécurité sociale et représentées par le Centre Technique des Institutions de Prévoyance (CTIP).

Type de contrats :
Il existe deux types de contrats : les contrats individuels et les contrats collectifs.
Les contrats individuels sont souscrits par des particuliers, tandis que les contrats collectifs sont souscrits dans la plupart des cas par des employeurs, au profit d'un ou plusieurs salariés. Concernant les contrats collectifs, plusieurs modes d'adhésion sont possibles :
- L'adhésion obligatoire : l'ensemble des salariés définis de manière objective est obligatoirement affilié.
- L'adhésion facultative : le contrat s'applique aux seuls salariés désirant bénéficier des garanties proposées.

D'après une étude du marché réalisée par la DRESS 2020, et selon le graphique ci dessous, il en ressort que les institutions de prévoyance sont spécialisées sur les contrats santé collectifs, lesquels représentent 87 % des cotisations collectées en 2019. Les mutuelles sont quant à elles largement positionnées sur les contrats santé individuels (69 % de leur activité). Les sociétés d'assurances sont dans une position intermédiaire avec 54 % des cotisations collectées au titre de contrats collectifs.

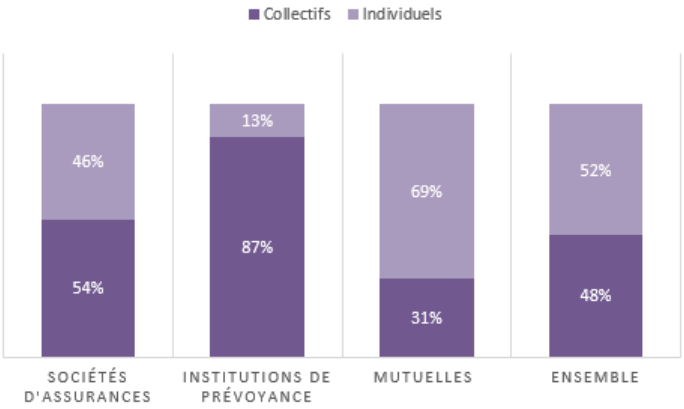

FIGURE 1.1 – Part des contrats collectifs et individuels dans l'ensemble des cotisations collectées en santé par les différents types d'organismes (en % des cotisations collectées 2019).

1.1.3 Les grands postes de soins

Les différents sinistres que le contrat d'assurance complémentaire est amené à rembourser peuvent être classés suivant leur nature. Cela nous permet d'isoler six postes majeurs de dépenses correspondant à des états ou pathologies des assurés.

L'hospitalisation :
Ce poste regroupe les frais de séjour lors d'une hospitalisation, les honoraires de chirurgie, anesthésie, le forfait journalier, et des prestations dites de "confort" telles que la chambre particulière ou les frais d'accompagnant. C'est le poste de dépenses dont le coût est le plus important : il s'agit de prestations d'un montant élevé. Cependant, les taux de remboursement du régime obligatoire sont également élevés.

Les soins courants :
Il s'agit essentiellement des dépenses de consultations de médecins généralistes et spécialistes, des frais de laboratoire, d'analyses, d'imagerie et des actes techniques médicaux. Ce poste ne présente pas forcément des frais élevés mais la fréquence des prestations est importante.

La pharmacie :
Il s'agit des prescriptions de médicaments (remboursés par le régime obligatoire). Pour ce poste aussi, ces prestations ont une fréquence importante, avec, pour certains traitements, un coût élevé.

L'optique :
Ce poste comporte les garanties : lunettes (verres et monture), lentilles et opérations de la myopie. Les remboursements du régime obligatoire sont très faibles et la dépense peut s'apparenter à une "dépense esthétique", les lunettes étant considérées comme un objet de mode (cas d'un renouvellement de lunettes sans changement de vue ou d'achat de lentilles non remboursées par le régime obligatoire). La part du régime complémentaire est donc importante. L'aléa combiné à l'utilité de la prestation pour l'assuré fait que ce poste est sujet à des dérives de "consommation".

Le dentaire :
Ce poste regroupe les soins dentaires, les prothèses dentaires, et l'orthodontie. Les prothèses et l'orthodontie étant peu remboursés par rapport au coût réel, il y a une nécessité de couverture complémentaire et la garantie peut être importante.

Les autres dépenses :
Nous regroupons ici toutes les prestations garanties qui n'appartiennent pas aux postes détaillés ci-dessus. Il s'agit notamment du remboursement des cures thermales prescrites et remboursées par le régime obligatoire, le remboursement des séances de médecines douces (par exemple l'acupuncture ou la chiropractie).

1.1.4 Principe de remboursement : décomposition d'un acte médical

Le remboursement de frais de soins de santé est basé sur un tarif de référence, différent pour chaque acte médical, et sur un taux de remboursement. C'est ainsi que nous obtenons le montant de remboursement de la Sécurité sociale. C'est le 1er palier de la pyramide de remboursement. Viens ensuite en relais du régime obligatoire, la prise en charge de la complémentaire santé définie en fonction des garanties souscrites. Et enfin, un reste à charge pour l'assuré. Pour mieux comprendre le mécanisme voici une illustration de prise en charge d'un sinistre :

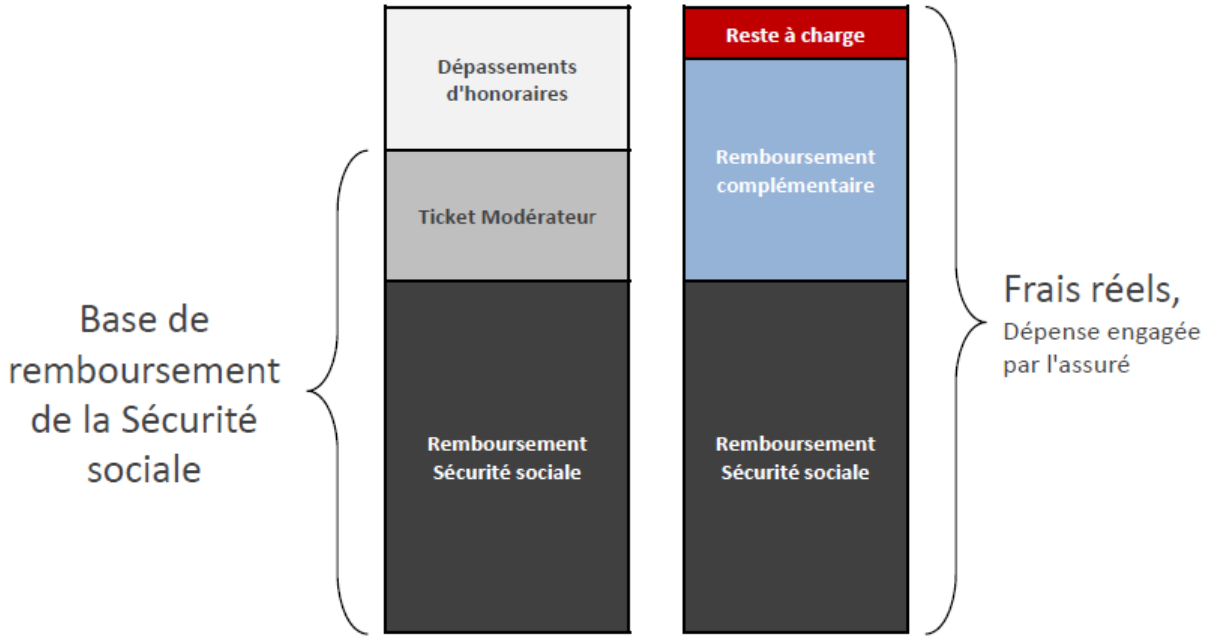

FIGURE 1.2 – Décomposition d'un sinistre.

— Les Frais Réels (FR) : cette variable désigne le montant global dépensé par un individu pour un acte médical déterminé ;

— La base de Remboursement de la Sécurité Sociale (BRSS) : pour un acte médical, correspond à un montant référence, exprimé en euros, remboursé totalement ou partiellement par la Sécurité Sociale ;

— La participation forfaitaire de 1€ est due par les assurés depuis le 1er janvier 2005. Elle est demandée à toutes les personnes âgées de plus de 18 ans, elle s'applique pour toutes les consultations et actes réalisés par un médecin, mais également sur les examens radiologiques ou analyses de biologie médicale. Elle est limitée à 4€ par jour pour un même professionnel de santé ;

— Le montant effectivement remboursé ou Remboursement Sécurité Sociale (RSS) étant déterminé par un taux de remboursement appliqué sur la BRSS ;

— Le ticket Modérateur (TM) : la différence entre la BRSS et le RSS, i.e. la part du montant BRSS non remboursé par la Sécurité Sociale ;

— Les dépassements d'honoraires qui ne sont pas pris en charge par l'Assurance Maladie ;

— Le montant remboursé par l'organisme complémentaire (RC) : il dépend des niveaux des garanties souscrites par l'assuré et peut notamment comprendre la prise en charge du ticket modérateur ;

— Le reste à Charge (RAC) : c'est le montant restant à régler par l'assuré pour rembourser les frais réels de ses soins, après remboursement de la Sécurité Sociale (RO) et de son organisme complémentaire (RC).

Comme nous l'indique les graphiques ci-dessous (source DRESS 2020), l'assurance maladie obligatoire couvre en grande partie les dépenses en santé, laissant à la charge 13% de ces frais aux complémentaires.

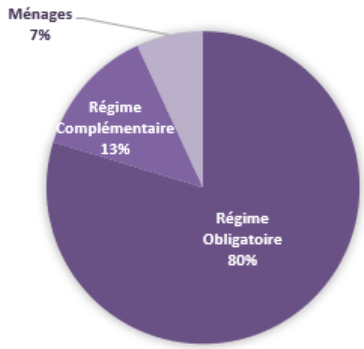

FIGURE 1.3 – Les dépenses de santé 2019 selon les différents acteurs

Si cette répartition entre acteurs est assez stable depuis plusieurs années il n'en demeure pas moins que cette prise en charge varie fortement d'un poste à l'autre. Les niveaux de recouvrement de la Sécurité sociale présentent des disparités importantes selon le poste considéré. Le graphique ci-dessous illustre bien la situation par grands postes de soins.

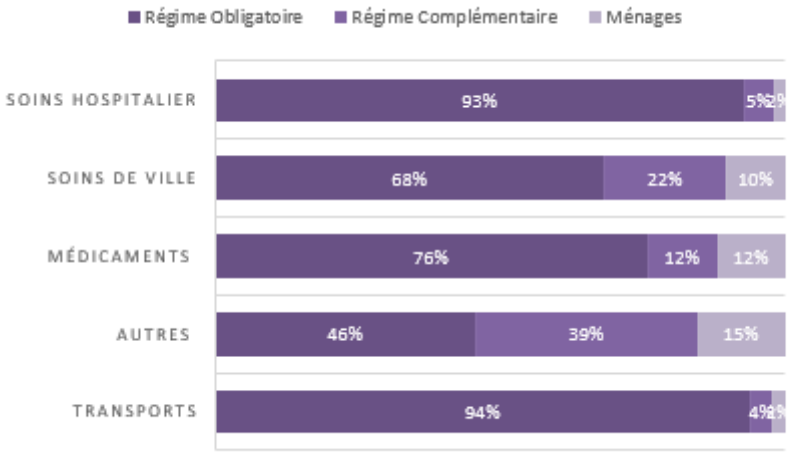

FIGURE 1.4 – Prise en charge des dépenses de santé 2019 selon les différents postes

La consommation de soins hospitaliers est majoritairement absorbée par la Sécurité sociale à hauteur de 93% tandis que les organismes complémentaires en supportent 5%.
A contrario, l'optique que l'on retrouve dans le poste « Autres » est très faiblement pris en charge par le régime obligatoire. Les frais engagés sur ce poste sont pris en grande partie par les complémentaires ainsi que les ménages.

1.1.5 Les évolutions réglementaires

Depuis 20 ans, la réglementation liée aux contrats d'assurance complémentaire santé n'a cessé d'évoluer. Nous assistons à de nombreuses réformes. Les politiques ayant conscience de l'importance du système santé dans les préoccupations des Français comme l'indique l'étude IPSOS qui place au second rang, l'avenir du système social (IPSOS 2020).
Néanmoins, on observe un phénomène de balancier entre cette volonté de rendre l'assurance Santé accessible à tous les français et l'objectif de maîtrise des coûts.

Le contrat responsable :
Un contrat responsable, comme son nom l'indique, a pour objectif principal de responsabiliser le patient sur ses dépenses en santé.
Ainsi, les complémentaires santé « responsables » doivent respecter un cahier des charges dicté par le gouvernement.
Concrètement, cela signifie que certains remboursements sont obligatoires, d'autres interdits ou plafonnés par la loi.
Par ailleurs, des garanties plancher, à savoir minimales, ont été instaurées. L'assureur ne peut alors pas rembourser en-deçà de ces paliers.
Cette notion est issue de la loi du 13 août 2004 relative à l'Assurance Maladie. L'objectif de cette loi était de mieux encadrer les dépenses de santé des Français afin de limiter le déficit de la Sécurité Sociale.
En parallèle de la mise en place du parcours de soins coordonnés autour du médecin traitant, de franchises et de la participation forfaitaire d'un euro, a ainsi été créée la notion de contrat responsable. Celui-ci consiste principalement à inciter les patients à respecter le parcours de soins coordonnés afin d'être mieux remboursé.
Les caractéristiques du contrat responsable ont évolué avec la loi de financement rectificative de la Sécurité Sociale (LFSS) pour 2014, puis par un décret d'application publié le 18 novembre 2014.

Jusqu'ici, les contrats responsables ne devaient contenir obligatoirement que des garanties plancher. La réforme a également instauré des plafonds de remboursement en optique et en cas de dépassements d'honoraires.
Le régime fiscal et social des contrats responsable apparait plus favorable et permet :

- Une exonération des charges sociales pour les contributions versées par l'employeur dans la limite d'un certain plafond,
- Un taux réduit de taxe de solidarité additionnelle rénovée (TSA rénovée) de 13,27% au lieu de 20,27%,
- Pour les salariés : une déduction de la part salariale dans le calcul de l'impôt sur le revenu.

L'ANI :
La loi ANI (Accord National Interprofessionnel) de 2013 est entrée en vigueur le 1er janvier 2016. Sa principale mesure est d'obliger les employeurs privés y compris les associations à proposer une mutuelle à l'ensemble de leurs salariés. C'est ce qu'on appelle la mutuelle d'entreprise. Par ailleurs, l'accord prévoit un financement de cette mutuelle par les employeurs à hauteur de 50% minimum. Une convention collective pourra cependant prévoir un système plus avantageux.
Cela permet alors à chaque salarié de bénéficier d'une complémentaire santé à moindre coût. La mutuelle viendra ainsi compléter les remboursements de la Sécurité sociale. Cet ANI a également étendu à 12 mois la portabilité des garanties santé pour les salariés quittant l'entreprise et pouvant prétendre aux indemnités chômage.
Ainsi, et comme le montre la Figure ci-dessous, 21% des établissements avec au moins un salarié au 31/12/2015 proposaient une complémentaire santé avant que la loi ne les y oblige et ont modifié l'offre existante ou l'ont élargie à l'ensemble des salariés, conformément à la loi généralisant la complémentaire santé en entreprise. Ces établissements regroupent 36% des salariés.

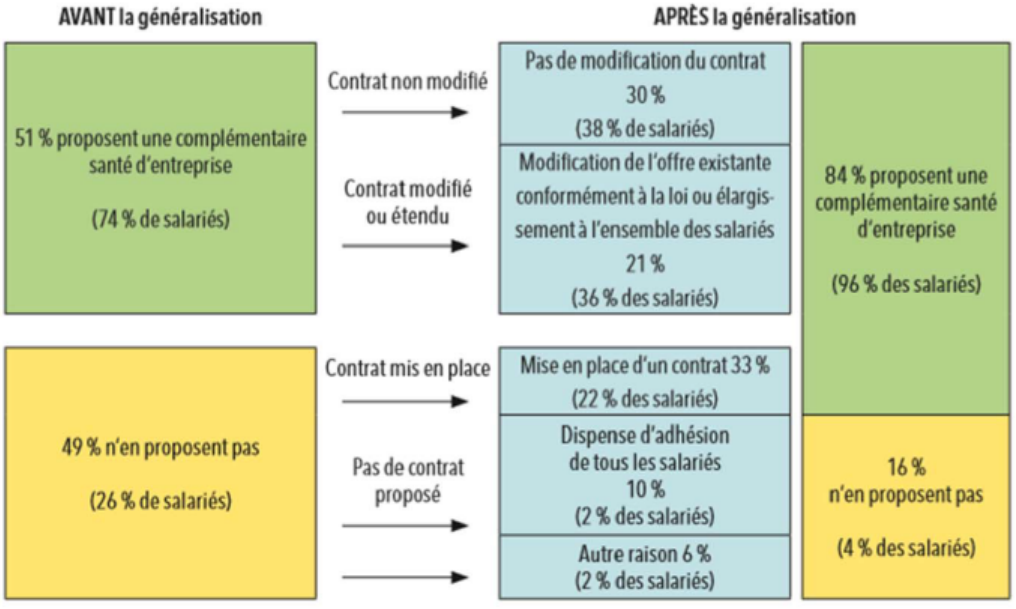

FIGURE 1.5 – Part des établissements proposant une complémentaire santé avant et après la généralisation

La réforme du 100% Santé :
Le 100% Santé ou réforme du reste à charge zéro (RAC 0) est une promesse électorale d'Emmanuel Macron. Avec une entrée en application progressive à partir du 1er avril 2019. Cette loi vise un double

objectif. Le premier est de faciliter l'accès à des équipements de santé coûteux pour les Français, comme les lunettes, les prothèses dentaires ou les aides auditives, en diminuant ou supprimant leur reste à charge. Diminuer cette dépense de santé permettra d'endiguer le renoncement aux soins qui reste assez élevé pour les soins optiques, les prothèses dentaires et les audioprothèses. Le second objectif de cette réforme est de mettre en place davantage de mesures de prévention pour anticiper les problèmes de santé liés à ces postes de soins

La RIA :
La Résiliation Infra-Annuelle (RIA) a été instituée par la loi « relative au droit de résiliation sans frais de complémentaire santé » du 14 juillet 2019. Elle s'inscrit dans la stratégie du gouvernement d'accroître la concurrence sur le marché de l'assurance complémentaire santé pour favoriser l'accès aux soins pour tous. À compter du 1er décembre 2020, la RIA donnera la possibilité de résilier et de changer le contrat de complémentaire santé en cours d'année à l'issu d'une année de souscription. Autrement dit, après 1 an d'ancienneté, il ne faudra plus attendre la date d'échéance du contrat de complémentaire santé pour le résilier.

En résumé de cette section, le système de santé en France, très dynamique d'un point de vue législatif, se compose d'un régime obligatoire auquel s'ajoute un régime complémentaire. Ce dernier représente un marché partagé par les trois acteurs que sont les institutions de prévoyance, les sociétés d'assurances et les mutuelles.

1.1.6 Le marché de la santé

Nous allons ici nous intéresser au marché santé des collectives en compétition avec celui de l'individuel. L'objectif de cette section est d'avoir une vue macro sur un historique de plusieurs années du comportement de la branche santé tous acteurs confondus. Nous allons voir comment le marché a été influencé par les évolutions réglementaires et les nouveaux modes de consommation.
Chaque année, la Direction de la recherche, des études , de l'évaluation et des statistiques (DRESS 2020) publie un rapport sur la situation financière des organismes complémentaires assurant une couverture de santé. Nous regardons tour à tour les cotisations, les prestations et le résultat technique de ces deux branches.

Les cotisations :
Depuis le 1er janvier 2016, toutes les entreprises sont tenues de proposer à leurs salariés une couverture complémentaire collective en santé. Sur le marché des collectives, l'augmentation de la masse des cotisations collectées confirme la progression des contrats collectifs par rapport aux contrats individuels, mouvement de fond antérieur à la mise en place de la généralisation de la complémentaire santé d'entreprise.
D'après le graphique ci-dessous, en 2019, les contrats collectifs représentent 48 % des cotisations collectées en santé, contre 44 % en 2015.
Ainsi, depuis 2015, dernière année avant la mise en place de la généralisation de la complémentaire santé d'entreprise, les contrats collectifs ont gagné 4,1 points de parts de marché.
Cette hausse de la part des contrats collectifs est relativement régulière depuis 2016, où elle avait enregistré une plus forte hausse (de l'ordre de +0,8 point par an après un pic à +1,8 point en 2016).
En favorisant le dynamisme de l'activité en collectif, la généralisation de la complémentaire santé d'entreprise aurait donc contribué à la hausse de la part des contrats collectifs. Cette tendance à la hausse de la part des contrats collectifs est cependant visible depuis au moins 2011.
Ainsi, la généralisation de la complémentaire santé d'entreprise ne semble pas avoir généré une transformation brutale du marché, mais plutôt l'avoir accentuée. En effet, de nombreuses entreprises couvraient déjà leurs salariés via des contrats collectifs avant 2016.

La réforme a donc conduit à accroître la part de salariés couverts par une complémentaire collective, en partie du fait de salariés nouvellement couverts, mais principalement via un transfert de salariés couverts par une couverture complémentaire individuelle vers une complémentaire collective.

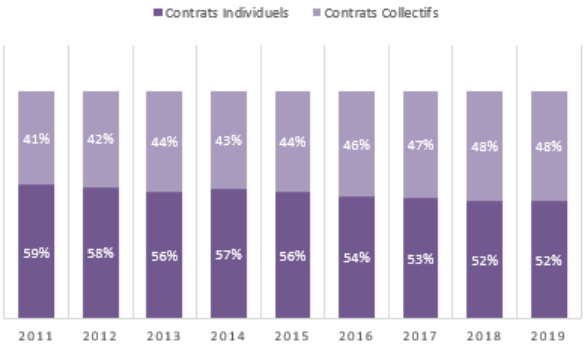

FIGURE 1.6 – Part des contrats collectifs et individuels dans l'ensemble des cotisations en « frais de soins » entre 2011 et 2019 En % des cotisations collectées.

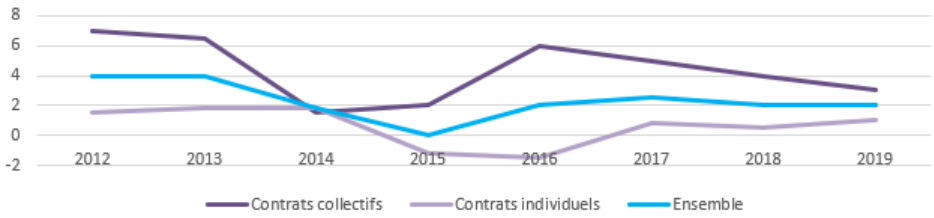

FIGURE 1.7 – Évolution de la masse des cotisations en santé entre 2012 et 2019. (en %)

Les prestations :
Depuis 2018, la croissance des prestations est légèrement supérieure à celle des cotisations.
Cela pourrait s'expliquer en partie par l'augmentation de la part de marché des contrats collectifs, qui reversent plus de prestations en pourcentage des cotisations que les contrats individuels.
Ainsi, en 2019, 47 % des prestations servies par les organismes complémentaires sur la consommation des soins et biens médicaux l'ont été au titre de contrats individuels et 53 %, au titre de contrats collectifs après respectivement 48 % et 52 % en 2018.

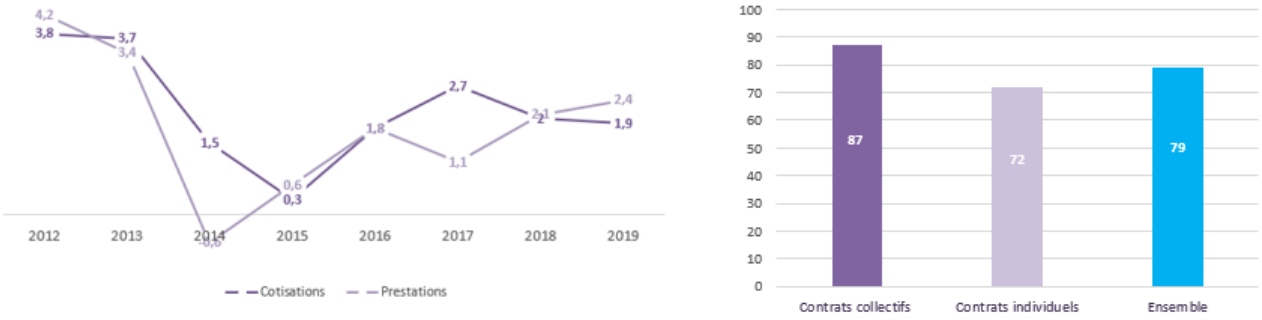

FIGURE 1.8 – De gauche à droite : Évolution des cotisations et prestations en santé entre 2012 et 2019 ; % de prestations sur cotisations collectées en 2019.

Le résultat technique :
Le résultat technique en santé est au global excédentaire en 2019. Il s'élève à 462 millions d'euros, soit 1,2 % des cotisations collectées hors taxe. Mais ce chiffre est une moyenne qui dissimule un écart significatif entre contrats collectifs et individuel.
Depuis 2011, les contrats collectifs sont techniquement en moyenne déficitaires.
Ils présentent un déficit de l'ordre de 4 % des cotisations tandis qu'à l'inverse, les contrats individuels continuent à dégager en moyenne des excédents de l'ordre de 6 % des cotisations. En 2019, l'écart de rentabilité continue à se creuser entre ces deux types de contrats.

en m€	Individuels	Collectifs	Total
Produits	19 798	18 660	38 458
Charges	18 592	19 404	37 996
Resultat technique	1 206	-744	462

FIGURE 1.9 – Résultat technique 2019 en santé en millions d'euros

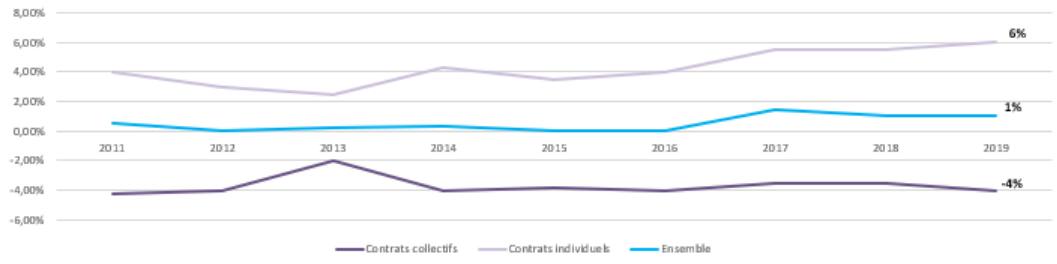

FIGURE 1.10 – Résultat technique en santé entre 2011 et 2019 En pourcentage des cotisations collectées

Conclusion :
La santé en France occupe une place importante. En 2019, la consommation de soins et biens médicaux s'élève à 208 milliards d'euros, soit 8.6% du Produit Intérieur Brut (DRESS' 2020).
Elle est souvent au coeur de débats politiques et économiques et donne lieu à de nombreuses évolutions législatives. Le déficit de la Sécurité Sociale laisse penser un désengagement avenir qui serait compensé par le régime complémentaire.
Sur ce marché de la complémentaire santé, nous observons une part de marché des collectives en progression au détriment des contrats individuels.
La DRESS nous donne également un historique du résultat technique par type de contrats et attire notre attention sur un déficit de la branche santé collective.

1.2 Présentation de l'étude

Introduction

Nous avons abordé en section précédente, l'omniprésence de la couverture médicale qu'elle soit de nature obligatoire ou complémentaire.
Avec l'ANI, cette dernière rendue obligatoire chez les populations actives, elle s'est ainsi généralisée.
Les assureurs de la place sont confrontés, depuis quelques années, à un marché dynamique d'un point de vu réglementaire, concurrentiel mais également chahuté par de nouveaux modes de consommation.
On observe une consommation en soins et biens médicaux en augmentation. Ceci se traduit par une dérive sur les portefeuilles des assureurs.
L'objectif de cette étude est de proposer une méthode basée sur un modèle mathématique robuste pour mesurer, évaluer le niveau de dérive à venir et répondre aux besoins des assureurs en termes de pilotage et suivi du portefeuille.
Pour cela, nous allons brièvement rappeler l'importance de la tarification qui est centrale en matière d'assurance santé. Nous poursuivrons le chapitre en énumérant les facteurs intrinsèques au risque santé. Ils sont déterminants dans l'estimation au plus juste du tarif, mais également dans le pilotage du portefeuille et plus précisément lors de l'évaluation de la dérive en santé collective.

1.2.1 La tarification

La tarification est un enjeu majeur pour l'assureur pour garantir la rentabilité de son activité.
Dans un contexte où la concurrence est rude, l'assureur va établir des tarifs attractifs pour rester compétitif tout en veillant à couvrir les engagements futurs auxquels il fera face. Définir le meilleur tarif c'est donc définir le bon niveau de ratio compétitivité/rentabilité.
Ceci passe nécessairement par la connaissance fine de son portefeuille, autrement dit du mode de consommation de ses assurés.
Dans cette partie, nous rappelons les principes de base de la tarification d'un contrat en santé.
La souscription d'un contrat d'assurance par un assuré se matérialise par le versement d'une cotisation à l'organisme assureur, de manière que ce dernier accepte le transfert de risque. Le niveau du montant de la cotisation dépend alors du niveau de risque transféré ainsi que d'autres facteurs définis par chaque organisme d'assurance. De manière générale, la détermination de la cotisation doit respecter des critères fondamentaux :

- Les cotisations doivent être suffisantes pour couvrir les risques assurés par les produits commercialisés et plus globalement qui doivent respectées les règles de souscription fixant le niveau de rentabilité attendu.
- Les cotisations doivent intégrer l'ensemble des frais liés à la gestion des couvertures distribuées.

La tarification autrement dit la détermination de la prime se définit par les éléments suivants :

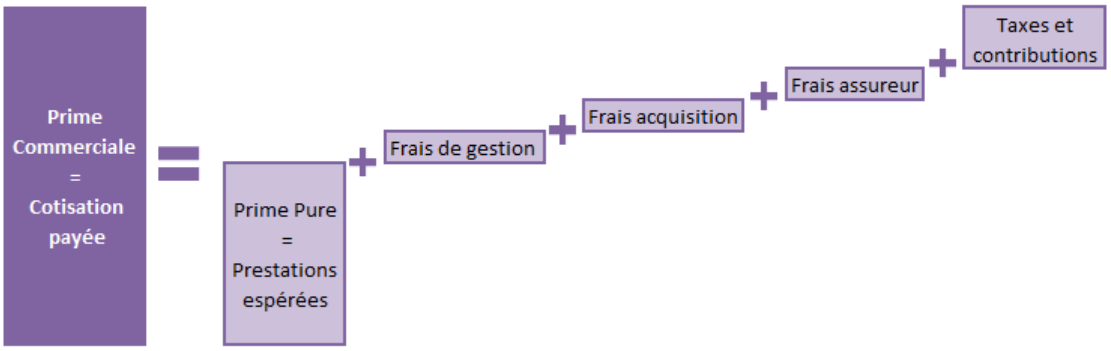

FIGURE 1.11 – Eléments constitutifs d'une prime

on a ainsi :

$$\text{Prime commerciale} = \text{Prime Pure} \times \frac{1 + taxes}{\text{1-Taux de chargements}}$$

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.
Le mode d'expression de la cotisation peut se faire en pourcentage de PMSS, ce qui signifie que chaque année la cotisation est revalorisée. D'autres modes d'expressions sont possibles : % salaires, forfaitaire...
Il existe plusieurs structures de cotisations :

- Uniforme : tous les adhérents payent la même cotisation quel que soit le nombre de bénéficiaires de chacun. (Solidarité maximum)
- Isolé/Famille : séparation de la cotisation en deux niveaux :
  - Isolé : cotisation pour les adhérents seuls
  - Famille : cotisation pour les adhérents avec leur conjoint et leurs enfants.
- Isolé/Duo/Famille : séparation de la cotisation en trois niveaux.
- Adulte/Enfant : pour éviter les effets d'antisélection dus au choix de la cotisation.

— Adulte/Conjoint/Enfant : pour éviter les effets d'antisélection du conjoint.
— Adulte+Enfant/Conjoint facultatif : cas particulier pour certaines conventions collectives.

1.2.2 Les facteurs discriminants en santé collective : quels impacts sur la consommation ?

Le but de cette partie est d'énumérer les facteurs pouvant avoir un impact sur la tarification et donc sur la consommation. En effet, l'évaluation du risque santé est d'autant plus complexe qu'il possède des caractéristiques propres.
Il s'agit de facteurs tels que le niveau de garantie, l'âge, le sexe, le lieu de résidence, le revenu, le secteur d'activité mais encore le risque d'antisélection ou l'aléa moral et la liste n'est pas exhaustive, peuvent avoir un fort impact sur la consommation.

L'antisélection :
Le risque d'antisélection se retrouve dans la plupart des branches d'assurance. En assurance santé, il se traduit par le fait que ce sont principalement des assurés se sachant malades qui vont souscrire. Ainsi, leur consommation présentera des écarts avec la consommation moyenne et la difficulté résidera dans la mesure de ces écarts. En assurance collective, on retrouve ce risque lors d'une adhésion facultative.

L'aléa moral :
Le risque moral est un risque inhérent à l'assuré. Il se définit par le fait qu'un assuré va modifier son comportement face au risque en fonction de sa couverture. Ce risque est beaucoup plus important en assurance santé complémentaire que pour d'autres branches d'assurance. En effet, le fait d'être couvert à un niveau de garantie supérieur entraîne une autre vision du risque de la part de l'assuré et l'incite à consommer plus particulièrement sur le poste optique. De plus, le risque moral est d'autant plus important en fonction du revenu. En effet, le reste à charge a un effet modérateur qui est fonction du niveau de vie. Ainsi, intuitivement, le risque moral est d'autant plus fort que le revenu est important.

Le niveau de garantie :
Le niveau de garantie est un facteur déterminant car plus le niveau de garantie est élevé, plus l'assuré sera tenté de consommer.

L'âge :
L'âge est incontestablement le facteur le plus discriminant en santé. L'âge de l'assuré a une influence significative sur la consommation. On observe une courbe en forme de "W" avec un pic de consommation à la naissance, à l'adolescence avec l'orthodontie, puis une phase de croissance dont le rythme s'accélère avec l'âge.

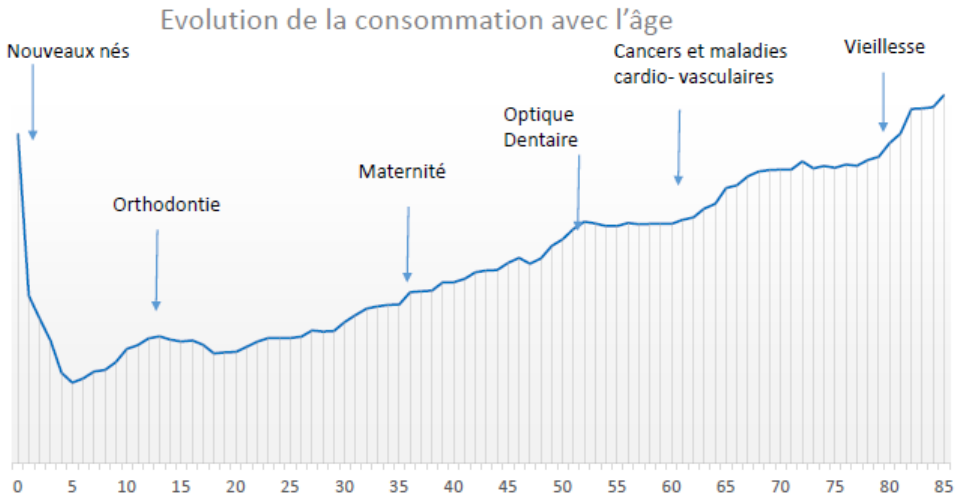

FIGURE 1.12 – Évolution de la consommation avec l'âge

Le genre :
En moyenne, la consommation des femmes est plus élevée que chez les hommes. Néanmoins sur certains postes on observe une consommation plus élevée chez l'homme : c'est le cas pour les honoraires d'hospitalisation.

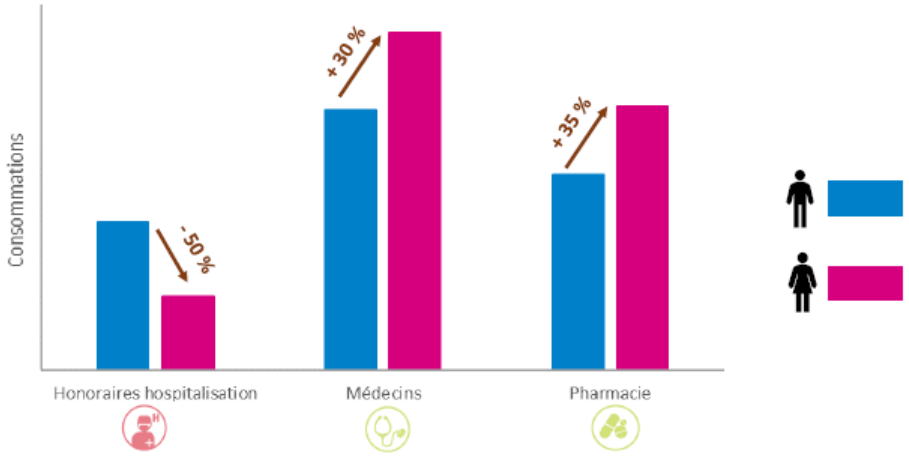

FIGURE 1.13 – Consommation par genre

Le lieu de résidence :
En France, la consommation médicale est plus importante dans certaines régions comme l'Ile de France et PACA que dans les régions rurales (Limousin, Auvergne...). Les premières disposent en effet d'une offre médicale abondante, accessible et proposent des prestations plus chères liées notamment au coût de la vie et à la capacité financière des résidents : même en Ile de France, le coût moyen d'une consultation est plus élevé au coeur de la capitale que dans une banlieue défavorisée.

CSP :
Il existe une inégalité sociale : les plus riches et les personnes ayant un niveau d'études plus élevé sont plus sensibles à la prévention et renoncent moins aux soins. Ils vivent mieux et plus longtemps que les personnes plus pauvres ou avec un niveau d'éducation moins élevé.

1.2.3 Le pilotage technique : projections et hypothèse de dérive

Point d'entrée en santé collective, nous venons de voir que la tarification est déterminante dans la rentabilité du portefeuille. Cette rentabilité s'inscrit dans le temps d'où la nécessité de surveiller le portefeuille.
Ceci passe nécessairement par l'établissement de comptes de résultats qui peuvent être de deux natures : réels ou prévisionnels.
L'établissement des comptes réels et prévisionnels vont permettre le pilotage technique du portefeuille et l'analyse de la rentabilité.
L'analyse de la rentabilité se traduit par l'indicateur : le ratio S/P autrement dit le quotient du montant estimé des sinistres rapporté aux cotisations nettes de taxes et de chargements.
On écrit alors :

$$S/P = \frac{\text{Charge de sinistres}}{\text{Cotisations nettes}}$$

Ce S/P donne lieu à trois niveaux possibles :

- S/P > 1, la charge de sinistres est supérieure aux cotisations nettes, le contrat est déficitaire. Le contrat va devoir être redressé pour revenir à l'équilibre.
- S/P < 1, la charge de sinistres est inférieure aux cotisations nettes, le contrat est bénéficiaire autrement dit rentable.
- S/P $= 1$, la charge de sinistres est égale aux cotisations nettes, le contrat est tout juste à l'équilibre.

En outre, ce ratio est un excellent indicateur pour le suivi technique, cela permet d'identifier un besoin éventuel de revalorisation tarifaire des contrats, d'anticiper la poursuite d'une tendance à la hausse des dépenses de santé ou d'un ralentissement, d'estimer les effets de certaines réformes de l'Assurance Maladie, d'anticiper une aggravation du risque due par exemple à une épidémie ou une augmentation de maladies chroniques et de prendre en compte la dérive naturelle causée par le vieillissement de la population.
Ce ratio peut être observé à différents moments (trimestriel, semestriel ou annuel) et différentes survenances (passées, futures).
C'est dans ce contexte que l'on comprend l'importance d'établir des comptes prévisionnels avec une vision aussi précise que possible.
Autrement dit, l'assureur doit pouvoir évaluer le niveau de sinistralité de son portefeuille pour une survenance future.
De manière plus concrète voyons comment cela se passe.
Pour estimer le compte de résultats en cours d'année, une des méthodes de projection est la déformation du compte de résultats de l'année précédente.
Pour ce faire nous disposons de :

- Paramètres connus :
  - le chiffre d'affaires hors taxes N.
  - le taux de chargement contractuel.
  - la charge de sinistres N.
  - le résultat technique et le S/P N sont déduit des éléments précédents.

- Hypothèses retenues :

— le taux de chute c'est-à-dire le taux de résiliation des affaires en portefeuille.
— le taux de revalorisation tarifaire éventuelle.
— le chiffre d'affaires hors taxes issu de la production nouvelle, c'est-à-dire les primes encaissées au titre des affaires nouvelles.
— la dérive de sinistralité, c'est-à-dire l'alourdissement de la charge de sinistre santé.

Ainsi avec ces éléments nous pouvons définir le compte de résultat de l'année N comme suit :

— Le chiffre d'affaires hors taxe de l'année N+1 :

$$CA_{N+1} = CA_N \times (1 - \text{taux de chute}) \times (1 + revalorisation) + CA_{\text{affaires nouvelles}}$$

— le ratio S/P hors taxes de l'année N+1 :

$$S/P_{N+1} = S/P_N \times \frac{1 + \text{dérive de sinistralité}}{1 + \text{revalorisation}}$$

L'hypothèse la plus variable est celle de l'estimation de la **dérive de sinistralité**, l'évolution du coût du risque de notre portefeuille.
Sur un portefeuille identique sur deux exercices clôt (N-1 et N), la dérive de sinistralité est déterminée :

$$\text{Dérive de sinistralité N} = Evolution N/(N-1) = \frac{\text{Coût moyen N}}{\text{Coût moyen N-1}} - 1$$

Où

$$\text{Coût moyen} = \frac{\text{Consommation}}{\text{Nombre de bénéficiaires}}$$

Nous nous proposons dans ce mémoire de définir à travers l'étude de différents modèles, celui qui sera le meilleur au sens prédictif à déterminer la dérive de sinistralité en santé.

Conclusion
Partant d'une vue macro du marché de la santé, la complémentaire collective est une branche assurantielle en déficit qui s'inscrit dans le temps. En cause, le cadre législatif très contraignant. Une concurrence accrue rendant l'exercice de tarification et de pilotage compliqué. Une consommation qui évolue et croît dans le temps. Dans ce contexte complexe, l'assureur doit évaluer de façon précise le risque et l'engagement qu'il prend. Bien connaitre son portefeuille est un atout majeur pour définir le comportement futur et d'anticiper une dérive de sinistralité.

1.3 Présentation des données

La troisième et dernière partie de ce chapitre sera consacrée à l'analyse, aux traitements et aux statistiques sur les bases de données à notre disposition. Cette partie étant une étape nécessaire et obligatoire et ce doit être réalisée en amont de la phase de modélisation dont le point d'entrée est « la donnée » et dont la qualité doit être au rendez-vous pour assurer des résultats concluants.

1.3.1 Les bases internes assureurs

Présentation générale des bases internes

Les données utilisées dans le cadre de notre étude sont issues d'un portefeuille de contrats de complémentaire santé collective, en délégation de gestion pour le compte de Groupama Gan Vie. Ces données sont disponibles à travers trois tables :

- Table des contrats : contient l'ensemble des contrats de la population couverte.
- Table des bénéficiaires : il s'agit de la démographie détaillée de la population couverte.
- Table des prestations : on retrouve les informations relatives à la consommation médicale de la population couverte.

Deux années d'études sont disponibles : 2018 et 2019.

La table des contrats :
Cette table disponible sur les exercices 2018, 2019 comporte 17 variables. Elle présente les caractéristiques propres aux contrats, comme par exemple :

- Numéro de contrat,
- Date d'effet du contrat,
- Date de résiliation du contrat,
- Type de population (actifs, inactifs),
- Niveau du contrat (base, surcomp),
- Nombre d'adhérents à la date d'exercice,
- Code postal de la société ...

La table des bénéficiaires :
Cette table disponible sur les exercices 2018, 2019 comporte 20 variables. Elle présente les caractéristiques propres aux bénéficiaires, comme par exemple :

- Numéro de contrat,
- Type de population (actifs, inactifs),
- Numéro d'adhérent,
- Numéro de bénéficiaire,
- Type de bénéficiaire,
- Date de naissance du bénéficiaire,
- Prénom du bénéficiaire,
- Sexe du bénéficiaire,
- Portabilité,
- Les 5 premiers chiffres du numéro de Sécurité Sociale...

La table des prestations :
Cette table disponible sur les exercices 2018, 2019 comportes 55 variables. Elle présente les caractéristiques propres à la consommation médicale, comme par exemple :

- Numéro de contrat,
- Type de population (actifs, inactifs),

— Collège,
— Numéro d'adhérent,
— Numéro de bénéficiaire,
— Type de bénéficiaire,
— Famille de l'acte de soin,
— Sous Famille de l'acte de soin,
— Acte de soin,
— Acte de soin sous Noémie,
— Libellé de l'acte de soin,
— Date de réalisation du soin,
— Date de règlement du soin,
— Portabilité,
— Montant frais réels,
— Montant de base de la Sécurité Sociale,
— Montant de remboursement de la Sécurité Sociale,
— Montant du ticket modérateur,
— Montant remboursement complémentaire santé...

Dans le cadre de ce mémoire, toutes les variables ne seront pas exploitées car non indispensable dans notre étude.
Des liaisons entre les différentes tables au travers de variables communes peuvent se faire comme l'illustre la figure ci-dessous :

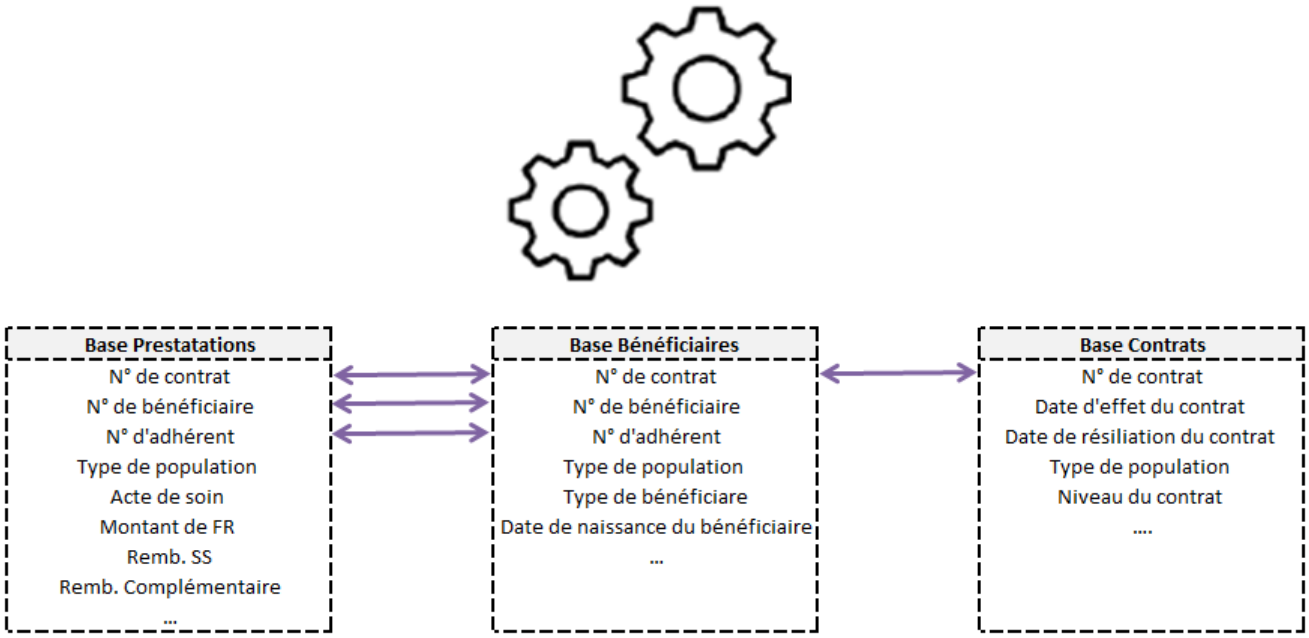

FIGURE 1.14 – Les jointures entre les bases de données

Traitements des données internes

Comme précisé précédemment nous sommes sur un périmètre de santé collective en délégation de gestion. Nous disposons de deux années de survenance durant lesquelles des « entrées/ sorties » ont pu être enregistrées durant cette période.

Périmètre d'étude :

Dans un premier temps, nous présenterons les statistiques du portefeuille sur les deux exercices en y conservant l'évolution, c'est à dire les "entrées/sorties" constatées à chaque période.
Par la suite et pour donner du sens aux statistiques nous serons contraints de prendre un panel de contrats présents sur toute la période d'analyse et de faire l'hypothèse structurante que les « entrées / sorties » de salariés ont des caractéristiques identiques.
C'est une méthode assez usuelle en assurances collectives quand on ne connait pas le détail des populations et les niveaux de garanties.

Les données manquantes :

La table des bénéficiaires présentes certaines valeurs manquantes sur la variable sexe.
Sur la base de contrats communs présents en 2018 et 2019, nous avons un total de 27% bénéficiaires concernés par ces valeurs manquantes répartis comme suit :

Adhérents	Conjoints	Enfants	Total
0%	3%	78%	27%

FIGURE 1.15 – % des bénéficiaires avec valeurs manquantes

La problématique rencontrée concerne principalement les enfants.
Cette variable étant une variable sensible et importante pour la suite des travaux, il semble important de procéder à une correction.
Comme vu précédemment, la table des bénéficiaires est celle qui renseigne sur les caractéristiques des bénéficiaires. Entre autres, nous disposons du prénom du bénéficiaire et de son genre (masculin ou féminin).
Ainsi, à partir des données des deux tables de bénéficiaires 2018 et 2019, nous avons créé une table à deux variables : prénoms et genre. Celle-ci sera utilisée en correspondance pour définir le sexe des bénéficiaires manquants.
Cette méthode va permettre de corriger 83% des valeurs manquantes. Les dernières valeurs manquantes ont été complétées à dire d'expert.

Outils :

Toute l'analyse statistique s'est réalisée par l'exploitation des bases de données sous le logiciel SAS.
Compte tenu de la volumétrie importante des bases, notamment celles des prestations (3Go chacune) nous avons agrégé certaines lignes à l'aide d'une clé de concaténation.
Cette agrégation n'altère en rien la pertinence des informations de la base dans le cadre de notre étude de la consommation médicale.

Statistiques descriptives internes

Maintenant que l'on a présenté les tables et leurs variables, nous pouvons présenter les statistiques pour se faire une meilleure vision du portefeuille et de la consommation médicale de la population couverte.

La démographie

Comme nous l'avons spécifié en chapitre 1, le portefeuille étudié concerne des contrats de complémentaire santé collective, en délégation de gestion pour le compte de Groupama Gan Vie. Pour commencer, il est intéressant de connaitre combien de contrats sont présents sur les deux dates d'exercices. Les entrées et sorties sont à l'origine de ces variations d'une année sur l'autre.

Exercice	2018	2019
Nombre de contrats	185	206

FIGURE 1.16 – Nombre de contrats présents au portefeuille

Regardons à présent le nombre de personnes présentes en portefeuille mais également son exposition au cours de l'exercice qui est définie comme le temps de présence dans l'année de chaque bénéficiaire.

		Total bénéficiares	Adhérents	Conjoints	Enfants
2018	Nombre	87 019	41 706	15 646	29 667
	Exposition	76 564	36 160	14 202	26 202
2019	Nombre	97 196	47 654	16 939	32 603
	Exposition	85 968	41 143	15 310	29 515

FIGURE 1.17 – Démographie de la population couverte : en nombre et exposition de bénéficiaires.

La répartition hommes/femmes globale et par type de bénéficiaires est présentée sur le tableau suivant. Sur la population des adhérents, nous avons une répartition à 62% d'hommes et 38% de femmes.

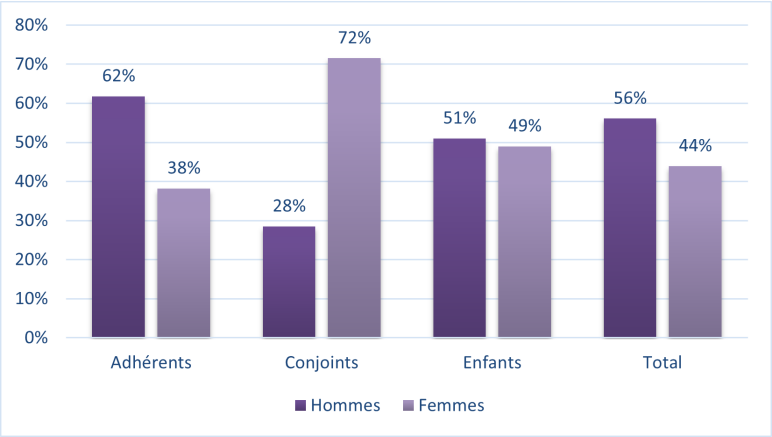

FIGURE 1.18 – Répartition par sexe des bénéficiaires

Le facteur âge a un réel impact sur la consommation médicale.
L'âge moyen du portefeuille est de 33 ans. Cette information prise seule ne peut être pertinente dans une analyse de la consommation médicale.

Exercice	Moyenne portefeuille	Adhérents	Conjoints	Enfants
Age moyen	33	43	46	12

FIGURE 1.19 – Age moyen du portefeuille et par bénéficiaires

La pyramide des âges du portefeuille permet de rendre compte la manière dont est répartie la population.
Elle est relativement bien répartie en fonction de l'âge jusqu'à environ 65 ans. La population par âge devient moins importante ce qui peut entrainer des résultats peu représentatifs sur les âges élevés dans la suite de l'étude.

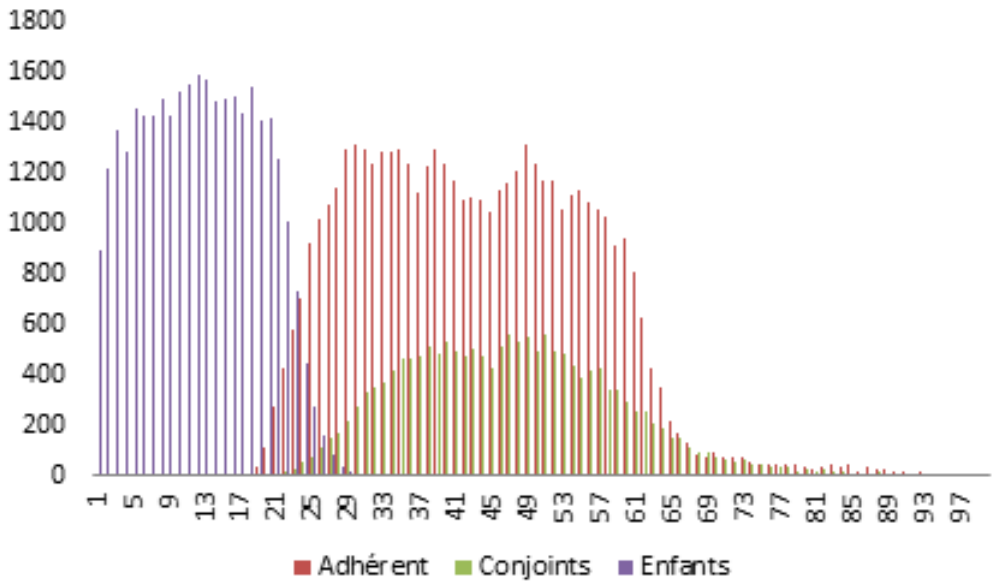

FIGURE 1.20 – Pyramide des âges du portefeuille

Enfin, sur le risque santé, la caractéristique "région" a un impact sur la consommation de soins et biens médicaux.
En effet, si sur certaines régions on observe des déserts médicaux, d'autres à l'inverse proposent une offre abondante.
Notre portefeuille est fortement reparti sur les régions Ile-de-France et Auvergne-Rhône-Alpes.

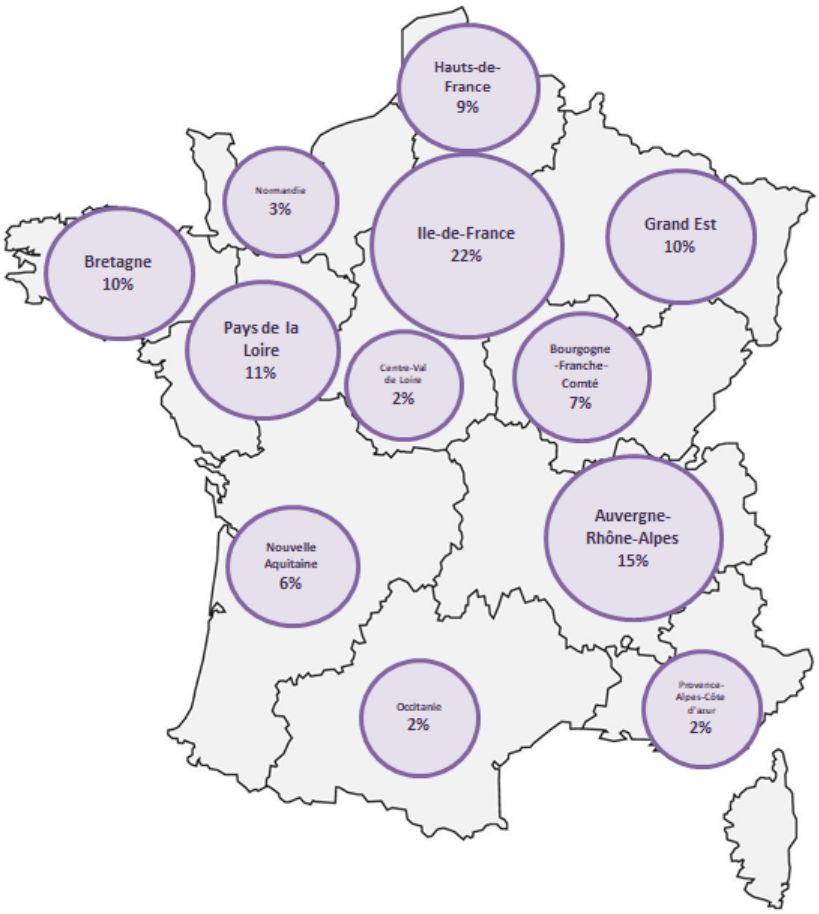

FIGURE 1.21 – Répartition du portefeuille par régions (exposition par bénéficiaire)

Les consommations

Dans cette section nous allons nous intéresser à la consommation médicale. Dans un souci de ne pas alourdir ce mémoire de données statistiques nous avons fait le choix de vous les présenter sur l'exercice 2019.
La part des bénéficiaires consommant est logiquement élevée. En effet, au cours d'une année, la probabilité d'avoir recourt à sa complémentaire santé est très élevée.

Adhérents	Conjoints	Enfants	Total
95%	93%	90%	93%

FIGURE 1.22 – Proportion de consommant sur l'exercice 2019.

La répartition sur les 6 grands postes de consommation est différentes suivant que l'on regarde les frais engagés, ou le remboursement de la sécurité sociale ou encore le remboursement de la complémentaire

santé.

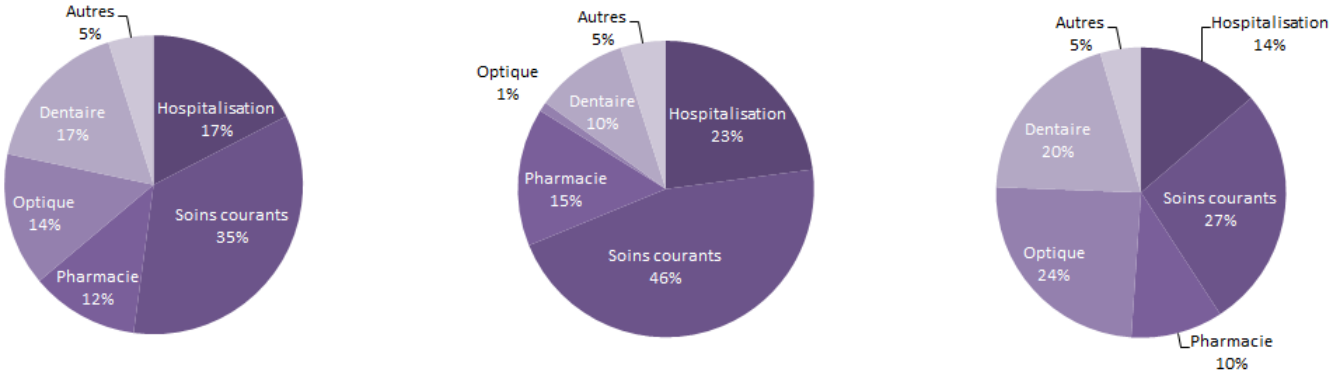

FIGURE 1.23 – Part des différents postes : frais réels (camembert 1), remboursement sécurité sociale (camembert 2), remboursement complémentaire santé (camembert 3).

Ceci est confirmé par le graphique ci-dessous où la part pris en charge par la sécurité sociale sur le poste optique est très inférieure à celle de la complémentaire. La tendance est inversée sur l'hospitalisation : la part de la sécurité sociale est plus importante que celle de la complémentaire.

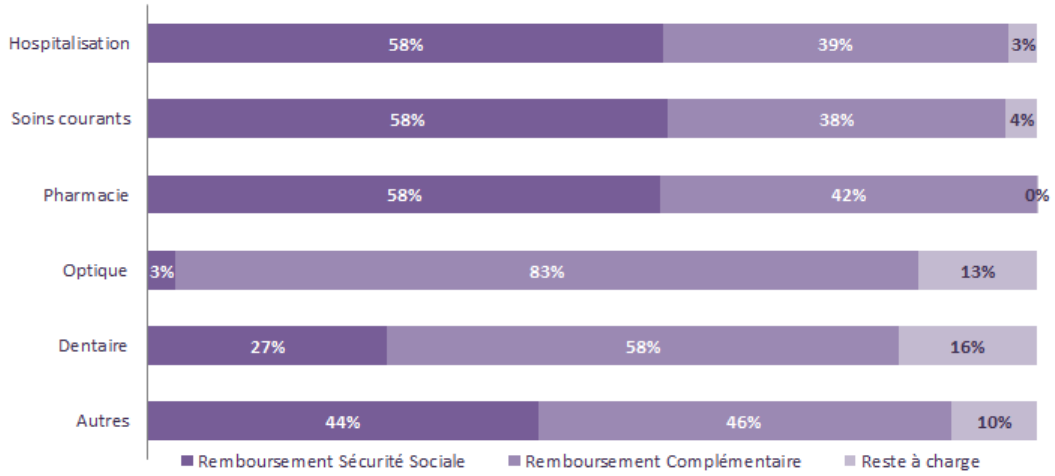

FIGURE 1.24 – Prise en charge des dépenses de santé 2019 selon les différents postes

Il est intéressant d'avoir une vision de la consommation moyenne par âge. En effet, il s'agit là du facteur le plus discriminant du risque santé : la santé des individus se dégradant avec l'âge.
D'après le graphique ci-dessous, nous avons une courbe qui est assez bien représentative de la consommation moyenne par âge avec des caractéristiques telles qu'une hausse chez les adolescents, on pense notamment à l'orthodontie courante chez cette tranche d'âge mais également à l'optique où les corrections sont plus importantes. Suivie d'une légère baisse chez les 20-30 ans où les besoins sont moindres. Puis à partir de 30 ans, la courbe est de nouveau en croissance notamment sur le poste maternité chez les femmes. Enfin, c'est sans surprise que nous constatons une pente accentuée sur la courbe à partir de 60 ans.

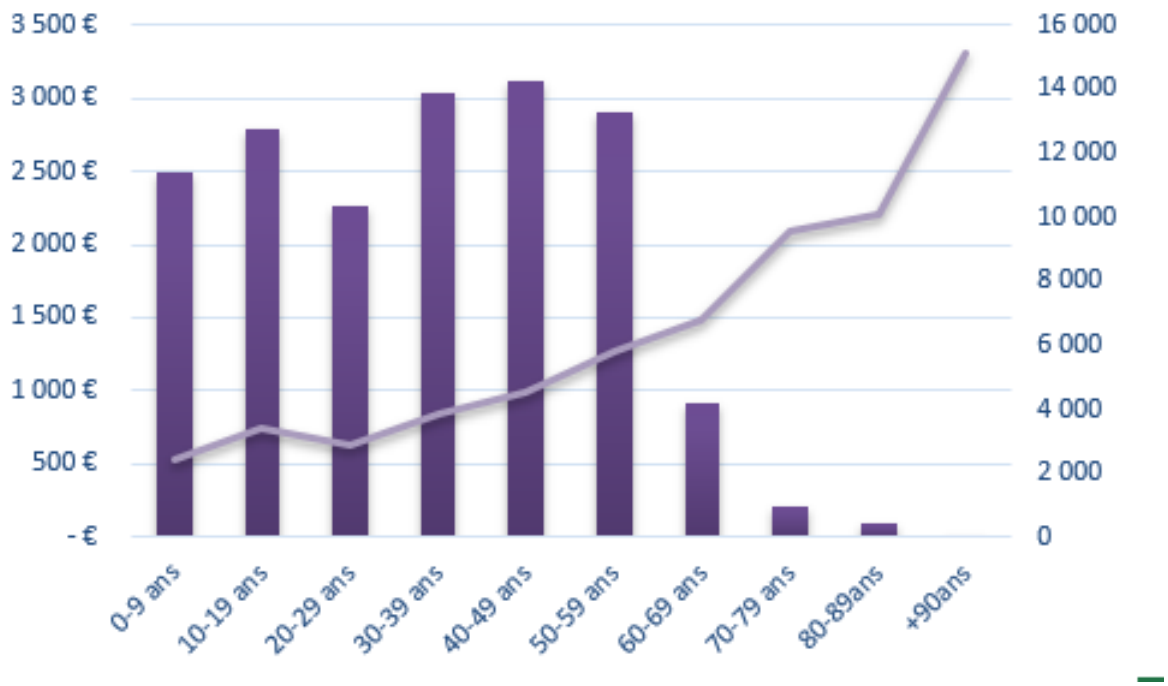

FIGURE 1.25 – Consommation moyenne par âge

Les dépenses

Regardons à présent les statistiques sur les bases des prestations.
En premier lieu, la répartition des dépenses en santé puis sur le poste pharmacie par bénéficiaire indique une consommation majoritaire chez les adhérents suivie à part quasi égale entre les conjoints et les enfants.

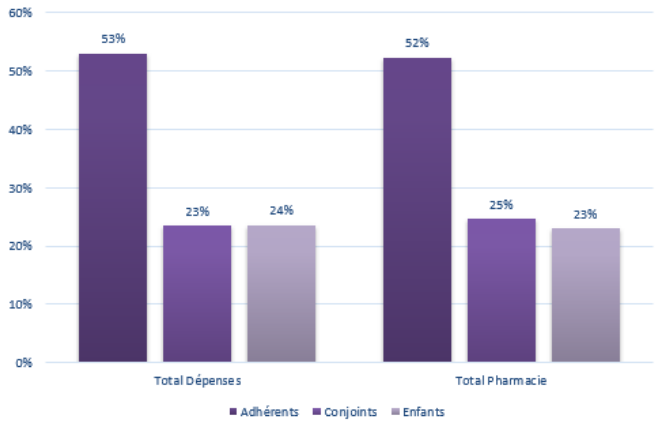

FIGURE 1.26 – Répartition des dépenses totales et pharmacie par bénéficiaire

La répartition par genre indique une plus forte consommation chez les femmes sur les dépenses au global mais également sur le poste pharmacie.

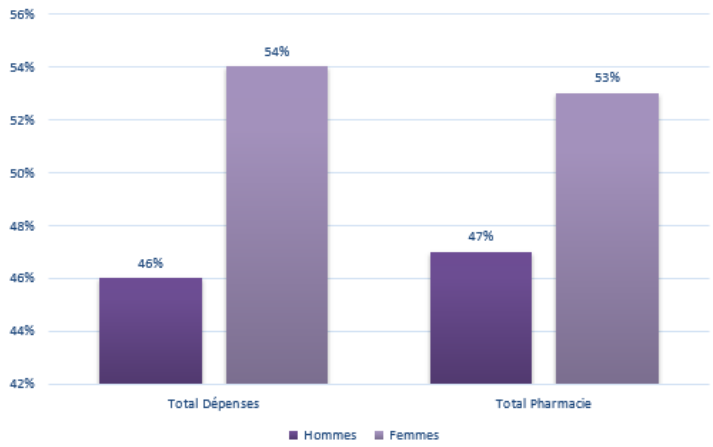

FIGURE 1.27 – Répartition des dépenses totales et pharmacie par genre

Enfin, c'est la région Ile-De-France qui arrive en tête de fil sur la répartition des dépenses par région.

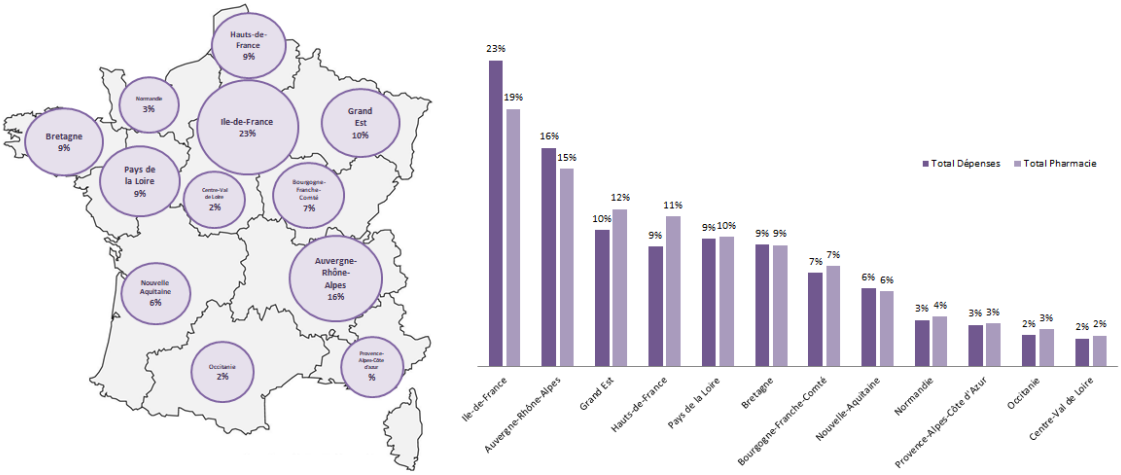

FIGURE 1.28 – Répartition par région des dépenses totales et pharmacie

L'analyse des années 2018 et 2019 nous permet d'évaluer l'évolution des dépenses par grands postes durant cette période.
Pour ce faire nous avons sélectionné sur le panel de contrats disponibles, ceux présents aux deux dates.
Nous avons ensuite calculé un coût moyen par grands postes défini comme suit :

$$\text{Coût moyen} = \frac{\text{Dépenses}}{\text{Nombre de bénéficiaires}}$$

Ici "Dépenses" représentent le total prestation remboursée par le régime complémentaire.

Enfin l'évolution des prestations entre deux dates N et N-1 s'obtient de la façon suivante :

$$\text{Évolution}\frac{\text{N}}{\text{N-1}} = \frac{\text{coût moyen N}}{\text{coût moyen N-1}} - 1$$

Globalement on observe une hausse des dépenses de 3.8% entre 2018 et 2019.
La majorité des grands postes de soins sont en hausse, exception faite sur le poste de soins pharmaceutiques qui est déflationniste d'environ -0.3%.

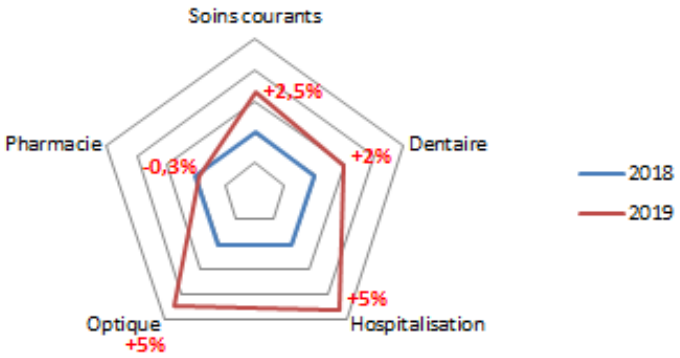

FIGURE 1.29 – Évolution des remboursements complémentaires par grands postes sur les exercices 2019 vs 2018

Sur les postes au global, on observe une hausse des prestations mois après mois à l'exception du mois de juin.
On observe une évolution différente sur le poste pharmacie avec une courbe en dents de scie.

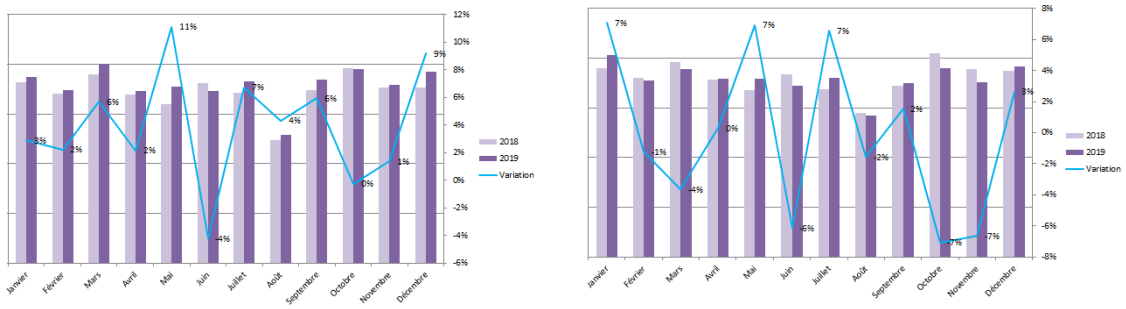

FIGURE 1.30 – Évolution des remboursements complémentaires par mois sur les exercices 2019 vs 2018 (à gauche tous postes confondus, à droite sur le poste soin pharmaceutique)

La portabilité

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. A 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.
Le graphique ci-dessous nous donne l'évolution des prestations réglées au titre de contrats issus de la portabilité.

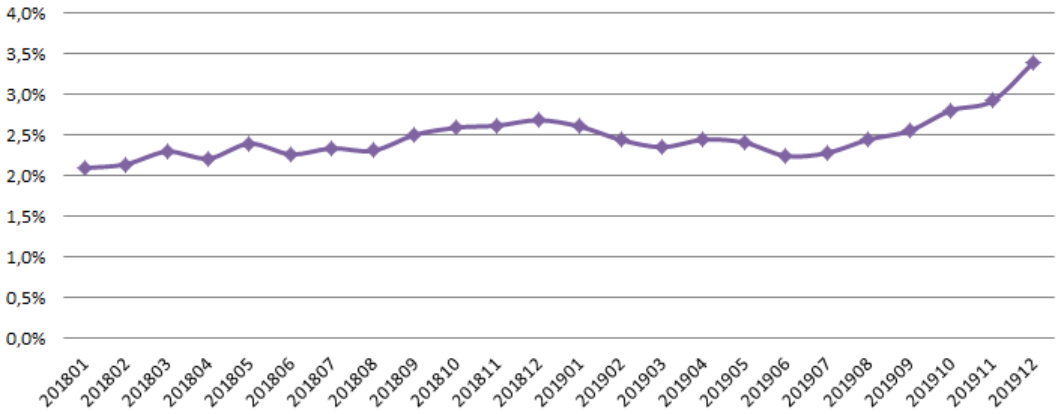

FIGURE 1.31 – % de prestations réglées au titre de la portabilité (moyenne mobile 3 mois)

On observe une hausse des prestations réglées au titre de la portabilité sur le T4 2019.

1.3.2 Les bases externes : l'Open Data DAMIR

Présentation générale

C'est en 1999, que le portail SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie) voit le jour. Le SNIIRAM a pour but de permettre une analyse des dépenses et parcours de soins des assurés des régimes généraux. Il regroupe plusieurs bases de données avec 3 niveaux d'accès différents, défini selon le niveau de confidentialité des données.
En 2009, l'Assurance Maladie a ouvert un accès public aux bases de données sur les dépenses tous soins confondus des Français. Avec une volonté des pouvoirs publics de favoriser l'information des professionnels et les prises de décision en matière de politique publique de santé. Ces bases ont vocation à mieux connaître le système de santé, pour mieux l'utiliser, pour en débattre démocratiquement et pour l'améliorer. Elle s'adresse donc aussi bien aux gestionnaires de l'Assurance Maladie, qu'aux responsables politiques en matière de santé, aux praticiens, et dans une moindre mesure au public.
L'accès libre de ces Open Data conduit ces bases à avoir un niveau de confidentialité de niveau 1. Ceci veut dire que les données sont anonymes et pour permettre le respect de l'anonymat des patients, les données ont été agrégées.
Chaque ligne représente la somme des actes et des montants associés à toutes les variables catégorielles. Ces dernières concernent aussi bien les caractéristiques individuelles (région, tranche d'âge...) que les caractéristiques propres à l'acte consommé (base de remboursement, mois de consommation, nature du prescripteur...).
Il est donc impossible d'identifier un individu. Cette méthode d'agrégation laisse cependant une granularité rendant l'exploitation statistique possible.
L'Open Data DAMIR est l'acronyme du fichier qui regroupe les Dépenses d'Assurance Maladie Inter Régime.
Ce jeu de données concerne l'ensemble des prestations prises en charge par l'Assurance Maladie Obligatoire. Ainsi, l'ensemble des dépenses de remboursement, tous régimes confondus, de l'Assurance Maladie est couvert par cette base de données, à l'exception d'une grande majorité des prestations hospitalières du secteur public.
Ces données étant collectées après le remboursement par la Sécurité Sociale, il est impossible de savoir si une personne a bénéficié, par la suite, d'un remboursement complémentaire.
Comme nous venons de le dire, l'Open Data DAMIR est public. Elle est en accès direct depuis le site AMELI 2020.
On y retrouve un fichier explicatif de toutes les variables de ces bases. Il explique également le changement de nomenclature intervenu à partir de l'année 2015 concernant des ajouts et des modifications de certaines variables.
Chaque année est représentée par ses 12 bases mensuelles. Ces mois correspondent à des mois de règlement du remboursement, et ne correspond donc pas forcément au mois de délivrance de la prestation. Il en va de même pour les années de délivrance des soins. Pour exemple, un acte effectué en décembre 2018, pourra être réglé en février 2019. Cet acte sera donc présent dans la base mensuelle de février de l'année 2019.
Les données sont mises à jour annuellement : au mois de juin. Ce qui veut dire que tous les fichiers mensuels de l'année N seront disponibles en juin N+1.
La volumétrie des bases est gigantesque. Chaque base mensuelle est de l'ordre de 3Go, c'est entre 20 à 22 millions de lignes par mois. Une année c'est donc environ 220 millions de lignes de prestations.
Toutes les bases ont le même format et se composent de 55 variables de type qualitatives et quantitatives.

Axe d'analyse	Libellé de variables
Période	Année de Soins
	Mois de Soins
Bénéficiaires	Sexe du Bénéficiaire
	Tranche d'Age Bénéficiaire au moment des soins
	Qualité du Bénéficiaire
	ZEAT de Résidence du Bénéficiaire
	Région de résidence du Bénéficiare
	Modulation du Ticket Modérateur
	Top Bénéficiaire CMU-C
Prestations	Nature de Prestation
	Nature d'Assurance
	Nature de l'Accident du Travail
	Type d'Enveloppe
	Complément d'Acte
	Motif d'Exonération du Ticket Modérateur
	Taux de Remboursement
	Code Secteur Privé/Public
	Type de Prise en Charge Forfait Journalier
	Indicateur TAA Privé/Public
	Code Qualificatif Parcours de Soins (sortie)
	Nature du Destinataire de Règlement affiné
	Type de Remboursement
Prescripteur	Catégorie du Prescripteur
	Nature d'Activité PS Prescripteur
	ZEAT du PS Prescripteur
	Région du prescripteur
	Statut Juridique PS Prescripteur
	ZEAT d'Implantation Etb Prescripteur
	Région d'implantation de l'établissement
	Catégorie Etb Prescripteur
Exécutant	Catégorie de l' Exécutant
	Spécialité Médicale PS Exécutant
	Nature d'Activité PS Exécutant
	ZEAT du PS Exécutant
	Région de l'Exécutant
	Statut Juridique PS Exécutant
	Mode de Fixation des Tarifs Etb Exécutant
	ZEAT d'Implantation Etb Exécutant
	Région d'implantation de l'établissement
	Catégorie Etb Exécutant
	Discipline de Prestation Etb Exécutant
	Mode de Traitement Etb Exécutant
Dépenses	Coefficient Global de la Prestation Préfiltré
	Dénombrement de la Prestation Préfiltré
	Quantité de la Prestation Préfiltrée
	Montant du Dépassement de la Prestation Préfiltré
	Montant de la Dépense de la Prestation Préfiltrée
	Montant Versé/Remboursé Préfiltré
	Coefficient Global
	Dénombrement
	Quantité
	Montant du Dépassement
	Montant de la Dépense
	Montant Versé/Remboursé
	Base de Remboursement

FIGURE 1.32 – Les variables de la base DAMIR

Extraction des données sur la pharmacie

Variables et Actes retenus

Comme vu précédemment, l'Open Data DAMIR contient toutes les prestations versées par l'Assurance Maladie. Ainsi, un premier travail d'isolement des dépenses entrant dans le périmètre de la pharmacie a dû être réalisé avec pour objectif la sélection des variables pertinentes et l'extraction des lignes de prestations relatives au poste de soin pharmaceutique.
Pour le choix des variables et parmi les 55 disponibles dans la base DAMIR, 13 seront conservées dans le cadre de notre étude.

Variables Qualitatives	Variables Quantitatives
Année de Soins	Montant de la Dépense
Mois de Soins	Montant Versé/Remboursé
Sexe du Bénéficiaire	Montant du Dépassement
Tranche d'Age Bénéficiaire au moment des soins	
Région de résidence du Bénéficiare	
Top Bénéficiaire CMU-C	
Nature de Prestation	
Taux de Remboursement	
Catégorie du Prescripteur	
Catégorie de l' Exécutant	

FIGURE 1.33 – Les variables de l'étude

Parmi les variables retenues l'une d'elle est la nature de prestation. Elle donne l'information sur le type d'acte réalisé. Il existe environ 800 types d'actes. Chaque acte fait intervenir les 3 acteurs suivants :

- Le patient : appelé aussi bénéficiaire de l'acte. Il est connu par son appartenance à une tranche d'âge, par son sexe, son lieu de résidence (région de résidence) et s'il est bénéficiaire ou non de la CMU.
- Le professionnel de santé : classé en exécutant et prescripteur. Le professionnel de santé est décrit par sa catégorie, sa spécialité, son statut juridique et sa localisation.
- L'organisme d'assurance maladie (OAM)

Les filtres nécessaires à l'extraction des données de la pharmacie ont été déterminés à l'aide du fichier explicatif de toutes les variables de la base. Ainsi, nous y trouvons un onglet spécifique à la variable acte médical ou encore appelé nature de la prestation.

Période d'observation retenue

Notre étude nous impose d'avoir un jeu de données avec une profondeur dans le temps suffisante. Toutefois, pour des questions de volumétrie de données et de temps de traitement souvent long, nous avons été confrontés à trouver un équilibre entre ces deux contraintes. L'étude portera donc sur les données de 2014 à 2019. L'année 2020 étant une année atypique en raison de la crise sanitaire, il n'est donc pas pertinent de la conserver. L'année 2015 a également écarté de l'étude. En effet, nous avons constaté des volumes de prestations très bas.

Traitement des données

Le traitement des données s'est réalisé en plusieurs étapes. Chaque étape occasion des temps de traitements long. En pratique, cela donne lieu à plusieurs étapes :

- **Etape 1 : Site Ameli.fr (AMELI 2020)**
  Récupération des fichiers mensuels de la base Open DAMIR en .csv sur le site Ameli.fr.

— **Etape 2 : Téléchargements**
Les fichiers téléchargés sont zippés : il faut par la suite les extraire.
— **Etape 3 : Import des fichiers sous SAS**
L'import des fichiers .csv vers la librairie SAS ne peut se faire via la procédure "proc import" sous SAS. La solution de contournement est la connexion au GRID depuis FileZilla.
— **Etape 4 : Sélection des variables et actes**
A partir des fichiers mensuels, sélection des variables retenues et filtre sur les actes liés à la pharmacie.
— **Etape 5 : Concaténation**
Nous avons agrégé par année les bases mensuelles propres à la pharmacie. Une deuxième concaténation des bases annuelles permet d'obtenir une base globale des 7 années d'étude. La base DAMIR ne peut contenir de lignes identiques, puisqu'il s'agit d'une base agrégée selon les variables d'origine. Cependant, à ce niveau de traitement, nombreuses sont celles qui ont été supprimées. Il est donc possible d'agréger à nouveau chacune des 6 bases en tenant compte des variables restantes.

Les données manquantes :

La base obtenue contient sur certaines lignes une ou plusieurs valeurs manquantes sur les caractéristiques des bénéficiaires. Ces dernières concernent l'âge, le sexe et la région du bénéficiaire.
Cela concerne 6% de lignes sur le nombre total de lignes propres à la pharmacie et 5% en volume de prestations totales sur la pharmacie. La variable région est la caractéristique qui ressort le plus en donnée manquante suivie par la variable âge et enfin la variable sexe. Enfin l'analyse se poursuit par le croisement de chaque variable à valeurs manquantes avec les autres variables. Nous constatons qu'il s'agit de variables absentes aléatoirement.

	Age	Genre	Région	% en nombre de lignes sur le nombre de lignes totales pharmacie	% de prestations (FR) sur les prestations totales pharmacie
3 variables manquantes	Absente	Absente	Absente	0,00%	0,00%
2 variables manquantes	Absente	Absente	Présente	0,00%	0,00%
2 variables manquantes	Absente	Présente	Absente	0,01%	0,02%
2 variables manquantes	Présente	Absente	Absente	0,00%	0,00%
1 variable manquante	Absente	Présente	Présente	0,20%	0,07%
1 variable manquante	Présente	Absente	Présente	0,00%	0,00%
1 variable manquante	Présente	Présente	Absente	5,79%	4,74%
0 variable manquante	Présente	Présente	Présente	94,21%	95,26%

FIGURE 1.34 – Données manquantes par variables en % de lignes et % de volume de prestations sur la pharmacie

Pour la suite des travaux et malgré une volumétrie faible de données manquantes, nous avons décidé de les corriger.
Pour cela nous avons utilisé Le package MICE (Multivariate Imputation by Chained Equations) du logiciel R. Ce package implémente une méthode pour traiter les données manquantes à condition qu'elles soient manquantes de manière aléatoire. A chaque variable est associée un modèle d'imputation, conditionnellement aux autres variables du jeu de données.

Statistiques descriptives

Vue d'ensemble

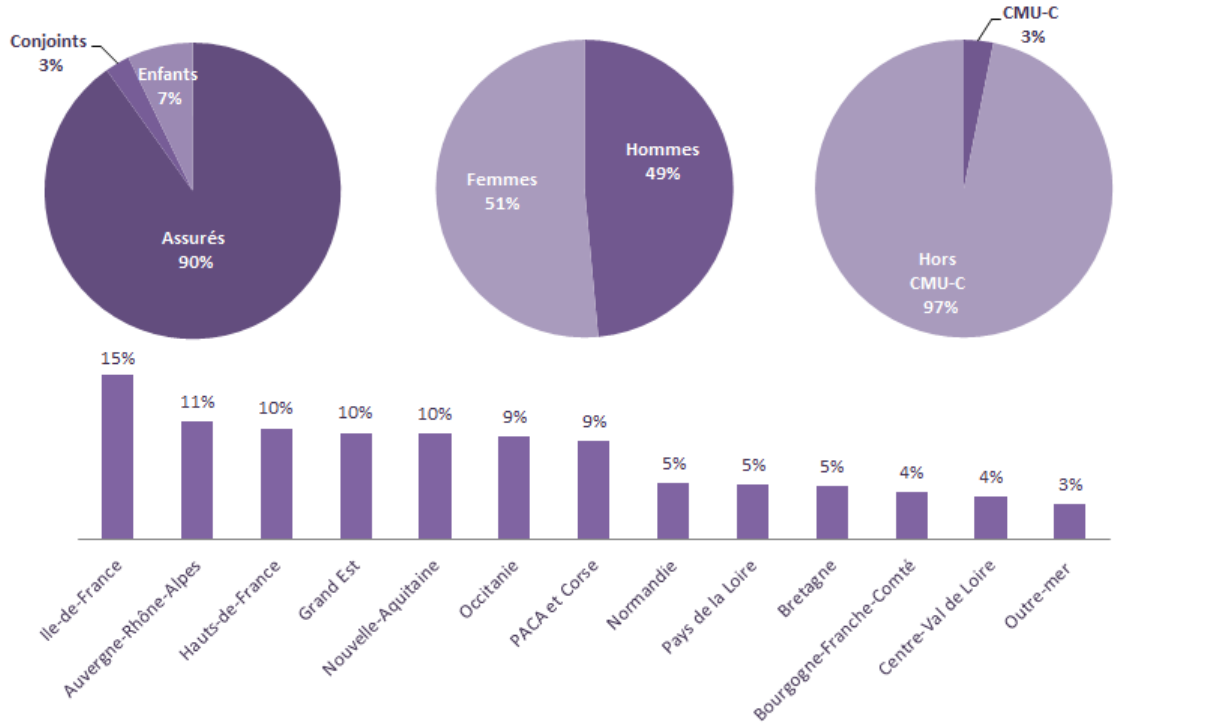

FIGURE 1.35 – De gauche à droite ; de haut en bas ; Répartition des dépenses en pharmacie par bénéficiaires ; genre ; Statut CMU-C ; région

1.3.3 Comparaison et mise en commun

L'objectif de cette partie est de définir notre périmètre d'étude.
La base DAMIR étant une base nationale, elle représente donc la consommation en soins de l'ensemble des Français.
Rappelons que cette étude a pour objet de mesurer la dérive de sinistralité sur le portefeuille d'un assureur. Pour utiliser les données DAMIR nous devons donc les mettre au format, avec les mêmes caractéristiques que notre base assureur.

Dans la partie précédente, nous avons présenté la phase d'extraction des données issues de la base DAMIR sur la pharmacie. A partie de cette base nous avons conservé les lignes dont le reste à charge après remboursement de la Sécurité Sociale soit supérieur à 0. Nous avons également exclu de notre périmètre les personnes bénéficiant de la CMU-C. Enfin, nous avons procéder à une segmentation basée sur les caractéristiques de notre portefeuille assureur avec la constitution de 3 segments :

— segment âges,

— Segment sexes,
— Segment régions.

Nous avons ensuite pioché au sein de ces segments pour constituer notre périmètre d'étude à l'image de notre portefeuille assureur.

Chapitre 2

Cadre théorique des modèles utilisés

Au chapitre précèdent, nous avons décrit la notion de dérive en santé et présenté une méthode intuitive basée sur l'observation de la dérive d'année en année.

Dans ce chapitre, nous nous proposons d'étudier une méthode de projection via les séries temporelles afin d'identifier dans quelle mesure ces méthodes peuvent être un complément d'analyse aux méthodes actuelles.
Dans un premier temps, nous présenterons le cadre théorique d'étude des séries temporelles univariées, puis nous décrirons et analyserons les données mises à notre disposition ainsi que les différentes méthodes de projections en séries temporelles applicables à ces dernières. Enfin nous choisirons les différents modèles qui s'ajustent au mieux à notre problème.

2.1 Introduction

Cette partie a pour objectif de présenter le cadre théorique des séries temporelles, puis nous détaillerons les différents modèles de projections utilisés afin de projeter la dérive de consommation de santé sur notre portefeuille d'étude.
Les séries temporelles recouvrent un large éventail de phénomènes de la vie réelle et se retrouvent dans de nombreux domaine. Elles ont été pendant longtemps étudiées en économétrie et depuis quelques années leur domaine d'application tend à s'élargir.

2.1.1 Définitions

Une **série temporelle**, est une collection finie de réalisations d'observations $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ représentant l'évolution d'une quantité au cours du temps, n représente l'étendue de la série temporelle.
Les observations peuvent être collectées à intervalle de temps t régulier ou non. Dans ce mémoire nous étudierons des séries temporelles régulières collectées sur un pas de temps mensuel.

Un **processus stochastique** $X_t, t \in T$, est une famille de variables aléatoires X_t définies sur le même espace de probabilité.
Lorsque les observations de la série temporelle étudiées sont des réalisations de variables aléatoires, on modélise cette série comme des processus stochastiques.
Une série temporelle de données peut être modélisée mathématiquement afin d'analyser, d'étudier son comportement, pour comprendre son évolution passée et prédire son comportement futur à partir des valeurs passées observées de cette même variable.
La série temporelle $X_t, t \in T$ d'un processus stochastique est donc composée d'une partie déterministe

et une partie aléatoire.

La composante déterministe :

— Une **série de tendance (T_t)** qui correspond à une évolution moyenne à long terme de la série, elle peut être déterministe ou stochastique.
La tendance est déterministe si T_t est une fonction déterministe de type :
  — Linéaire : $T_t = a + bt$
  — Quadratique : $T_t = a + bt + ct^2$
  — Logarithmique : $T_t = log(t)$

— **Une saisonnalité (S_t)** qui correspond à la présence d'un phénomène périodique qui se répète au fil du temps. Si la série (S_t) présente une saisonnalité de période d, il existe d tel que $S_{t+d} = S_t$

La composante aléatoire :

Une composante aléatoire (ϵ_t) qui correspond au résidu du modèle. Elle est l'expression de la partie de la série temporelle que la décomposition ne permet pas d'expliquer.

Structures de modèle

On peut décomposer une série temporelle à partir de la tendance, la saisonnalité et le résidu en trois types de structure de modèle :
— Une structure dite additive, dans ce cas le modèle s'écrit :

$$X_t = T_t + S_t + \epsilon_t,$$

— Une structure dite totalement multiplicative

$$X_t = (T_t + S_t) \times \epsilon_t,$$

— Une structure mixte, à la fois multiplicative et additive :

$$X_t = (T_t \times S_t) + \epsilon_t,$$

2.1.2 Définitions et propriétés importantes des séries temporelles

Le bruit blanc

Un **bruit blanc fort**, $(\epsilon_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires indépendantes, identiquement distribuées (iid) de moyenne nulle et de variance finie.
$\forall t \in \mathbb{Z}$, : $\mathbb{E}(\epsilon_t) = 0$, et $\mathbb{E}(\epsilon_t^2) = \sigma^2$.

Un **bruit blanc faible**, $(\epsilon_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires non-corrélées, identiquement distribuées de moyenne nulle et de variance finie.
$\forall (t, t') \in \mathbb{Z}^2, t \neq t' : \mathbb{E}(\epsilon_t) = 0$, $\mathrm{Cov}(\epsilon_t, \epsilon_{t'}) = 0$ et $\mathbb{E}(\epsilon_t^2) = \sigma^2$.

Remarque : les variables aléatoires d'un bruit blanc faible ne sont pas nécessairement indépendantes. On note que dans un bruit blanc, il n'y a pas de dépendance temporelle.
L'objectif de la modélisation des séries temporelles est d'obtenir un résidu de type bruit blanc, qui ne contient plus d'information temporelle, soit un signal stationnaire, décorrélé et aléatoire.

La stationnarité

Une série $\{X_t\}$, est **stationnaire au sens fort** si la distribution conjointe de $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ est identique à $(X_{t_{t+1}}, X_{t_{t+2}}, \ldots, X_{t_{t+k}})$ quel que soit k le nombre d'instants considérés, $(t_1, t_2, \ldots, t_k)$ les instants choisis et t le décalage.

Autrement dit si la structure d'un processus stochastique évolue avec le temps, le processus n'est pas stationnaire.
L'étude de la stationnarité d'une série est un pré-requis à la modélisation des séries temporelles car les démarches de modélisation qui en découlent diffèrent qu'on soit en présence de stationnarité ou non.
En pratique, la stationnarité forte est difficile à vérifier, on lui préfère la stationnarité faible.

Une série X_t, est **stationnaire au sens faible** si :
— $E(X_t) = \mu$, constante indépendante de t
— $Var(X_t) = \sigma^2 \neq \infty$, constante indépendant de t
— $\gamma(h) = Cov(X_t, X_{t+h})$, ne dépend que de h mais pas de t.

L'espérance étant constante, la série ne doit donc pas présenter de tendance.
Le processus doit être d'ordre 2 c'est à dire de variance finie.
L'auto-covariance entre 2 éléments de la série ne dépend que de l'ampleur du décalage h et non de la position dans le temps.
Dans le cas où $h = 0$, l'autocovariance est égale à la variance qui est nulle.

$(X_t), t \in Z$, est un **processus non-stationnaire TS (Trend Stationary)** s'il peut s' ecrire sous la forme :

$$X_t = f(t) + Y_t,$$

où f est une fonction déterministe et $(Y_t), t \in Z$ un processus stationnaire.

$(X_t), t \in Z$ est un **processus non-stationnaire DS (Difference Stationary)** s'il est stationnaire après différenciation :

$$\Delta dX_t = (I - B)dX_t,$$

où

$$BX_t = X_{t-1}.$$

Si d = 2, cela signifie que le processus défini pour tout $t \in Z$ par

$$\Delta 2X_t = X_t - 2X_{t-1} + X_{t-2},$$

est stationnaire.

Ces processus peuvent néanmoins être rendus stationnaires par une suite de transformations usuelles (désaisonnalisation, différenciation, transformation non linéaire...). Une méthodologie classique est celle de Box-Jenkins

Autocovariance et autocorrelation

On appelle fonction d'auto-correlation d'une série chronologique la fonction $h \mapsto \rho(h) = \frac{\gamma(h)}{\gamma(0)}$.

Elle correspond au fait que, dans une série temporelle, la valeur prise par la série à un instant t peut être corrélée aux mesures précédentes (au temps t-1, t-2, t-3, ...) ou aux mesures suivantes (à t+1, t+2, t+3, ...). Une série autocorrélée est donc corrélée à elle-même, avec un décalage donné.
Dans la pratique, on utilise la fonction d'auto-correlation empirique.
Puis, on calcule les auto-corrélations pour différent décalage k. Ce qui permet de construire un auto-corrélogramme empirique, qui permet d'apprécier visuellement le décalage avec le plus de corrélations.

L'auto-corrélation partielle

Dans la suite de l'étude, on analyse l'auto-corrélation partielle.
On appelle fonction d'auto-corrélation partielle, l'auto-corrélation calculée pour un décalage "k" fixé indépendamment des autres décalages. De même, on peut construire un auto-corrélogramme empirique partiel.

L'opérateur retard

Afin de faciliter l'écriture des séries temporelles on a coutume d'exprimer l'évolution d'une série en fonction de son passé. On utilise l'opérateur retard noté L ou B qui lie tout élément de la série temporelle à son observation précédente : $X_t = BX_{t-1}$.
De même, cet opérateur retard peut s'étendre aux décalages > 1
On écrira par exemple $X_{t-k} = B^k X_t$.
Afin de réaliser cet objectif, une première étape de modélisation de la série est nécessaire. Cette étape consiste à sélectionner, parmi une famille de modèles correspondant à des approximations de la réalité, celui qui décrit le mieux la série en question.

Dans ce mémoire nous étudierons les modèles de séries temporelles suivants :
— les lissages exponentiels,
— les modèles du type ARIMA,
— les modèles de régression (régression linéaire, modèles non-paramétriques. . .).

2.2 Les lissages exponentiels

2.2.1 Principe du lissage exponentiel

Introduit par Holt (1957) et par Brown (1962), le lissage exponentiel est un modèle de prévision qui consiste à ajuster à une chronique de série temporelle, une estimation locale de ce que va être sa valeur future. On va utiliser les valeurs passées de la série chronologique pour prédire le futur. En effet, pour un horizon $h > 0$, on souhaite trouver X_{t+h} à partir des réalisation $X_1, \ldots, X_t$. On cherche donc :

$$X_{t+h|t} = E[X_{t+h}|X_1, \ldots, X_t].$$

Les modèles de lissage exponentiel prévoient plusieurs type de modélisations de $X_{t+h|t}$. On note $\epsilon_t = X_t - X_t|_{t-1}$, l'erreur de prévision à l'instant $t-1$.
En fonction des caractéristiques de la série temporelle à modéliser on distingue plusieurs types de lissage exponentiel :
— Le lissage exponentiel simple : pour des séries sans tendance ni saisonnalité.
— Le lissage exponentiel double : pour les séries qui ont une tendance linéaire.
— Le lissage de Holt-Winters : pour les séries qui ont à la fois une tendance et une saisonnalité.

— Enfin on peut aussi faire des modèles ayant une composante saisonnière mais sans aucune tendance distincte.

L'objectif de la modélisation est d'estimer les paramètres propres à chaque équation. Il est possible de sélectionner un modèle à partir d'un critère de type AIC, AICc ou BIC.

2.2.2 Lissage exponentiel simple

La série (X_t) peut être modélisée par un lissage exponentiel simple si elle peut s'écrire de manière suivante : $X_t = l_{t-1} + Z_t$, avec Z_t un bruit blanc gaussien et l_{t-1} le niveau de connaissance à l'instant t.

D'après le cours de Angelina ROCHE 2018-2019 sur les lissages exponentiels, on a

$$\hat{X}_{t+1|t} = \hat{X}_{t|t-1} + \alpha(X_t - \hat{X}_{t|t-1}),$$

avec $0<\alpha<1$

$$\hat{X}_{t+1|t} = \alpha X_t + (1-\alpha)\hat{X}_{t|t-1},$$
$$\hat{X}_{t+1|t} = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \ldots + (1-\alpha)^t \hat{X}_{1|0},$$

La solution est exprimée par :
$\hat{X}_{t+h|t} = \hat{l}_t$, avec $\hat{l}_t = \alpha X_t + (1-\alpha)\hat{l}_{t-1}$.
pour tout h>0, Le paramètre α est estimé par maximisation de la vraisemblance.

2.2.3 Lissage exponentiel double (méthode de Holt)

En 1957, Holt a étendu la notion de lissage exponentiel simple au cas du lissage exponentiel linéaire. L'innovation apportée est la possibilité d'estimer localement la série par une droite.
La série (X_t) peut être modélisée par un lissage exponentiel double si la prédiction de X_{t+h} sachant $X_1, \ldots, X_t$ peut s'écrire de manière suivante :

$$\hat{X}_{t+h|t} = l_t + hT_t$$

avec,
— T_t la pente définie par : $T_t = \beta^* \times (l_t - l_{t-1}) + (1-\beta^*) \times T_{t-1} = T_{t-1} + \alpha\beta e_t$, avec $\beta = \alpha\beta^*$;
— l_t le niveau défini par : $l_t = \alpha X_t + (1-\alpha)(l_{t-1} + T_{t-1}) = l_{t-1} + T_{t-1} + \alpha e_t$.

La solution est donnée par :

$$(\hat{l}, \hat{T}) = argmin_{l,T} \sum_{i=0}^{\infty} (1-\alpha)^i (X_{t-i} - (l - T_i))^2.$$

2.2.4 Lissage exponentiel triple (Méthode de Holt-Winters)

Le lissage exponentiel triple est une généralisation du lissage double, qui permet entre autres de proposer les modèles suivants :

— Tendance linéaire locale,
— Tendance linéaire locale + saisonnalité (cas du modèle additif),
— Tendance linéaire locale×saisonnalité (cas du modèle multiplicatif).

Dans ce cas, deux paramètres de lissage entrent en jeu et on ajuste au voisinage de t une fonction linéaire $l_t + hT_t$, h étant l'horizon de prévision.
La série (X_t) peut être modélisée par un lissage exponentiel triple si la prédiction de X_{t+h} sachant $X_1, \ldots, X_t$ peut s'écrire de manière suivante : $X_{t|t-1} = l_{t-1} + T_{t-1} + S_{t-m}$ avec :
pour $\alpha, \beta, \gamma \in]0, 1[$,
— l_t le niveau égal à $l_t = \alpha(X_t - S_{t-m}) + (1 - \alpha)(l_{t-1} + T_{t-1})$.
— T_t la pente égale à $T_t = \beta(l_t - l_{t-1}) + (1 - \beta^*)T_{t-1}$.
— S_t la saisonnalité égale à $S_t = \gamma(X_t - l_{t-1} - T_{t-1}) + (1 - \gamma)S_{t-m}$.

Les paramètres d'estimation (σ , α, β, γ) peuvent être estimés par maximisation de la vraisemblance d'un modèle espace-état [1]. Les intervalles de prédiction sont dérivés à partir de ces modèles.

2.3 Les modèles ARMA

Les modèles ARMA (Auto Regressive Moving Average) apportent une approche complémentaire dans l'analyse et la prévision de séries chronologiques. Nous avons vu précédemment que le lissage exponentiel triple était basé sur une description de la tendance et de la saisonnalité au sein des données, les modèles ARMA visent quant à eux à décrire les autocorrélations entre les observations. Ces modèles capturent donc une suite de différentes structures temporelles.

2.3.1 Le processus autorégressif (AR)

Une série temporelle est un processus autorégressif d'ordre p si elle peut s'écrire de la manière suivante :

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t,$$

Avec $\epsilon_t \sim BB(0, \sigma^2)$ et $\phi_p \neq 0$

Il s'agit d'une régression de la variable par rapport à elle-même, d'où le terme d'auto-regression.
Un processus X_t est AR(p) s'écrit comme combinaison linéaire de ces p valeurs précédentes.
En utilisant l'opérateur retard définit au paragraphe précédent, le terme (1) devient :

$$c + \epsilon_t = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t.$$

On appelle

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$

l'opérateur d'auto-regression.

2.3.2 Le processus Moyenne Mobile (MA)

Une série temporelle est un processus moyenne mobile d'ordre q s'il peut s'écrire comme une combinaison linéaire de q bruits blancs.
On note d'après le cours de Angelina' ROCHE 2018-2019 :

1. Forecasting with Exponential Smoothing - The State Space Approach

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q},$$

Avec $\theta_q \neq 0$
En utilisant l'opérateur retard, on note :

$$X_t = \mu + (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q) \times \epsilon_t.$$

Soit

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q,$$

qui est appelé l'opérateur moyenne mobile.

Remarque : Un processus MA(q) est toujours stationnaire car combinaison linéaire de bruits blancs.

2.3.3 Le processus autorégressif et moyenne mobile (ARMA (p,q))

Une série temporelle est un processus ARMA (p,q) s'il est stationnaire et peut s'écrire comme combinaison d'un processus AR(p) et MA (q)
avec p $\geq$ 0, q $\geq$ 0.
On note :

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q},$$

avec $\epsilon_t \sim BB(0, \sigma^2)$

On démontre que :

$$X_t = \mu + \frac{\Theta(B)}{\Phi(B)} \times \epsilon_t.$$

2.3.4 Le modèle ARMA saisonnier (SARMA)

X_t est un processus SARMA(p,q) $\times$ (P,Q), de période s, s'il est un processus stationnaire de la forme :

$$\Phi(B)\Gamma(B^s)X_t = \Theta(B)\Omega(B^s) \times \epsilon_t,$$

avec $\epsilon_t \sim BB(0, \sigma^2)$

$$\Theta(B) = 1 - \nu_1 B - \nu_2 B^2 - \ldots - \nu_p B^p,$$

et

$$\Omega(B^s) = 1 + \omega_1 B + \omega_2 B^2 + \ldots + \omega_q B^q,$$

l'identification des modèles SARMA(p,q) $\times$ (P,Q) nécessite de rechercher toutes les combinaisons possibles (p,q) $\times$ (P,Q). Ceci peut être extrêmement long en temps de calcul. Une solution possible est l'utilisation des algorithmes génétiques.

2.3.5 Les processus intégrés ARIMA et SARIMA

Dans la pratique, la plupart des séries ne sont pas stationnaires. Dans certains cas, on arrive à transformer les séries non stationnaires en séries stationnaires par différenciation simple ou saisonnière. Les processus sont dits "intégrés".

Les Processus ARIMA et SARIMA sont des processus qui deviennent des processus ARMA et SARMA après une différentiation simple ou saisonnière. Un processus Xt est un SARIMA(p,d,q)(P,D,Q) s'il s'écrit d'après le cours de Angelina" ROCHE 2018-2019 :

$$\Phi_s(B^s)\Phi(B)(1-B)^d(1-B^s)^D X_t = \Theta(B)\Theta_s(Bs) \times \epsilon_t,$$

Le terme $(1-B)^d$ représente la différentiation simple et $(1-B^s)^D$ la différenciation saisonnière.

Durant la phase de modélisation, nous allons être confrontés à plusieurs défis qui seront de :
— définir un modèle avec un nombre fini de paramètres,
— estimer les paramètres de ce modèle,
— vérifier la qualité d'ajustement du modèle, comparer différents modèles,
— effectuer des prédictions.

2.3.6 Démarche de Box-Jenkins :

Les étapes de la démarche de Box-Jenkins sont résumées dans la figure ci-dessous :

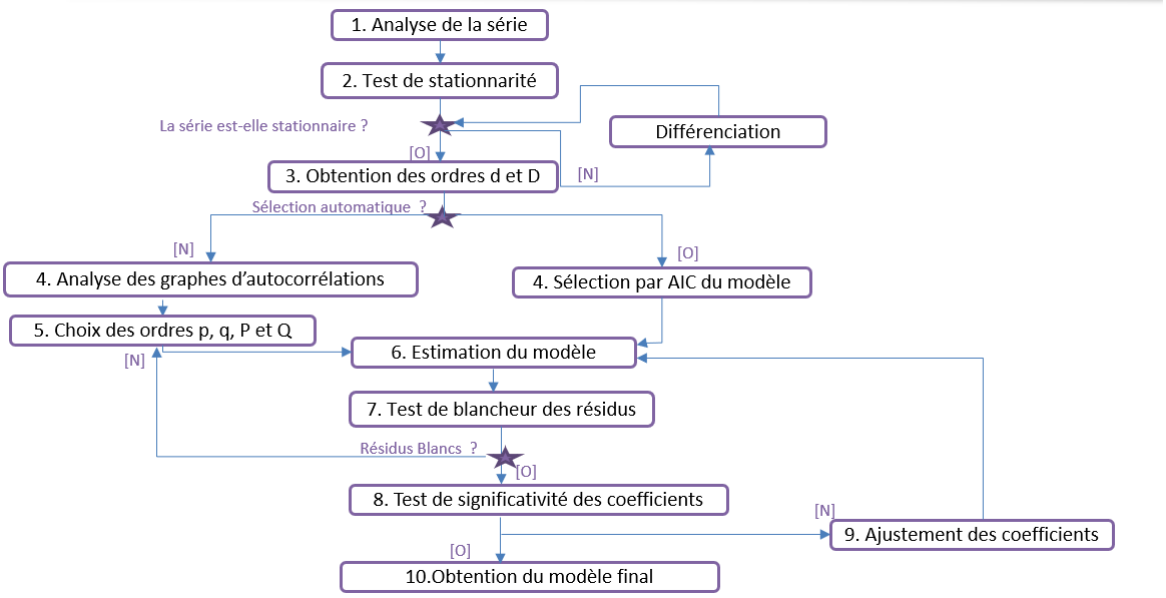

FIGURE 2.1 – Les différentes étapes de la démarche de Box-Jenkins

Tests sur les résidus

— Tests d'autocorrélation :
Il s'agit d'un test visuel basé sur l'analyse de l'auto-corrélogramme.
L'hypothèse nulle (H0) : ϵ_t est un bruit blanc
et l'hypothèse alternative (H1) : ϵ_t nest pas un bruit blanc.
On sait que pour un ordre h fixé, si chacune des autorrélations empiriques
$\hat{\rho(h)} \notin [\frac{-1,96}{\sqrt{n}}; \frac{1,96}{\sqrt{n}}]$
pour un seuil de 5% par exemple, donc si les $\hat{\rho(h)}$ sont à l'extérieur de la bande de décision alors l'autocorrelation théorique $\rho(h)$ est non nulle et la série n'est pas un bruit blanc.

— Tests autocorrélation du portemanteau :
Les tests d'autocorrélation du portemanteau testent l'indépendance asymptotique des variables d'autocorrélations empiriques.
On pose comme hypothèse nulle (H0) : $\rho_1, \rho_2, \ldots, \rho_h = 0$
et l'hypothèse alternative (H1) : Au moins un des $\rho_1, \rho_2, \ldots, \rho_h$ est nul.

Deux statistiques sont souvent considérées :

— Statistique de Box-Pierce (1970) :

$$Q - BP(h) = T \sum_{k=1}^{h} \hat{\rho}^2 - k,$$

— Statistique de Ljung-Box (1978) :

$$Q^* = T(T+2) \sum_{k=1}^{h} (T-k)^{-1} \hat{\rho}^2 - k.$$

Avec :
h est le décalage maximal considéré et T est le nombre d'observations. La statistique de Ljung-Box est plus précise pour les petits échantillons.
Nous utiliserons exclusivement cette dernière dans notre démarche de modélisation.

— Tests de normalité :

Afin de pouvoir rajouter des intervalles de confiance autour des prévisions, on peut tester la normalité des résidus.
On pose comme hypothèse nulle (H0) : ϵ_t est gaussien
et l'hypothèse alternative (H1) : ϵ_t nest pas gaussien.

Dans le cadre de ce mémoire, pour tester la normalité des résidus, nous allons analyser le QQ-plot et effectuer un test de Shapiro-Wilk.
La représentation graphique du nuage de points formé par les quantiles théoriques de la loi N(0,1) et par les quantiles empiriques de la série temporelle s'ajustent à la droite d'équation
$y = \sigma x + \mu$ si $\epsilon_t \sim \mathcal{N}(\mu, \sigma)$ sous (H0).
La statistique de test de Shapiro-Wilk est le coefficient de détermination du nuage de point.

2.4 Les modèles de régression

La modélisation représente le comportement d'une grandeur naturelle par une expression comportant une partie déterministe (une fonction) et une partie aléatoire. La partie déterministe est ce qui permet de décrire le comportement de la moyenne du phénomène (le comportement moyen). La partie aléatoire est le différentiel entre la vraie valeur de la variable étudiée et la partie déterministe. La modélisation se « fabrique » donc sur deux plans : déterministe, en essayant d'ajuster une forme mathématique à la variation « en moyenne » du phénomène, par des méthodes que nous verrons plus loin ; aléatoire, en donnant une forme à la variabilité du phénomène autour de sa moyenne, en d'autres termes, en donnant une forme au hasard.

Nous introduisons dans cette section un rappel des modèles linéaires, puis nous présenterons le modèle linéaire généralisé qui est sans nul doute l'outil le plus général, le plus utile et, par conséquent, le plus utilisé de la panoplie des instruments dévolus à la modélisation. Enfin, et dans le cadre de notre étude, le modèle qui nous intéressera sera le modèle de regression sur des données longitudinales.

2.4.1 La régression linéaire multiple

Présentation du modèle

Dans bons nombres d'études, on souhaite prédire et/ou expliquer les valeurs d'une variable quantitative Y à partir des valeurs de p variables $X_1, X_2, \ldots, X_p$.
On dit alors que l'on souhaite "expliquer Y à partir de $X_1, X_2, \ldots, X_p$", Y est appelée "variable à expliquer" et $X_1, X_2, \ldots, X_p$ sont appelées "variables explicatives".
Si une liaison linéaire entre Y et $X_1, X_2, \ldots, X_p$ est envisageable, on peut utiliser le modèle de régression linéaire multiple. Sa forme générique est :

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,$$

où $\beta_0, \ldots, \beta_p$ sont des coefficients réels inconnus et ϵ est une variable quantitative de valeur moyenne nulle, indépendante de $X_1, X_2, \ldots, X_p$, qui représente une somme d'erreurs aléatoires et multifactorielles.
Comme $\beta_0, \ldots, \beta_p$ caractérisent le lien existant entre Y et $X_1, X_2, \ldots, X_p$, on souhaite les estimer à l'aide des données. Soit par la méthode des moindres carrés ou par l'estimation du maximum de vraisemblance dans le cas gaussien.

L'écriture matricielle du modèle linéaire est la suivante :

$$Y = X\beta + \epsilon,$$

où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{1,1} & \cdots x_{p,1} \\ 1 & x_{1,2} & \cdots x_{p,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \cdots x_{p,n} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

Les hypothèses

Les hypothèses standards sur le modèle de régression linéaire sont :

X est de rang plein (donc $(\mathrm{X}^t X)^{-1} existe$),
ϵ et $X_1, X_2, \ldots, X_p$ sont indépendantes et $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$ où $\sigma > 0$ est un paramètre inconnu
En particulier, cette dernière hypothèse entraîne que

— ϵ est un vecteur gaussien ;
— $\mathbb{E}[\epsilon] = 0_n$;
— le modèle est homoscédastique : $\mathbb{V}[\epsilon_j] = \sigma^2 \ \forall j = 1, \ldots, n$;
— $\epsilon_i \perp\!\!\!\perp \epsilon_j \ \forall i \neq j$.

L'estimateur de β

L'estimateur des **moindres carrés ordinaires de** β est :

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

Il est construit de sorte que l'erreur d'estimation entre $\mathrm{X}\hat{\beta}$ et Y soit le plus petite possible au sens $\|.\|^2$:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2,$$

où $\|.\|$ désigne la norme euclidienne de R^n :

$$< a, b >= a^t b = b^t a = \sum_{i=1}^{n} a_i b_i,$$

avec $\|a\|^2 =< a, a >= a^t a = \sum_{i=1}^{n} a_i^2$,

Pour tout $j \in \{0, ..., p\}$, la j+1-ème composante de $\hat{\beta}$, notée $\hat{\beta}_j$, est l'estimateur des moindres carrés ordinaires de β_j

L'estimateur du **maximum de vraisemblance** dans le cas gaussien est défini de la sorte :

$$\begin{aligned} L_Y(y, \beta) &= \prod_{i=1}^{n} f_Y(y_i, \beta) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y - X\beta\|^2}{(2\sigma^2)}\right), \end{aligned}$$

avec f_Y la densité gaussien de Y, $y \in \mathbb{R}^n$,

Maximiser cette valeur, revient à minimiser son logarithme qui vaut :

$$\underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmax}} L_Y(y,\beta) = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2 = \hat{\beta},$$

L'estimateur des moindres carrés ordinaires de β est l'estimateur du maximum de vraisemblance de β.

2.4.2 La régression linéaire généralisée

La régression linéaire généralisée est une extension de la régression linéaire multiple, qui permet de s'affranchir de certaines contraintes fortes. En effet, ce modèle n'impose pas de satisfaire les trois dernières hypothèses vues précédemment. Il fait intervenir une fonction de lien qui va supprimer la linéarité qui était présente entre la variable à expliquer et les variables explicatives. Tout ceci est expliqué dans le cours de Charpentier 2013.

La genèse

La théorie des modèles linéaires généralisés a été formulée par John Nelder et Robert Wedderburn en 1972 puis de façon complète par McCullagh et Nelder en 1989 comme un moyen d'unifier les autres modèles statistiques y compris la régression linéaire. le modèle linéaire généralisé souvent connu sous les initiales anglaises GLM est une généralisation souple de la régression linéaire.
Les GLM permettent d'étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs $X_1, \ldots, X_p$.

Ces modèles englobent :
— le modèle linéaire général (régression multiple, analyse de la variance et analyse de la covariance)
— le modèle log-linéaire,
— la régression logistique,
— la régression de Poisson.

Présentation du modèle

Les modèles linéaires généralisés sont formés de trois composantes :

La composante aléatoire

C'est la variable de réponse Y, composante aléatoire à laquelle est associée une loi de probabilité appartenant à la famille exponentielle.
Notons $(Y_1, \ldots, Y_n)$ un échantillon aléatoire de taille n de la variable de réponse Y, les variables aléatoires $Y_1, \ldots, Y_n$ étant supposées indépendantes.

— Y_i peut être binaire (succès-échecs, présence-absence), dans ce cas il s'agit d'une loi de Bernoulli ou loi binomiale.

— Y_i peut être distribuée selon une loi de Poisson.

— Y_i peut être distribuée selon une loi normale.

La densité de la composante s'écrit sous la forme :

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

avec :

$\theta \in \mathrm{R}$ est le paramètre canonique ou de la moyenne,
$\phi \in \mathrm{R}$ est le paramètre de dispersion,
a fonction définie sur $\mathbb{R}$ non nulle,
b fonction définie sur $\mathbb{R}$ deux fois dérivables,

c fonction définie sur $\mathbb{R}^2$.

La composante déterministe

Les variables explicatives $X_1, \ldots, X_p$ utilisées comme prédicteurs dans le modèle définissent sous forme d'une combinaison linéaire la composante déterministe. Cette dernière exprimée sous forme d'une combinaison linéaire $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ (appelée aussi prédicteur linéaire) précise quels sont les prédicteurs.

La fonction lien

La fonction lien est une fonction g définie sur $\mathbb{R}$ strictement monotone et inversible.
Cette fonction décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_1, \ldots, X_p$ et l'espérance mathématique de la variable de réponse Y.
Elle spécifie comment l'espérance mathématique de Y notée μ est liée au prédicteur linéaire construit à partir des variables explicatives.
On peut modéliser l'espérance μ directement (régression linéaire usuelle) ou modéliser une fonction monotone $g(\mu)$ de l'espérance :

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

La fonction lien qui utilise le paramètre canonique dans la famille des modèles linéaires généralisés, est appelée la fonction de lien canonique.
A toute loi de probabilité de la composante aléatoire est associée une fonction spécifique de l'espérance appelée paramètre canonique.

Quelques exemples :

La fonction de lien $g(\mu)=\mu$ permet de modéliser l'espérance.
Les modèles utilisant cette fonction de lien ici lien identité ont une distribution normale.

La fonction de lien $g(\mu)=\log(\mu)$ permet de modéliser le logarithme de l'espérance.
Les modèles utilisant cette fonction de lien sont des modèles log-linéaires.

La fonction de lien $g(\mu)=\log(\frac{\mu}{1-\mu})$ modélise le logarithme du rapport des chances.
Elle est appelée logit et est adaptée au cas où μ est compris entre 0 et 1 (par exemple la probabilité de succès dans une loi binomiale)

Résolution du modèle

La résolution du modèle linéaire se fait par l'estimation des paramètres β_i du GLM. Pour cela, on utilise la méthode du maximum de vraisemblance.
La (log)-vraisemblance s'écrit dans le cas des modèles exponentiels :

$$logL(\theta_1, \ldots, \theta_n, \phi, y_i, \ldots, y_n) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right],$$

les paramètres β sont obtenus en dérivant cette fonction log-vraisemblance par rapport au paramètre β et d'écrire les conditions du premier ordre.

Si la fonction de lien utilisée n'est pas la fonction canonique, il n'existe pas de solution analytique pour résoudre le système final. La résolution de ces équations est donc effectuée numériquement, par des logiciels statistiques. L'algorithme de Newton-Raphson est le plus utilisé pour ce type de traitement.

2.4.3 La régression sur données longitudinales

Il est possible de modéliser un GLM dépendant du temps comme décrit dans le document Eilstein 2019
Les hypothèses du GLM , vues ci-dessus, se traduisent, dans le cas d'une variable indexée par le temps (processus), de la façon suivante :

La composante aléatoire Y_t est la variable expliquée au temps t (t = 1, 2,..., T), qui suit une loi de probabilité de la famille exponentielle comme vu plus haut.

La composante déterministe est représentée par les variables explicatives $X_{t_1}, X_{t_2}, \ldots, X_{t_p}$ où X_{t_j} est la valeur du facteur j au temps t.

$(\beta_1, \beta_2, \ldots \beta_j, \ldots \beta_p)$' est le vecteur des paramètres.

La fonction de lien g décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_{t_1}, X_{t_2}, \ldots, X_{t_p}$ et l'espérance mathématique de la variable de réponse Y_t.
μ_t est l'espérance de Y_t, au temps t, g(μ_t) est le prédicteur linéaire au temps t,

2.5 Critères de choix et validation de modèle

Sélection des variables

C'est une étape primordiale qu'est celle où l'on va déterminer les variables qui expliciteront le mieux la variable réponse Y. La sélection des variables pour la modélisation se fera en veillant à sélectionner des variables indépendantes les unes des autres. Ensuite, nous veillerons à sélectionner suffisamment de variables pertinentes pour rendre le modèle précis. Face à une infinité de modèle, il est possible d'effectuer une recherche exhaustive en explorant l'ensemble des combinaisons de variables et en choisissant celle qui optimise un critère défini.

Pour ce faire, il existe plusieurs algorithmes dont les plus courants :

— **La méthode ascendante ou forward** consiste en l'ajout à chaque étape de la variable qui conduit à l'optimisation du critère de choix. Si aucune variable ne permet l'optimisation du critère de choix ou que toutes les variables sont intégrées, alors on arrête.

— **La méthode descendante ou backward** consiste à la suppression à chaque étape de la variable qui conduit à l'optimisation du critère de choix. Si aucun retrait ne permet d'optimiser le critère ou que toutes les variables ont été retirées, alors on arrête.

— **La méthode progressive ou stepwise** consiste en une méthode de sélection ascendante dans laquelle on intercale entre deux étapes d'élimination (ceci permet d'éliminer des variables introduites en amont et n'ayant plus d'effet après introduction de nouvelles variables).

Pour tester la significativité d'une variable, en d'autres termes sa pertinence pour le modèle, il existe des critères de sélection. Deux des plus généralement utilisés :

— **AIC : Akaike Information Criterion** défini de la manière suivante :

$$AIC = -2ln(L(\beta)) + 2N,$$

où :
N est le nombre de variables du modèle ;
L est la vraisemblance du modèle.

L'AIC prend en compte à la fois la qualité de l'ajustement via la fonction de vraisemblance (qui dépend des coefficients du GLM) et de la complexité du modèle (via du nombre de variables).

— **BIC : Bayesian Information Criterion** défini de la manière suivante :

$$BIC = -2ln(L(\beta)) + 2Nln(n),$$

où n est le nombre d'observations.
Néanmoins, le BIC n'est pas forcément adapté à des bases de données de petites tailles.
En effet, celui-ci a tendance à choisir des modèles trop simples du fait de sa forte pénalisation.

Ces deux critères doivent être minimisés. Un modèle sera considéré comme meilleur s'il présente un critère plus petit par rapport aux autres modèles testés.

Adéquation du modèle

Deux statistiques sont utiles pour juger de l'adéquation du modèle aux données :

— **La déviance (scaled deviance)** :

$$D = 2(ln(L_{sat}) - ln(L)),$$

où la L_{sat} correspond à la log-vraisemblance d'un modèle saturé, modèle construit pour lequel le nombre de variables est égal aux nombres d'observations.

— **La statistique du khi-deux de Pearson** :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(Y_i)},$$

Si ces critères sont supérieurs au quantile d'une loi du χ^2 à (n-p-1) degrés de libertés, alors le modèle sera jugé de mauvaise qualité.
Pour tester la significativité du modèle, le **Test de Wald** est généralement utilisé.
Cela consiste à tester les hypothèses suivantes :

$$\begin{cases} H_0 : \forall i, \beta_i = 0 \\ H_1 : \exists i, \beta_i \neq 0 \end{cases}$$

Le modèle est significatif si l'hypothèse H_0 est rejetée, autrement dit si la $p - value$ de la statistique de Wald définie sous $N(0, 1)$ par :

$$W = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)},$$

est inférieure au seuil de significativité choisi.

Mesure de la qualité de prédiction

Notons $\epsilon_t := X_t - \hat{X}_{t|t-1}$ le résidu d'estimation. Si nous observons $X_1, \ldots, X_n$, nous avons les indicateurs ci-dessous permettant de valider les modèles de prédictions :

— l'erreur moyenne :

$$ME = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t;$$

— l'erreur quadratique moyenne :

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2},$$

— l'erreur absolue moyenne :

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\epsilon_t|,$$

— l'erreur en pourcentage moyenne :

$$MPE = 100\frac{1}{n}\sum_{t=1}^{n} \frac{\epsilon_t}{X_t}$$

— l'erreur relative moyenne

$$MAPE = 100\frac{1}{n}\sum_{t=1}^{n} \frac{|\epsilon_t|}{|X_t|}$$

— erreur moyenne normalisée

$$MASE = \frac{n-1}{n}\frac{\sum_{t=1}^{n} \epsilon_t}{\sum_{u=2}^{n} |X_u - X_{u-1}|}$$

— autocorrélation des erreurs à l'horizon 1 (ACF1).

Chapitre 3

Modélisation de la dérive pour le poste pharmacie

Afin de modéliser la dérive des dépenses en pharmacie, nous avons fait le choix de modéliser la charge des dépenses et puis de déduire de cette charge estimée pour l'année à projeter la dérive applicable.
La dérive d'une année t par rapport à celle qui la précède se déduit par le ratio :

$$Derive_t = \left(\frac{RAC_{estim(t)}}{RAC_{t-1}} \right) - 1$$

Avec RAC = reste à charge Sécurité Sociale.
RAC_{est} = reste à charge estimé.

La prise en charge des médicaments dans les différentes formules et peu importe le niveau de garantie par l'assureur est de 100%, les assurés sont donc remboursés en totalité de leurs dépenses liées aux médicaments à partir du moment où il y a une prise en charge par la sécurité sociale.

3.1 Analyse de la série temporelle de données

Dans cette partie nous nous proposons d'étudier la série chronologique du reste à charge Sécurité Sociale des dépenses en pharmacie sur la base DAMIR qui a été retraitée, échantillonnée afin de s'ajuster aux données de l'assureur en ce qui concerne le profil des assurés, la démographie, la répartition des dépenses cf chapitre 1.
Cette base de données collectée mensuellement entre 2014 et 2019, nous donne des informations sur les montants payés par les assurés, les montants réglés par la Sécurité Sociale à la maille, sexe, âge, région, nature de la prestation.
Nous avons créé une nouvelle variable qui représente le reste à charge sécurité sociale défini comme suit :
RAC (SS) = Montant réel - montant remboursé par la SS.
Les études ci-dessous seront faites à partir d'une partie de cette base de données constituée de 3 variables : l'année, le mois, le montant de reste à charge.

Pour des fins de modélisation, cette nouvelle base de données est-elle même scindée en 2 :
— Une base d'apprentissage = données des RAC entre de Janvier 2014 à Décembre 2018, qui nous permettra de définir et entrainer nos modèles.
— Une base de validation (données de l'année 2019), qui nous permettra de tester et valider nos modèles.

3.1.1 Exploration de la série temporelle

Avant toute démarche de modélisation, observons et étudions les représentations graphiques des RAC des dépenses en pharmacie engagées par les assurés sur la période de 2014 à 2019.

Le chronogramme

Sur le chronogramme du reste à charge des dépenses en pharmacie ci-dessous (Figure 3.1) qui est la représentation graphique de la série au fil du temps, nous observons une tendance baissière qui n'est pas uniforme sur toute la série.
En effet, la série décroit très rapidement entre 2014 et 2015, puis sa décroissance est modérée de 2015 à 2019.

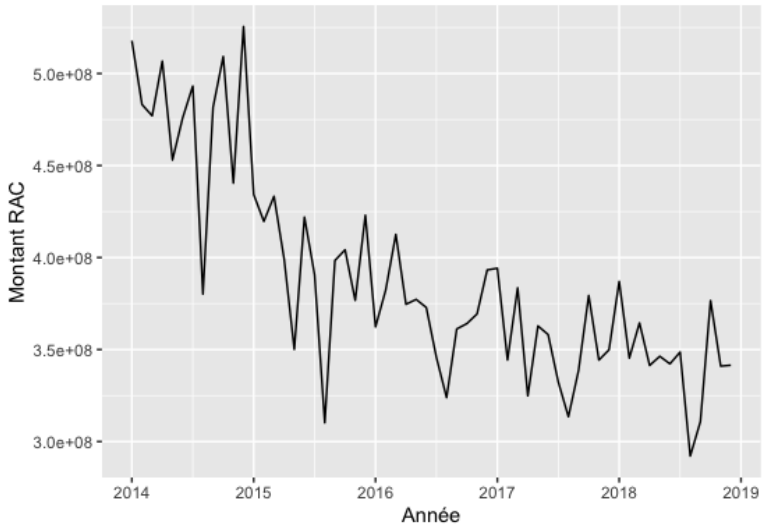

FIGURE 3.1 – Chronogramme du reste à charge

De même on observe des pics de consommation à la baisse au mois d'Août.
La variance de la série ne semble également pas uniforme, il y a une forte variabilité des RAC en début de période et le montant moyen se stabilise ensuite.

Le Month-Plot

Le Month-Plot ou Season-Plot est la représentation graphique simultanée pour chaque mois de la série des RAC des dépenses en pharmacie engagées par les assurés (Figure 3.2).

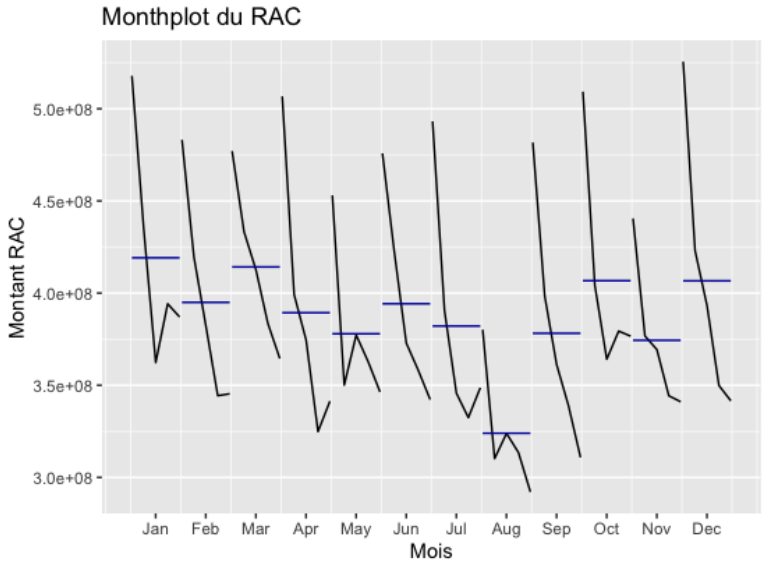

FIGURE 3.2 – Monthplot du RAC

La ligne horizontale en bleu représente le montant moyen des restes à charge pour chacun des mois et la longueur et les courbes verticales représentent l'importance du montant des RAC.
Dans une série sans saisonnalité les chronogrammes de chaque mois sont à peu près de la même forme, ce qui n'est pas le cas pour notre série, pour laquelle on observe un chronogramme plus petit au mois d'août, peut-être dû aux vacances d'été qui peuvent influencer à la baisse la consommation médicale ?

Le Lag-Plot

Le Lag-Plot est la représentation graphique des données décalées de k. On a en abscisse la donnée X_{t-k} et en ordonnée la donnée X_t. (Figure 3.3)

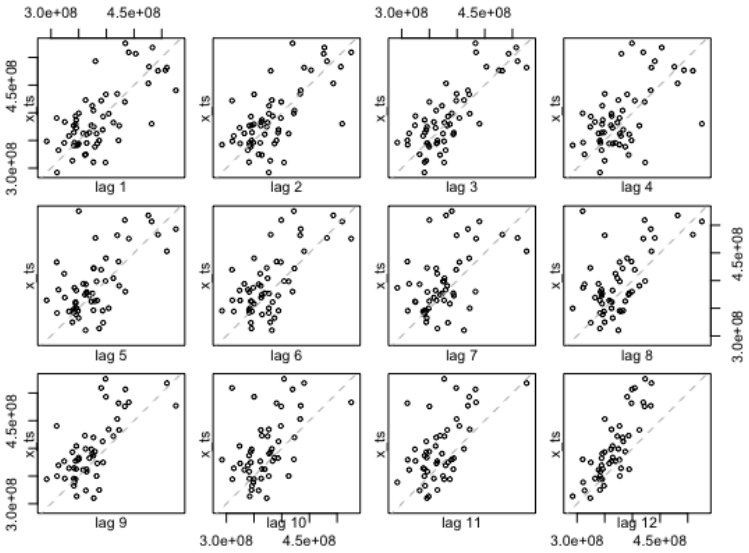

FIGURE 3.3 – Lag-plot du RAC

Le Lag-Plot nous donne des pistes d'analyse sur :

La distribution du modèle :
le tracé des retards est plutôt linéaire, ce qui fait penser à un modèle sous-jacent autorégressif.

Les valeurs aberrantes : Les décalages 8,10,11 et 5 montrent plusieurs valeurs aberrantes. Ils seront à éviter dans un éventuel exercice de différenciation pour obtenir de la stationnarité.

Le caractère aléatoire des données : En présence, d'un caractère aléatoire des données aucun motif n'apparait dans le tracé des Lag-plot. Ici, ce n'est pas le cas, on a une forme générale plutôt linéaire.

L'autocorrélation : Au décalage 12, nous avons une forte concentration de données sur la diagonale, ce qui suggère une forte auto-corrélation au lag 12.

La stationnarité

Les irrégularités de la série, la présence d'une tendance, éventuellement d'une saisonnalité (persistance de la baisse de montants observés sur le mois d'Août, sur toute la série, le Lag-Plot qui rejette toute structure aléatoire des données) nous font dire que la série n'est pas stationnaire. Nous analyserons davantage cette propriété dans la suite du mémoire.

3.2 Décomposition de la série temporelle

Comme vu aux paragraphes précédents, la série semble présenter une tendance et de la saisonnalité. Afin de projeter nos modèles nous allons décomposer cette série afin de décrire quelle est la nature de la tendance et de la saisonnalité mais également quel modèle sous-jacent lie ces composantes.

3.2.1 Choix du modèle

La littérature propose plusieurs méthodes pour estimer la nature du modèle (multiplicatif, additif ou mixte) d'une série temporelle.
Il est nécessaire de tester plusieurs méthodes de choix de modèle pour conclure, car on n'obtient pas toujours des conclusions uniformes.
Définir le type de modèle qu'on a est important car les méthodes à employer pour estimer et corriger la tendance et la saisonnalité dépend du modèle en présence.
De manière générale, on cherche à identifier si on a une variation saisonnière qui s'ajoute à la tendance, dans ce cas on peut conclure à modèle additif, en revanche si la variation saisonnière est proportionnelle à la tendance, on a affaire à un modèle multiplicatif ou mixte.

La méthode du profil :
Cette méthode consiste à superposer chaque série annuelle, sur le même graphique.
Si les courbes de profils sont parallèles alors le modèle est additif sinon il est multiplicatif.
Pour notre jeu de données (Figure 3.4) on constate que les profils sont parallèles par endroits et séquents à d'autres endroits.

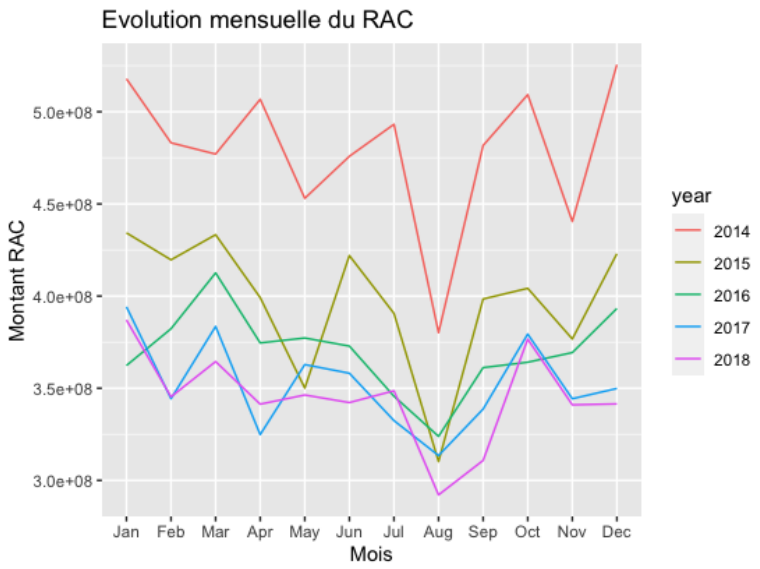

FIGURE 3.4 – Profil des chronogrammes annuel

Le test des profils ne permet pas de conclure fermement sur le type de modèle en présence.

La méthode de la bande :
Cette méthode consiste à représenter graphiquement la série chronologique, puis à tracer une droite passant respectivement par les minimas et par les maximas sur chaque période.
Si ces deux droites sont parallèles, nous sommes en présence d'un modèle additif. Dans le cas contraire, il s'agit d'un modèle multiplicatif.

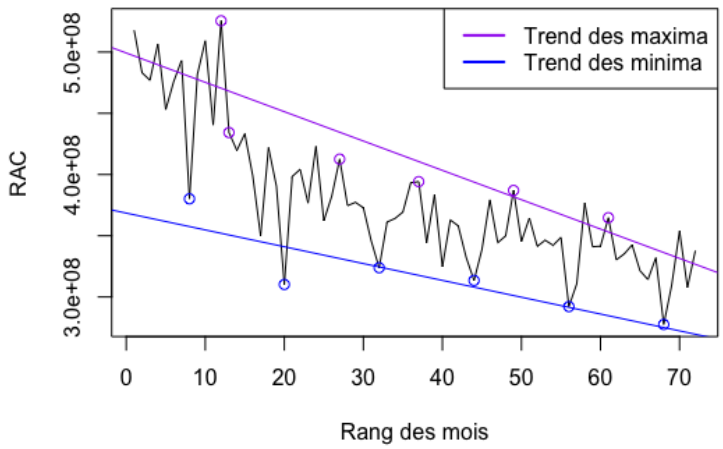

FIGURE 3.5 – Monthplot du RAC

Pour notre jeu de données, nous constatons que ces deux droites ne sont pas parallèles, elles sont de plus en plus séquentes ce qui tend à proposer un modèle multiplicatif.

La méthode analytique du tableau de Buys-Ballot :

Coefficients	Estimate	Std.Error	t.value	Pr(>|z|)
$\hat{b}$	-17630000	12550000	-1,405	0,2328
$\hat{a}$	0,1224	0,0329	3,721	0,0205

TABLE 3.1 – Coefficients du tableau de Buys-Ballot

Cette méthode consiste à calculer pour chaque période considérée, la moyenne et l'écart-type de la période. Puis de faire passer une droite des moindres carrés $\sigma = a\overline{x} + b$. Si le coefficient a est nul, alors on est en présence d'un modèle additif, sinon on a un modèle multiplicatif.

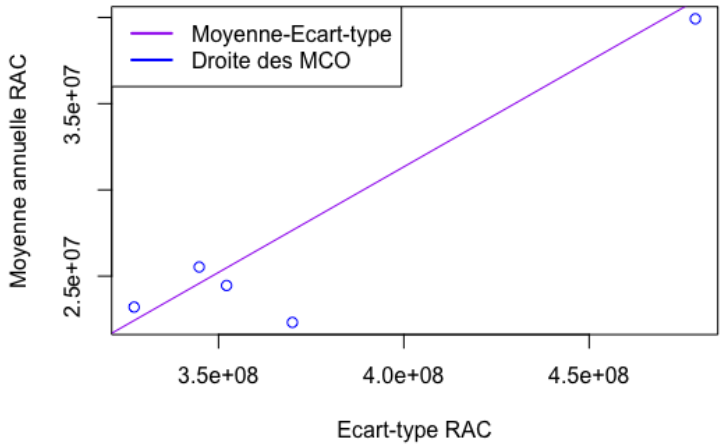

FIGURE 3.6 – Méthode analytique de Buys Ballot

La représentation de la droite des MCO est donnée par la Figure 3.6.
La valeur estimée du coefficient a est $\hat{a} = 0{,}1224$ avec une erreur de 0,0329.
Nous rejetons l'hypothèse de nullité du paramètre estimé $\hat{a}$ car on a $0 \notin [\ 0{,}0566\,;\ 0{,}1882]$, l'intervalle de confiance de $\hat{a}$.

Les 3 méthodes ci-dessus nous orientent donc plutôt vers le choix d'un modèle multiplicatif. En fonction de la nature de l'erreur ce modèle peut être :

— Multiplicatif pure : il sera du type $X_t = T_t \times S_t \times \varepsilon_t$,
— Multiplicatif partiel : $X_t = T_t \times S_t + \varepsilon_t$

3.2.2 Transformation des données par la méthode de Box-Cox

L'observation du chronogramme nous montre que la variance n'est pas constante dans le temps.
Il y a une forte variabilité des données pour les années 2014 et 2015 qui tend à se stabiliser par la suite.
Afin de stabiliser les données, nous allons effectuer une transformation de type Box-Cox qui aura également pour effet de permettre de se rapprocher de la normalité des erreurs.
On appelle transformation de Box-Cox, les transformations de $Z - t > 0$ par la fonction suivante :

$$F_\lambda(z) = \frac{z^\lambda - 1}{\lambda}, \text{ pour } (\lambda \neq 0)$$

$$F_\lambda(z) = ln(z) \text{ pour } (\lambda = 0).$$

Pour identifier le coefficient λ le plus adapté à la transformation de nos données, nous allons calculer la vraisemblance du modèle linéaire générée par :

$$F - \lambda(Z - t) = a - \lambda t + b - \lambda + \epsilon - t^\lambda.$$

On choisira le coefficient λ qui maximisera la vraisemblance.
La procédure *boxcox()* du package Mass du logiciel R permet d'automatiser cette démarche.
La vraisemblance maximale est atteinte pour une valeur de λ =-0.6548732.
Le modèle étant multiplicatif, cette transformation de Box-Cox a également rendu le modèle additif.
Une vérification par les méthodes de la bande, de Buys-Ballot a permis de confirmer le modèle additif sur les données transformées.

Il n'est pas nécessairement utile de réaliser la transformation qui correspond au meilleur λ. Dans la littérature, le choix d' une transformation logarithmique $ln(z)$ choisi pour le paramètre $\lambda = 0$ permet également de rendre également la série additive lorsqu'on a un modèle multiplicatif pure.

3.2.3 Estimation de la tendance

La tendance crée de la corrélation entre les éléments X_t de la série. Cette corrélation n'apporte pas d'information ou de caractère explicatif à la série.
Notre objectif est d'étudier la nature de cette tendance afin de l'éliminer dans le but de ne conserver que des liaisons à caractère explicatif dans le modèle.
La série corrigée de la tendance est dite "détendancialisée", elle est de la forme : $X_{CT,t} = S_t + \varepsilon_t$ pour un modèle additif.

La méthode des moindres carrés :
On suppose que $X_t = T_t + Z_t$ avec $T_t = \beta_0 + \beta_1 t + ... + \beta_d t^d$.
On souhaite estimer $\beta^* = (\beta_0^*, ..., \beta_d^*)^t$ à partir des $X_{t1}, ..., X_{tn}$.
La solution des moindres carrés est donnée par :

$$\hat{\beta} \in argmin_{\beta \in R^{d+1}} \frac{1}{n} \sum_{i=1}^{n} (X_{ti} - T_{ti})^2$$

La Figure 3.7, nous donne pour différentes valeurs "d" degré du polynôme, l'ajustement de la droite des moindres carrés aux données.

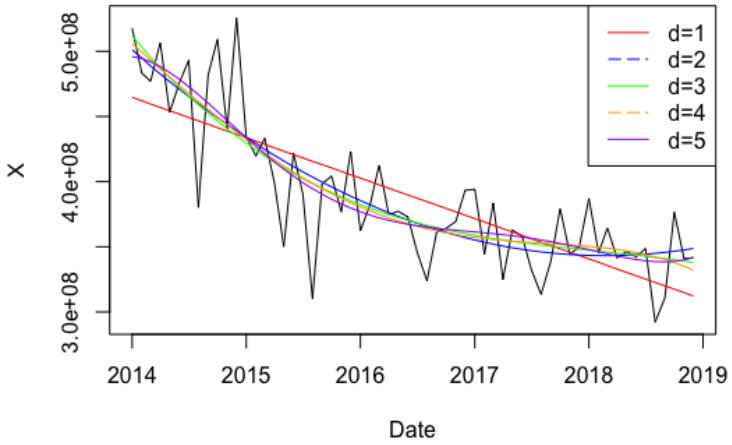

FIGURE 3.7 – Ajustement des polynômes estimés par MCO à la tendance

La tendance de notre jeu de données n'est pas linéaire, car pour d=1, la droite de tendance s'ajuste moins bien aux données. Par contre, les polynômes de degrés supérieurs de 2 à 5 semblent s'ajuster au mieux au modèle mais sont très proches pour qu'on puisse trancher visuellement sur le meilleur ajustement.
Afin de choisir le meilleur modèle de tendance, il nous faut un critère plus objectif.

Degré du polynôme	AIC	BIC
Dégré 1	-1673.092	-1666.809
Dégré 2	-1681.204	-1672.827
Dégré 3	-1679.889	-1669.418
Dégré 4	-1678.188	-1665.622
Dégré 5	-1677.042	-1662.381

TABLE 3.2 – Comparaison ajustement de la tendance par des polynômes

L'analyse des AIC des différents polynômes, nous permet de confirmer que c'est le polynôme de degré 2 qui s'ajuste le mieux, car c'est celui qui minimise l'AIC parmi les 5 proposés.
Les coefficients estimés sont les suivants : $\beta_0^* = 1.527$, $\beta_1^* = -3.307e^{-08}$, $\beta_2^* = 2.978e^{-10}$.
Soit le modèle : $T_t = 1.527 - 3.307e^{-08}t + 2.978e^{-10}t^2$.

Les résultats du test de Student de significativité des coefficients montrent que ces coefficients sont significatifs.

La méthode moyenne mobile :
L'opérateur moyenne mobile est un filtre appliqué sur la série afin de supprimer les fluctuations.
Plus schématiquement, il s'agit d'une moyenne calculée non pas sur toute la série mais d'une moyenne calculée sur chaque sous ensemble de taille q de la série.

L'opérateur moyenne mobile s'écrit :

$$\hat{T}_t = \frac{1}{2q+1} \sum_{k=-q}^{q} X_{t+k}$$

Nous allons utiliser la fonction *filter()* du logiciel R pour analyser cette tendance en posant q =12.

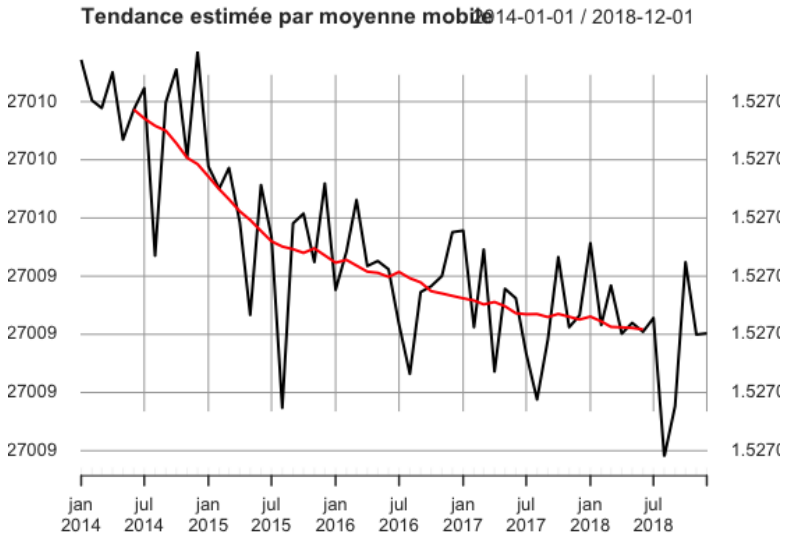

FIGURE 3.8 – Estimation de la tendance par MA

La tendance obtenue par moyenne mobile est représentée en rouge dans la Figure 3.8.

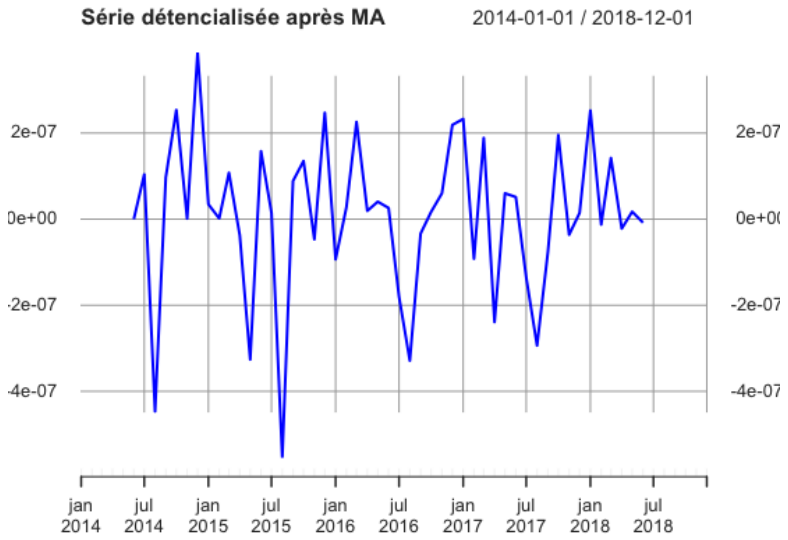

FIGURE 3.9 – Série détencialisée après MA

La Figure (Figure 3.9) représente la série détendancialisée suite à application du filtre moyenne mobile.

Polynômes locaux :
Cette technique d'élimination de la tendance consiste à approcher localement une fonction polynômiale au voisinage h, taille de fenêtre choisie. Nous allons utiliser la Fonction *Loess()* du logiciel R qui permet d'automatiser cette démarche. Ci-dessous la tendance estimée par Loess et la série détendancialisée.

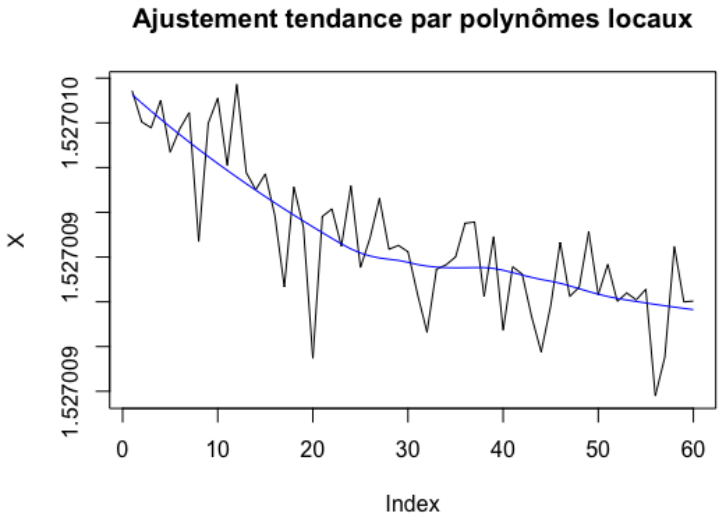

FIGURE 3.10 – Série détencialisée par polynômes locaux

Régression sur base de splines cyclique :
Nous allons utiliser une régression de type modèle additif généralisé pour estimer la tendance.
Nous allons utiliser la Fonction *Loess()* du logiciel R qui permet d'automatiser cette démarche.

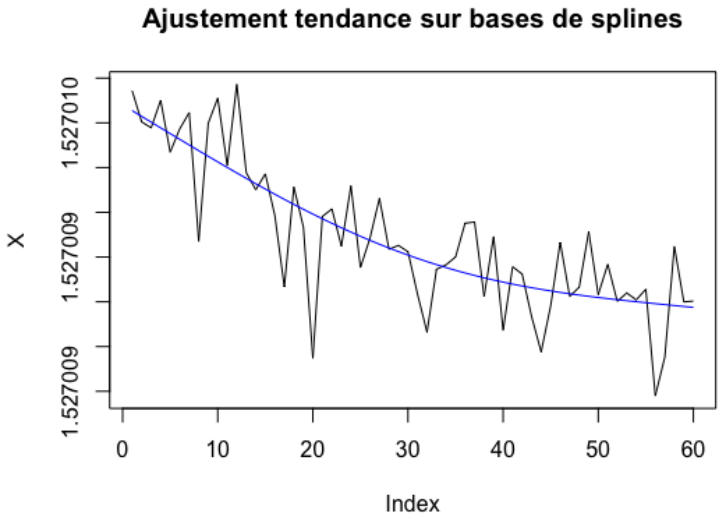

FIGURE 3.11 – Ajustement tendance sur bases de splines

Comparaison des méthodes d'estimation de la tendance

L'analyse graphique des différentes méthodes d'estimation de la tendance, nous montre que ces méthodes sont plutôt efficaces car elles s'ajustent visuellement à la tendance de la série. Mais nous ne nous permettent pas conclure sur le meilleur modèle.

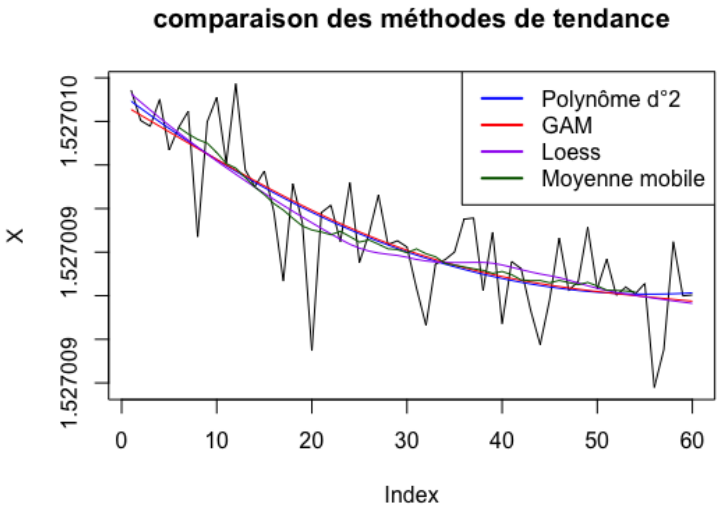

FIGURE 3.12 – Représentation graphique des tendances estimées

Afin de comparer les modèles d'estimation de la tendance, nous introduisons une métrique utilisée par HYNDMAN et ATHANASOPOULOS 2018. Elle mesure la force de la tendance définie par :
$F_t = max(0, 1 - \frac{Var(R_t)}{Var(R_t+T_t)})$
Avec R_t, la série détendancialisée et T_t, la série avant élimination de la tendance.
L'analyse de la force de la tendance sur les différents modèles de tendance étudiés sont regroupés dans la Table 3.3.

Polynôme de dégré 2	Regression sur bases de splines	Polynômes locaux	Moyenne mobile
0.6776903	0.6773374	0.68055	0.6908888

TABLE 3.3 – Variance résiduelle après détendancialisation de la série

Nous choisissons le modèle pour lequel la métrique F_t est la plus élevée, soit le modèle de détendancialisation par moyenne mobile.

3.2.4 Estimation de la saisonnalité

Maintenant qu'on a identifié le modèle de tendance applicable à nos données, nous allons corriger nos données des variations saisonnières sur la série détendancialisée par moyenne mobile.
Nous définissons une nouvelle série de données $Y_t = X_t - T_t$.

Les transformations appliquées afin d'estimer la tendance peuvent également s'utiliser pour la saisonnalité avec des paramètres différents.

Estimation paramétrique de la saisonnalité - décomposition par des séries de Fourier :

La forme naturelle de la saisonnalité fait penser à des fonctions trigonométriques. Nous allons utiliser une décomposition par des séries de Fourier afin de modéliser la saisonnalité.
Soit un processus y_t admettant une saisonnalité de période τ alors le modèle suivant est généralement proposé :

$$y_t = \sum_{j=1}^{q} a_j cos(\omega_j t) + b_j sin(\omega_j t) + \varepsilon_t$$

où $\omega_j = 2j\pi/\tau$,

Le paramètre "q" est identifié par une méthode de sélection de modèles et les coefficients a_j et b_j sont obtenus par estimation des moindres carrés.

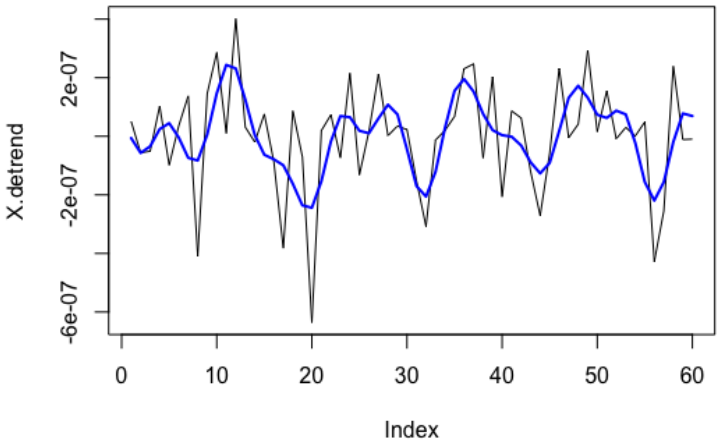

FIGURE 3.13 – Représentation graphique de la série détendancialisée et désaisonnalisée

Estimation non-paramétrique de la saisonnalité par Moyenne mobile :
On peut, en choisissant la bonne valeur de décalage, estimer la composante saisonnière d'un processus par différentiation.

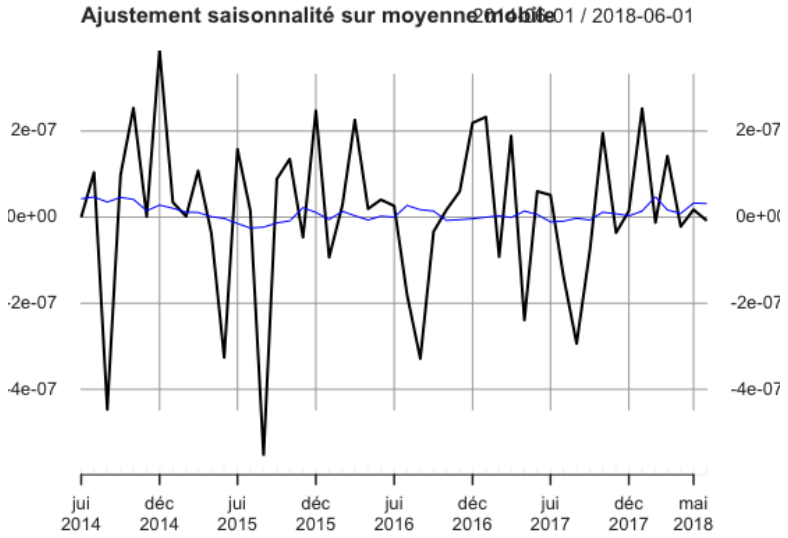

FIGURE 3.14 – Décomposition/Ajustement saisonnalité sur moyenne mobile

Estimation semi-paramétrique de la saisonnalité par spines cycliques :
Comme pour la tendance, il est possible d'identifier la saisonnalité à partir des spines cycliques. On utilisera la fonction *gam()* du logiciel R en faisant varier les paramètres.

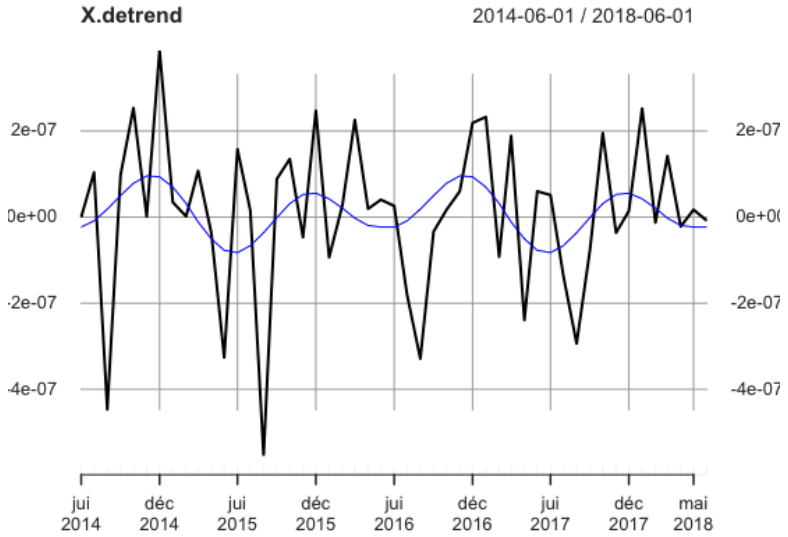

FIGURE 3.15 – Décomposition/Ajustement saisonnalité sur base de spines cycliques

Evaluation des résidus de la série décomposée :
Afin de comparer les modèles d'estimation de la saisonnalité, nous introduisons une métrique utilisée par Wang, Smith, Hyndman, identique à celle de la tendance qui est une mesure de la variance résiduelle dans le modèle après désaisonnalisation :

$$F_S = \max\left(0, 1 - \frac{\mathrm{Var}(R_t)}{\mathrm{Var}(S_t + R_t)}\right).$$

Avec R_t, la série désaisonnalisée et $S_t + R_t$, la série avant élimination de la saisonnalité.
L'analyse de la force de la saisonnalité sur les différents modèles de saisonnalité étudiés sont regroupés dans la Table 3.4.

Séries de Fourier	Moyenne mobile	Splines cycliques
0.36370510	0.01290729	0.12318971

TABLE 3.4 – Variance résiduelle après détendancialisation et désaisonnalisation de la série

Nous choisissons le modèle pour lequel la métrique F_s est le plus elevé, soit le modèle de désaisonnalisation par série de Fourier.

Nous obtenons une série corrigée des variations saisonnières par série de Fourier et de la tendance par moyenne mobile.
Cette dernière devrait être stationnaire et se rapprocher d'un bruit blanc. Nous allons en évaluer les propriétés.

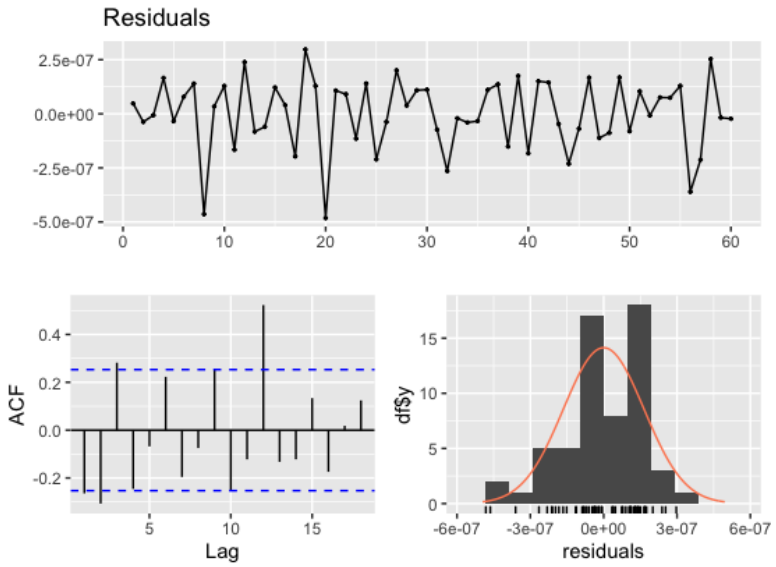

FIGURE 3.16 – Représentation graphique de la série détendancialisée et désaisonnalisée

Bien que le test de racine unité de KPSS conduit à ne pas rejeter l'hypothèse nulle de stationnarité car la statistique du test est inférieure au niveau critique à 5%, on constate sur l'autocorrélogramme de la série, un pic au lag 12, ce qui montre que la saisonnalité n'a pas été totalement capturée.
Nous ne retiendrons pas ce modèle pour prédire le reste à charge des dépenses en santé.
Ce type de modèle de décomposition suppose une connaissance exacte de la structure de modèle de la série temporelle.
Une source d'erreur potentielle dans cette décomposition peut être dû au fait qu'ayant supposé que les données après transformation de Box-Cox avaient rendu le modèle additif (confirmé par d'autres tests de la bande et Buys-Ballot non exposé dans ce mémoire) mais le modèle peut-être mixte et dans ces cas la décomposition manuelle peut être ardue.

3.2.5 Estimation conjointe de la tendance et de la saisonnalité

Dans cette section, nous allons utiliser des méthodes automatiques de décomposition des séries temporelles sous R. L'avantage de ces méthodes est qu'elles offrent en général une meilleure décomposition que la décomposition "manuelle" étudiée plus haut, elles sont rapides à mettre en oeuvre ce qui est un avantage en milieu professionnel mais l'inconvénient est qu'elles ne permettent pas de bien spécifier la nature et les paramètres des éléments qui composent la série.
Dans un objectif de prévision, ces méthodes sont suffisantes et adaptées.

La décomposition par moyenne mobile :
Nous allons utiliser la fonction *décompose()* du logiciel R pour mettre en œuvre cette procédure.
La fonction détermine d'abord la composante de tendance à l'aide d'une moyenne mobile et la supprime de la série temporelle. Puis, elle calcule automatiquement une fenêtre pour la saisonnalité, en faisant la moyenne pour chaque pas de temps, sur toutes les périodes possibles.
Enfin, la composante d'erreur est déterminée en supprimant la tendance et la saisonnalité de la série chronologique d'origine.
Cette méthode est adaptée si la série possède un nombre entier de périodes complètes.
En paramètre de cette fonction, nous pouvons spécifier le type de modèle (additif ou multiplicatif) souhaité pour mettre en oeuvre la décomposition.

En testant les 2 versions, nous arrivons à la conclusion que la décomposition par modèle additif est mieux adaptée pour les données modifiées par Box-Cox car l'ajustement visuel est légèrement meilleur et que la décomposition par modèle multiplicatif se prête mieux au jeu de données brutes. Cette analyse visuelle n'est pas toujours évidente, c'est pourquoi il faut connaitre le type de modèle en présence pour le spécifier au logiciel.
Nous présenterons ici uniquement la décomposition par modèle additif sur notre base de données transformée par Box-Cox.

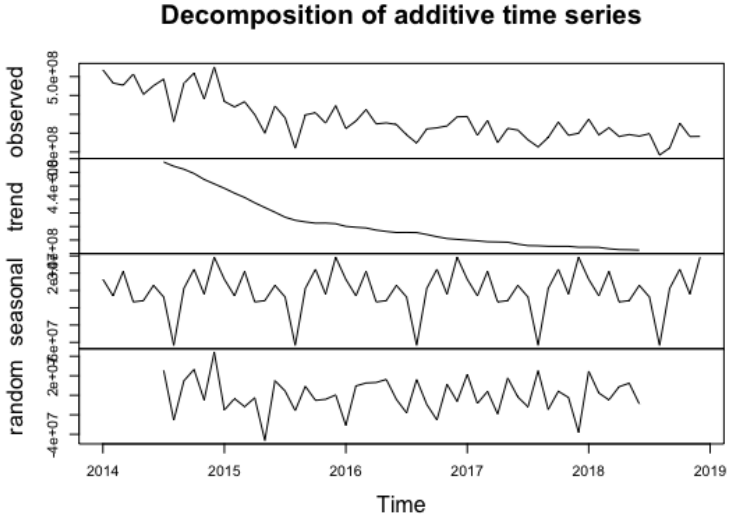

FIGURE 3.17 – Représentation graphique de la série détendancialisée et désaisonnalisée par *decompose()*

Les résidus sont visuellement centrés et erratiques, ils ne présentent pas de structure ou forme parti-

culière ce qui ressemble à un bruit blanc.
La tendance, est bien décroissance et ressemble à l'allure d'un polynôme de degrés 2 comme attendu.
Le signal saisonnier est bien régulier et défini un schéma annuel facilement identifiable.
Un zoom sur ce schéma périodique (Figure 3.18) permet de bien comprendre les variations infra-annuelles du montant des RAC. Hormis la baisse importante des prestations sur le mois d'août, les prestations des autres mois varient autour d'une moyenne.

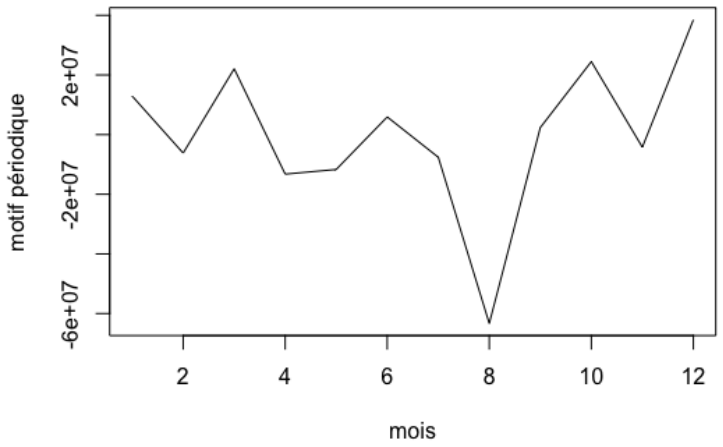

FIGURE 3.18 – Motif périodique infra-annuel

La Figure 3.19 est la représentation sur le même graphique de la série temporelle avec la moyenne m_t et la saisonnalité s_t estimées par la fonction *decompose()*.
On peut dire que le modèle $m_t + s_t$ s'ajuste bien aux données moyennes. Lorsqu'on a une variance de la consommation un peu plus forte comme en 2014 et 2015, l'estimation est sous-estimée.

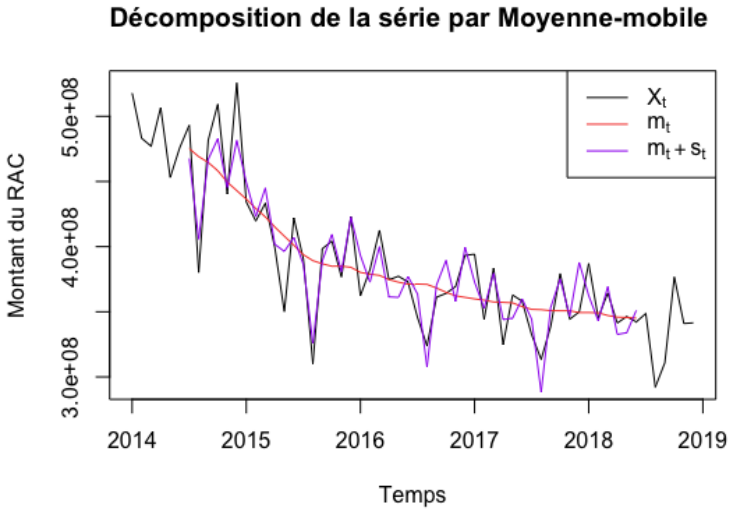

FIGURE 3.19 – Décomposition de la série par Moyenne-mobile

La décomposition par les polynômes locaux :
Nous allons utiliser la fonction *stl()* du logiciel R pour mettre en œuvre cette procédure.
STL est une méthode polyvalente et robuste pour décomposer des séries chronologiques. Elle utilise la méthode des polynômes locaux qui modélise les liens non linéaires. Cette méthode a été développée par R. B. Cleveland, Cleveland, McRae et Terpenning (1990). Les avantages de cette fonction par rapport à la fonction décompose est qu'elle prend en entrée plusieurs paramètres permettant de tester différents modèles notamment. L'utilisateur peut spécifier une saisonnalité différente au cours du temps en le précisant dans le modèle. La douceur de la tendance peut également être contrôlée par l'utilisateur.
L'inconvénient de cette fonction est qu'elle ne gère que les séries additives.
La composante saisonnière est trouvée en lissant à l'aide des polynômes locaux, chaque sous-série mensuelle.
Les valeurs saisonnières sont supprimées et le reste est lissé pour trouver la tendance. Le niveau global est retiré de la composante saisonnière et ajouté à la composante tendancielle. Ce processus est répété plusieurs fois. La composante restante correspond aux résidus de l'ajustement saisonnier plus tendance.

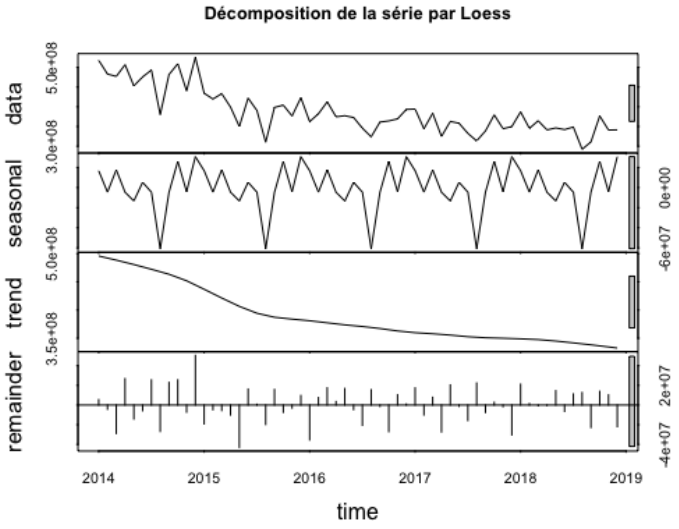

FIGURE 3.20 – Décomposition de la série par Loess

La Figure 3.20 ci-dessus nous donne l'ajustement de la série RAC des dépenses en pharmacie.
Cet ajustement est graphiquement très proche de celui donné par la fonction *decompose()*.
Le modèle s'ajuste bien aux données de départ.

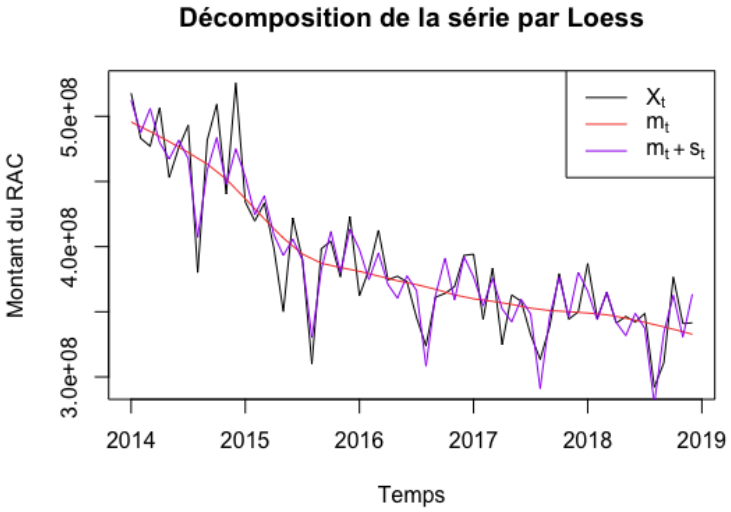

FIGURE 3.21 – Ajustement de la série par Loess

Comparaison des décompositions par les polynômes locaux et par moyenne mobile

Analyse graphique des autocorrélations des résidus :
L'autocorrélogramme empirique est significatif car quasiment tous les pics sont dans la zone d'acceptation. Les résidus varient autour de la valeur 0, et s'ajustent visuellement bien à une loi Normale.

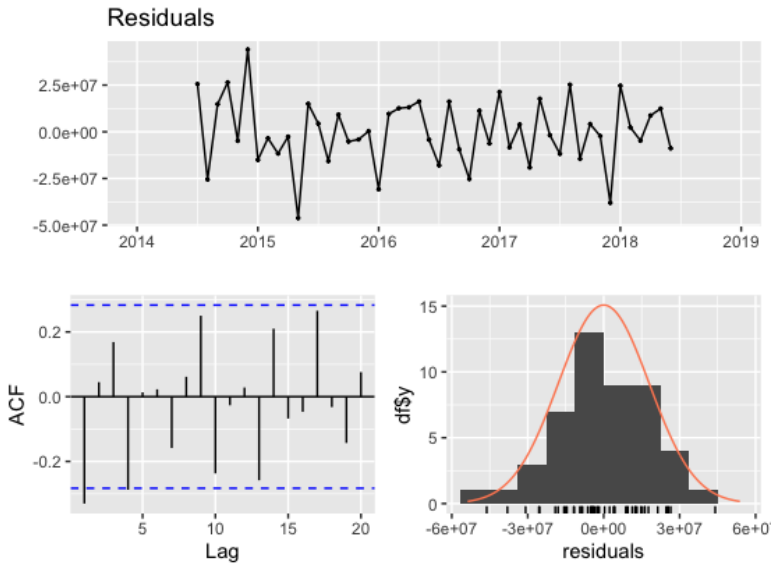

FIGURE 3.22 – Analyse des résidus de la série décomposé par Moyenne mobile

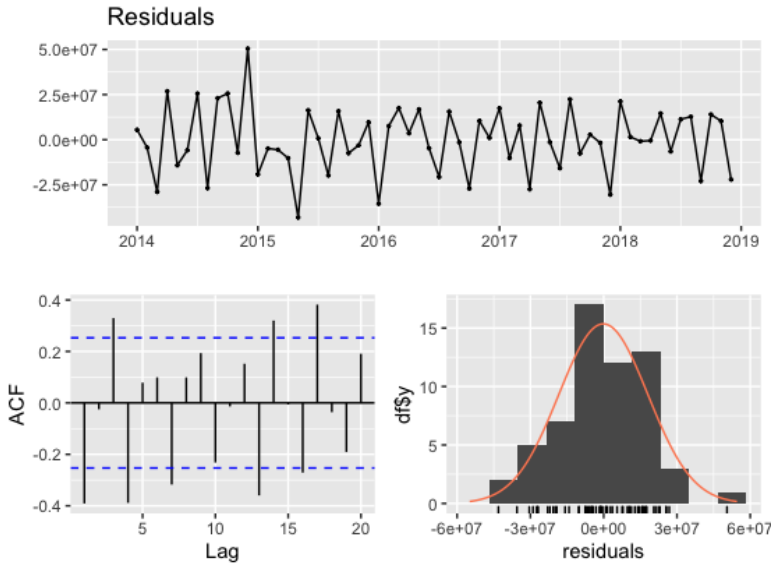

FIGURE 3.23 – Analyse des résidus de la série décomposée par Loess

Tests de blancheur des résidus :

Afin de déterminer si les résidus sont normaux nous allons observer leur QQ-Plot respectifs.

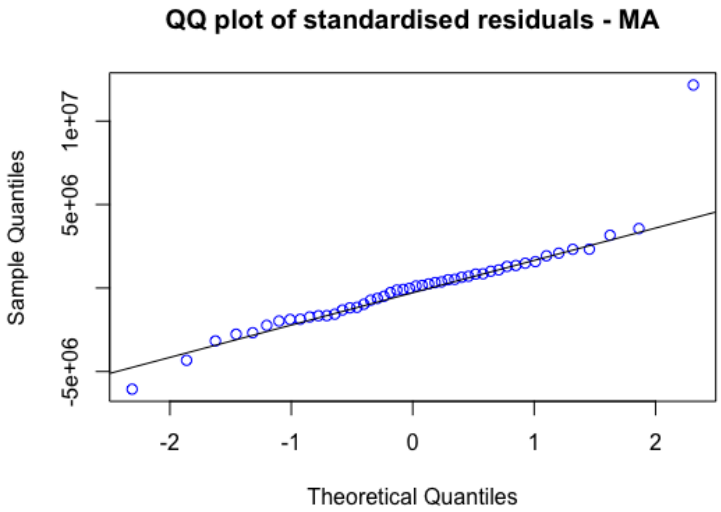

FIGURE 3.24 – QQ plot of standardised residuals - MA

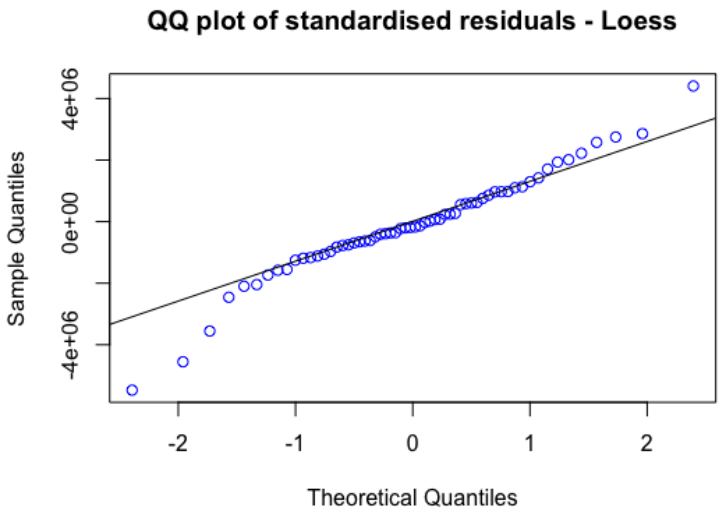

FIGURE 3.25 – QQ plot of standardised residuals - Loess

La Figure 3.24 nous montre que la projection MA est normale car les quantiles théorique de la loi normale s'alignent sur la droite de Henry.
Nous allons confirmer ce résultat avec des tests statistiques.

Tests statistiques de non corrélation et normalité des résidus :
Nous allons utiliser la statistique Ljung-Box Q (LBQ) au décalage 12. Cette statistique permet de tester l'hypothèse nulle selon laquelle les autocorrélations jusqu'au décalage 12 sont égales à zéro soit que les valeurs sont aléatoires et indépendantes jusqu'au décalage 12.
Afin de tester l'hypothèse de normalité des résidus, nous utiliserons le test d'adéquation à la loi normale de Shapiro–Wilk qui teste l'hypothèse nulle selon laquelle les résidus sont normalement distribués.

Modèle	Type de test	Statistique	P-Value
Decompose()	Ljung-Box Q (Blancheur)	16.439	0.1719
Decompose()	Shapiro-Wilk (Normalité)	0.96531	0.1653
Stl()	Ljung-Box Q (Blancheur)	31.079	0.001916
Stl()	Shapiro-Wilk (Normalité)	0.96643	0.09745

TABLE 3.5 – Test du portemanteau et de normalité : méthode de décomposition

Le modèle retenu est le modèle estimé par la fonction *décompose()*.
Les P-value des tests de normalité de Shapiro-Wilk et du bruit blanc de Ljung-Box conduisent à ne pas rejetter les hypothèses nulles de normalité et blancheur car elles sont supérieures au seuil α =0,05.

Le modèle *Stl()* avec une P-value inférieur à α =0,05 pour le test de Ljung-Box conclu une corrélation des résidus on a donc pas un bruit blanc.

Analyse de la prévision :
Sur notre échantillon de tests (données 2019), nous avons réalisé la projection des modèles *Stl()* et *decompose()*
Les performances ont été évaluées sur la base de deux mesures : l'erreur absolue moyenne (MAE) et

l'erreur quadratique moyenne (RMSE).

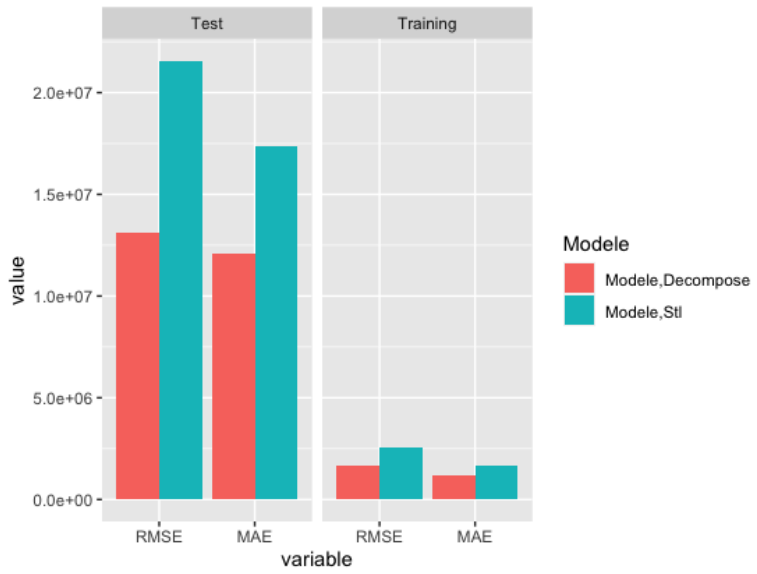

FIGURE 3.26 – Comparaison des RMSE et MAE des modèles decompose.

On constate que les niveaux moyennes de la MAE et la RMSE sont plus faibles pour l'échantillon d'apprentissage que l'échantillon de test. Le modèle MA minimise l'écart de prédiction en comparaison au modèle Loess.

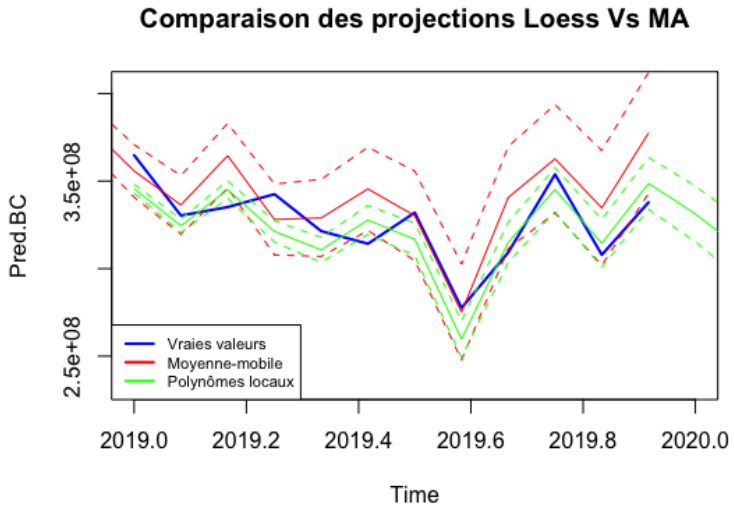

FIGURE 3.27 – Comparaison des projections Loess vs MA

L'intervalle de confiance à 95% du modèle MA est plus large que celui des polynômes locaux, et donc une probabilité plus grande que la vraie valeur soit dans cet intervalle. Hormis le mois de février où la vraie valeur ne se trouve ni dans l'intervalle du modèle MA, ni dans l'intervalle du modèle polynômes locaux.

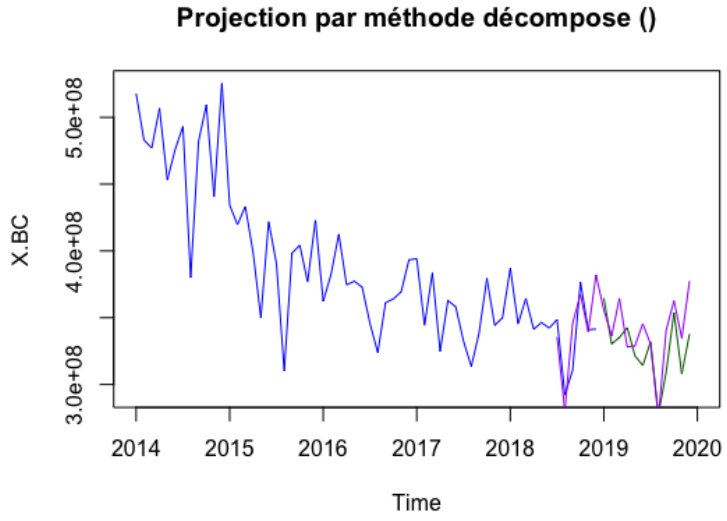

FIGURE 3.28 – Comparaison des indicateurs de performances entre les méthodes *STL* et *decompose*

Comme le montre le graphique ci-dessus, la fonction *décompose()* est celle qui minimise l'écart de prédiction.

3.3 Modélisation par lissage exponentiel

Comme vu précédemment, la série du RAC en pharmacie comporte une composante saisonnière et une tendance. La méthode de lissage la plus appropriée est donc celle de Holt-Winters.
Nous allons utiliser les fonction *ets()* du package Forecast de R en faisant varier les paramètres pour tester différents types de lissage.

La fonction *Forecast()* sera utilisée afin de prédire les données de l'année 2019.

3.3.1 Modélisation par lissage exponentiel sur les données brutes

Dans un premier temps nous allons tester la modélisation sur les données de reste à charge des dépenses en pharmacie brutes c'est à dire sans transformation par Box-Cox.

Modèles candidats :
Nous allons tester tous les modèles possibles : additif, multiplicatif et mixte afin d'identifier celui qui s'ajuste le mieux à notre jeu de données.

Nous nommerons par la suite, les modèles candidats en 3 lettres.
La première lettre désigne le type d'erreur, la deuxième la nature de la tendance et la troisième la nature de la saisonnalité. Chacune des lettres peuvent prendre les valeurs ("A" pour "additif", "M" pour "multiplicatif" et "N" pour "absence").
Par exemple le modèle "M,A,N" désigne un modèle avec erreur multiplicatif, tendance additive et sans saisonnalité soit le modèle de type : $X_t = T_t \times \varepsilon_t$.
Le tableau ci-dessous résume les AIC, BIC et AICc des différents modèles testés.

Nature du modèle	AIC	AICc	BIC
"A,N,N"	2335,469	2335,898	2341,752
"A,A,N"	2335,373	2336,484	2345,845
"A,A,A"	2307,434	2324,117	2345,132
"M,A,N"	2329,864	2331,449	2342,43
"M,A,A"	2307,006	2323,689	2344,705
"M,N,N"	2330,106	2330,534	2336,389
"M,M,M"	2296,725	2313,408	2334,423
"M,M,N"	2321,327	2322,912	2333,893
"M,A,M"	2298,658	2315,341	2336,356

TABLE 3.6 – Comparaison des modèles de lissage exponentiel

Nous remarquons sans surprise que les modèles sans tendance ni saisonnalité sont ceux avec l'AIC le plus élevé et ont la moins bonne qualité d'ajustement (les modèles exponentiels simple et double prédisent une droite). Nous ne retenons pas ces modèles car non adaptés à notre jeu de données.
L'AICc le plus faible (2313,408) est obtenu pour le modèle ("M,M,M") mais reste très proche de celui du modèle ("M,A,M"), ce qui ne permet pas de trancher entre les 2 modèles.

Analyse de l'autocorrélogramme :
Le test visuel basé sur l'analyse de l'auto-corrélogramme s'écrit de la manière suivante :
L'hypothèse nulle (H_0) :ε_t est un bruit blanc
L'hypothèse alternative (H_1) :ε_t n'est pas un bruit blanc

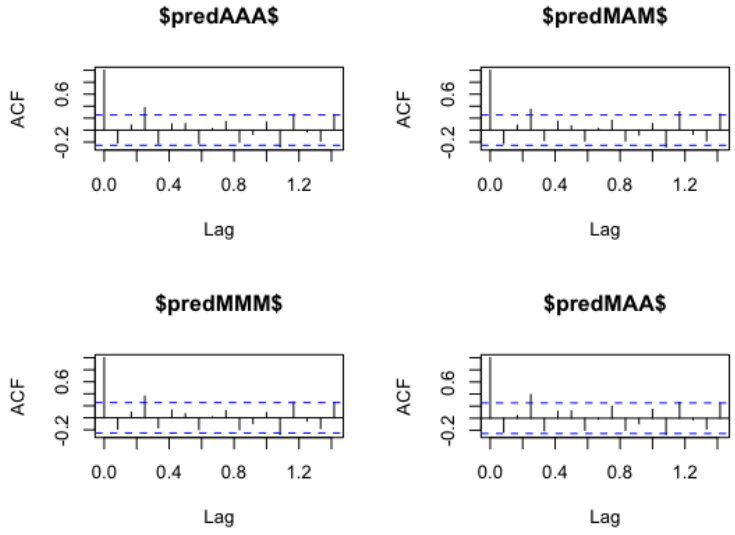

FIGURE 3.29 – Autocorrelogramme des résidus des modèles avec données brutes

Les quatre autocorrélogrammes des modèles ("M,M,M"), ("M,A,M"), ("M,A,A") , ("A,A,A"), se ressemblent, ils sont tous significatifs.

3.3.2 Modélisation par lissage exponentiel sur les données transformées par Box-Cox

Dans cette partie, nous étudierons l'impact de la transformation des données par Box-Cox sur la prédiction par lissage exponentiel.
Nous allons utiliser λ = -0.6549, tel que démontré plus haut. La méthode automatique de sélection de modèles exponentiels par la fonction *ets()* du logiciel R, nous donne le modèle ci-dessous :
Pour distinguer le modèle qui va suivre avec celui précédant étudié, nous allons le nommer ("A,Ad,A").
Ainsi, on a le modèle de type : ("A,Ad,A") $= X_t = T_t + S_t + \varepsilon_t$ avec les paramètres de lissage estimés suivants :

$\hat{\alpha} = 0.116$, $\hat{\beta} = 1e^{-04}$, $\hat{\gamma} = 1e^{-04}$ et $\hat{\phi} = 0.9782$.

Les coefficients initiaux sont :
l = 1.527, b = 0 , s = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) et σ= 0.

L'AIC, AICc et le BIC du modèle ("A,Ad,A") sur les données transformées par Box-Cox sont nettement plus faibles que ceux qu'on a pu observer sur les données brutes (Tableau 3.4)

AIC	AICc	BIC
-1631.852	-1615.169	-1594.154

TABLE 3.7 – Résultats du modèle de Holt-Winters sur données transformées par Box-Cox

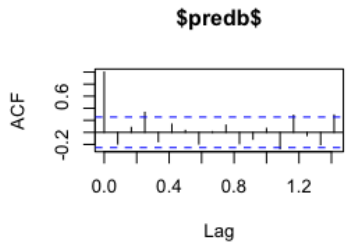

FIGURE 3.30 – Autocorrélogrammes empiriques des résidus des modèles avec données transformées par Box-Cox

On remarque que la plupart des résidus sont à l'intérieur de la bande de confiance : on ne rejette pas l'hypothèse (H_0) que les résidus sont des bruits blancs.

Test du porte-manteau
Nous allons réaliser un test de Ljung-Box avec comme hypothèses :

$$(H_0) : \rho_1, \rho_2, \ldots, \rho_h = 0 \text{ contre } (H_1) : \text{ Au moins un des } \rho_1, \rho_2, \ldots, \rho_h \text{ est nul .}$$

Modèle	X-squared	p-value
"A,A,A"	27.381	0.006809
"M,A,M"	22.227	0.03505
"M,M,M"	28.782	0.004244
"M,A,A"	23.19	0.02615
"A,Ad,A"	20.293	0.06175

TABLE 3.8 – Comparaison des tests du porte-manteau sur les modèles de lissage

D'après les résultats du test de Ljung-Box pour nos 5 modèles candidats (Table 3.8), on déduit que pour un niveau de décalage fixé à 12, seul le modèle avec données transformées est susceptible d'être un bruit blanc, car il est le seul modèle pour lequel la P-value de statistique de Ljung-Box est supérieure au seuil de 5%.

Tests de normalité :
Afin de pouvoir réaliser un intervalle de confiance, nous allons tester la normalité des résidus du modèle ("A,Ad,A") sur les données transformées.
On pose :
(H_0) : ε_t est gaussien contre (H_1) : ε_t n'est pas gaussien.

Sur le QQ-plot ci-dessus le nuage de points s'aligne sur la droite :

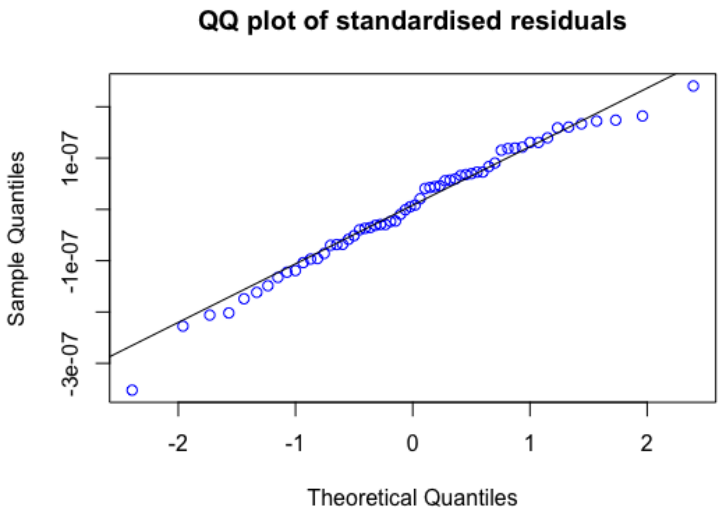

FIGURE 3.31 – QQ-plot ajustement de la série des RAC sur un modèle de lissage exponentiel de type (A,Ad,A)

la P-value donnée par le test de Shapiro-Wilk est de 0.4363 qui est supérieure au seuil fixé à 5%. On ne rejette pas l'hypothèse nulle (H_0).
Il y a de forte chance que les résidus de ce modèle soient normaux.

3.3.3 Prévision à l'aide du modèle choisi

Le modèle de lissage exponentiel choisi est celui de type additif sur les données après transformation de Box-Cox, en raison de la qualité des résidus.

Nous allons utiliser le modèle retenu afin d'évaluer les montants de RAC attendus sur l'année 2019. Ci-dessous l'ajustement graphique obtenu :

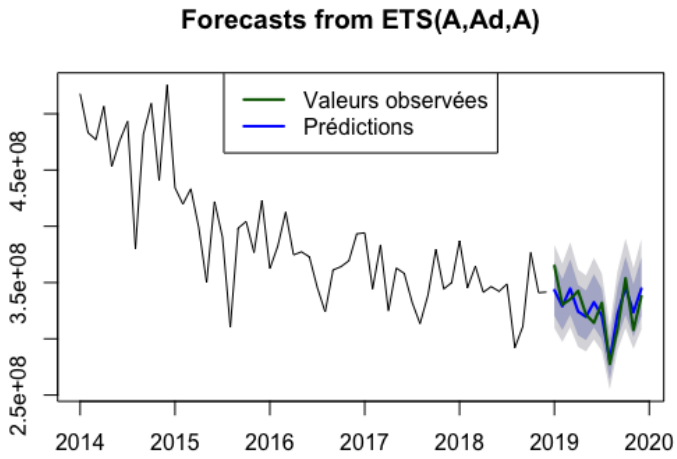

FIGURE 3.32 – Ajustement du modèle de lissage exponentiel de type (A,Ad,A) sur l'année 2019

Les estimations semblent bien s'ajuster aux données réelles de 2019.
La qualité de prévision sera discutée à la fin de ce chapitre en même temps que les modèles SARIMA et de décomposition.

3.4 Modélisation par Box-Jenkins

Les modèles ARIMA apportent une approche complémentaire à la modélisation des séries temporelles par rapport aux modèles de lissage exponentiel.
En effet, tandis que les modèles de lissage exponentiel sont basés sur la description de la nature de la tendance et de la saisonnalité, les modèles de type ARMA vont plus loin et proposent de modéliser l'autocorrélation dans les données.

3.4.1 Étude de la stationnarité

Le chronogramme et les différentes décompositions de la série ont mis en évidence la présence d'une tendance et d'une saisonnalité au sein de la série. On peut donc en déduire que la série est non-stationnaire. Nous allons confirmer cette supposition en analysant les autocorrélogrammes simples et partiels de la série des RAC.

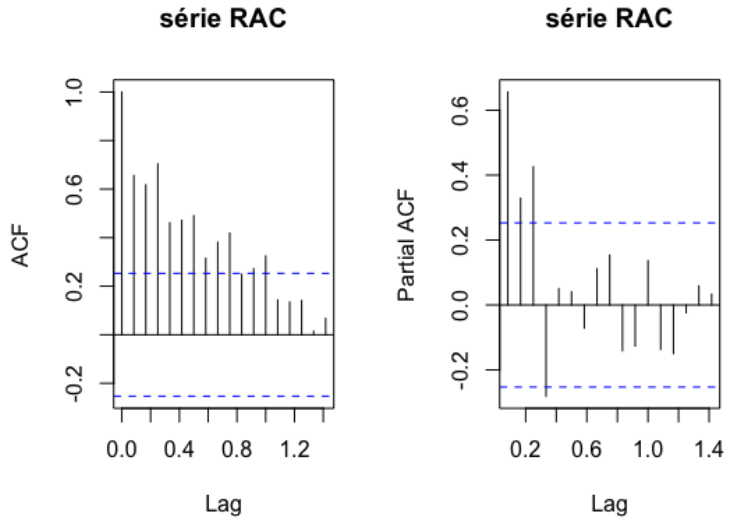

FIGURE 3.33 – Autocorrélogrammes simple et partiel de la série

L'autocorrélogramme de cette série confirme que les données de RAC des dépenses en pharmacie ne sont pas stationnaires.
En effet, pour une série chronologique stationnaire, on s'attend à ce que les traits de l'AFC se retrouvent majoritairement dans l'intervalle de confiance représenté par la bande bleue. Ici, on a plusieurs traits qui sont hors de l'intervalle de confiance et de plus l'AFC converge lentement vers 0.
On peut utiliser également des tests statistiques pour confirmer cette non-stationnarité.

Dans la littérature, il existe plusieurs tests de racine unitaire, basés sur des hypothèses différentes et qui peuvent conduire à des conclusions contradictoires.
Dans notre analyse, nous utilisons le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS), pour lequel l'hypothèse nulle est l'hypothèse que les données sont stationnaires.
Nous recherchons les preuves que l'hypothèse nulle soit fausse.
Par conséquent, pour des valeurs très faibles de p on va conclure à la non stationnarité.
Lorsque les données ne sont pas stationnaires, il faut passer par une étape préalable de stationnarité afin d'effectuer par la suite une modélisation de Box-Jenkins.

Afin de réduire la variance de la série, dans la suite de cette partie nous utiliserons une transformation de type logarithme tel que :

$$Y_t = ln(X_t)$$

On aurait pu utiliser la même transformation de type Box-Cox présenté dans le paragraphe 2, mais les résultats finaux ne montrent pas de différences significatives dans le choix du modèle selon qu'on applique une transformation de type 1/x ou ln(x).

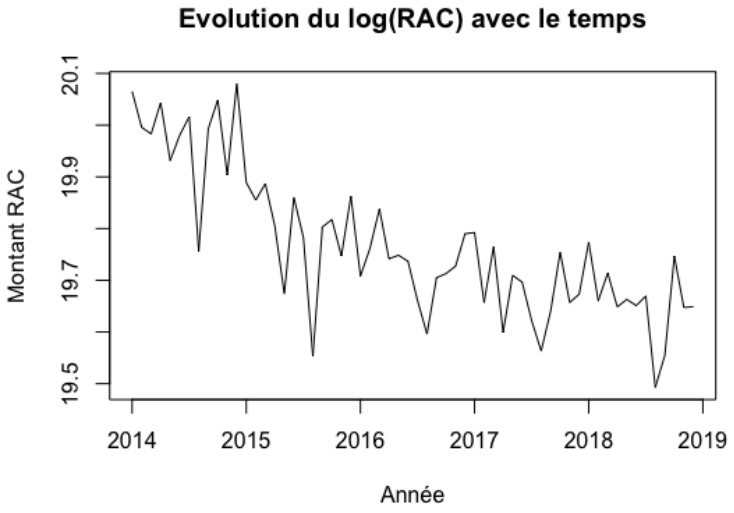

FIGURE 3.34 – Evolution du ln(RAC) avec le temps

Nous allons utiliser la fonction *ur.kpss()* du package urca du logiciel R sur les données transformées en ln(X).
Nous obtenons le résultat ci-dessous :

Type de test	Statistique	Valeur critique à 5%
KPSS (stationnarité)	1.2906	0.463

TABLE 3.9 – Test de stationnarité de la série ln(RAC)

La statistique de test obtenue est de 1.2906 qui est bien supérieure à la valeur critique de 5% (0.463), ce qui indique que l'hypothèse nulle est rejetée : les données ne sont pas stationnaires.

3.4.2 Stationnarisation

Afin d'étudier la série temporelle, la première étape de la méthode de Box-Jenkins est de stationnariser la série temporelle. Il existe plusieurs méthodes pour stationnariser la série, nous utiliserons la méthode des différences.
Nous allons appliquer à cette série une première différentiation afin de supprimer la marche aléatoire an appliquant l'opérateur de différence première (I-B) on obtient une nouvelle série :

$$DeltaY_t = Y_t - Y_{t-1} = c + \varepsilon_t$$

Le chronogramme de la série $DeltaY_t$ ne montre plus de tendance.

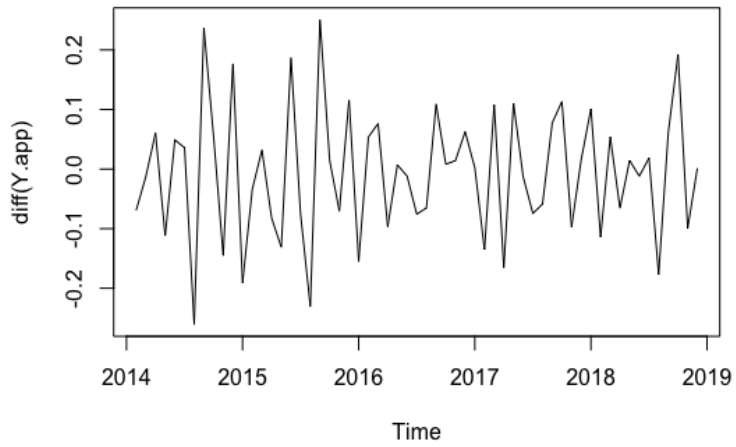

FIGURE 3.35 – Chronogramme de la série différenciée 1 fois

Le résultat du test de KPSS pour la série différenciée nous donne une statistique de test = 0.0816 qui est très inférieure à la valeur critique à 5% = 0.463. On ne rejette donc pas l'hypothèse nulle de stationnarité de la série. La procédure *ndiffs()*, nous permet de confirmer que le nombre de différence première pour atteindre la stationnarité est 1.
Cependant, la fonction d'autocorrélation présente des pics significatifs pour le premier retard et pour les retards aux pas 1 et 12. ce qui suggère une saisonnalité d'ordre 12.

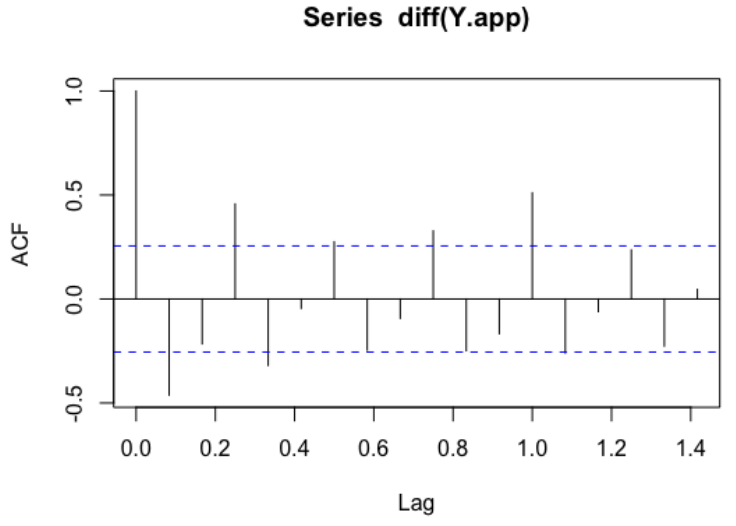

FIGURE 3.36 – autocorrélogramme de la série différenciée

Afin d'éliminer la saisonnalité résiduelle, nous allons appliquer un second filtre de différentiation saisonnière cette fois d'ordre 12. On définit une nouvelle série de valeurs :

$$Delta_{12}DeltaY_t = Y_t - Y_{t-12},$$

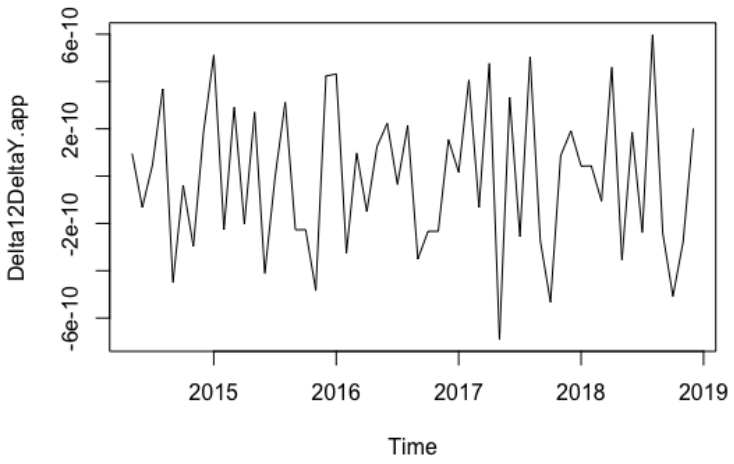

FIGURE 3.37 – Chronogramme de la série $Delta_{12}DeltaY_t$

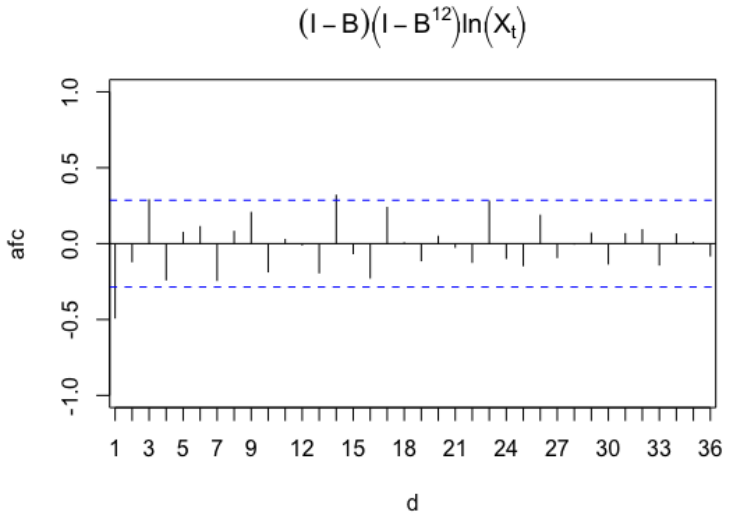

FIGURE 3.38 – Autocorrélogramme de la série $Delta_{12}DeltaY_t$

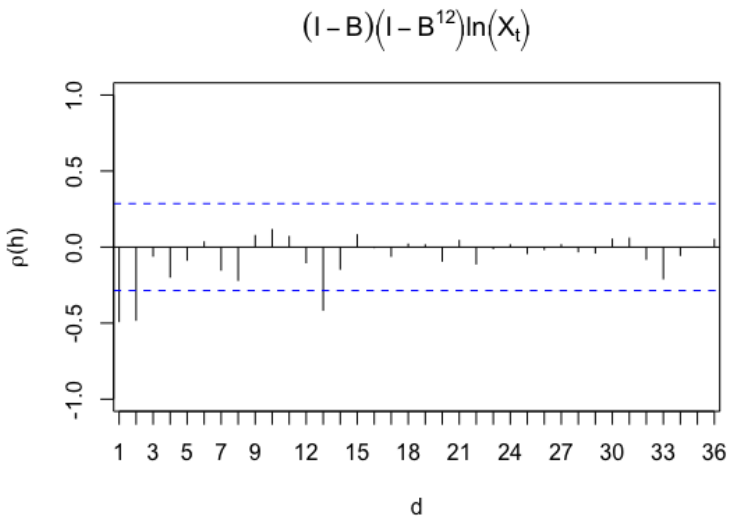

FIGURE 3.39 – Autocorrélogramme partiel de la série $Delta_{12}DeltaY_t$

L'autocorrélogramme de la série $Delta_{12}DeltaY_t$, montre que la série est maintenant stationnaire. Tous les pics sont dans l'intervalle de confiance. On ne rejette pas H_0.

Le résultat du test de stationnarité de KPSS nous donne une statistique de test = 0.0459 très inférieure à la valeur critique à 5% = 0.463.
On ne rejette pas l'hypothèse nulle de stationnarité de la série.

3.4.3 Identification et estimation des modèles potentiels

Comme démontré au chapitre précédent, la série chronologique $Delta_{12}DeltaY_t = Y_t - Y_{t-12}$ est stationnaire.
Les modèles candidats sont les modèles de type AR(p), MA(q) et ARMA(p,q).
L'autocorrélogramme partiel empirique de la série $Delta_{12}DeltaY_t$ permet d'éliminer les modèles AR(p) et MA(q) car leur forme n'est pas identifiable.

On déduit donc que $Delta_{12}DeltaY_t$ suit un ARMA (p,q), donc la série initiale du logarithme des restes à charge ln(RAC) suit un modèle SARIMA(p,d,q)(P,D,Q)[s], car cette série a été stationnarisée suite à intégration et à stationnarisation.

Dans la suite nous travaillerons sur la série ln(RAC).

Il nous faut maintenant identifier les coefficients (p,d,q)(P,D,Q)[s] du modèle SARIMA.

L'analyse de la fonction d'autocorrélation empirique et de l'autocorrélation partielle empirique montre des pics significatifs au décalage 1 ce qui suggère la présence d'une composante MA(1). De plus par différentiation, nous avons appliqué un filtre moyenne mobile d'ordre 1 pour éliminer la marche aléatoire et un filtre d'ordre 12 pour éliminer la saisonnalité.
On en déduit comme estimateurs des paramètres d,D et S : d = 1, D=1, S =12.

Premier modèle candidat : SARIMA(1,1,1)(1,1,1)[12]
De ce qui précède, nous allons dans un premier temps tester le modèle SARIMA(1,1,1)(1,1,1)[12].

Nous utilisons la fonction *arima()*, du package *forecast* disponible sous R.
Cette fonction permet d'identifier les valeurs des paramètres qui maximisent la vraisemblance. Elle donne les mêmes résultats que les estimations des paramètres à l'aide des moindres carrés.

Le modèle identifié permet d'obtenir les coefficients ar(1), ma(1), sar(1) et sma(1) comme le montre la Table 3.10 ci-dessous.

Test de significativité des paramètres
Le test de Student de significativité des paramètres ci-dessous nous montre que seuls les coefficients sar(1) et ma(1) sont significatifs au seuil de 5 %.

paramètre	Estimate	Std. Error	z value	Pr(>|z|)
ar(1)	-0.23217	0.19499	-1.1907	0.23379
ma(1)	-0.61428	0.13885	-4.4240	9.69e-06 ***
sar(1)	0.52782	0.25365	2.0809	0.03744 *
sma(1)	-0.99853	0.60766	-1.6433	0.10033

TABLE 3.10 – Test de Wald de significativité des coefficients

Nous allons tester le modèle suivant en supprimant le coefficient le moins significatif ar(1), soit le modèle SARIMA(0,1,1)(1,1,1)[12].

Nous allons procéder pas à pas, en éliminant le coefficient le moins significatif, puis en testant la significativité globale des paramètres à l'aide du test de Student jusqu'à l'obtention des modèles qui nous donnent des coefficients totalement significatifs à 5%.

Mesure de la qualité globale des modèles
Quatre modèles ont été implémentés suite à la procédure de sélection définie ci-dessus en Table 3.11.
Afin de les comparer, nous allons mesurer la qualité globale des modèles proposés à partir des critères d'informations AIC et BIC, car ils permettent d'avoir un bon équilibre biais-variance en pénalisant les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie.
Le critère de choix final sera le critère AIC corrigé car il intègre donc une pénalité supplémentaire pour les paramètres additionnels permettant d'éviter un sur-ajustement par rapport à l'AIC.

Numéro	Modèle	AICc	BIC	Paramètres sign.	Paramètres non sign.
1	SARIMA(1,1,1)(1,1,1)[12]	-110.42	-102.63	ma(1), sar(1)	ar(1),sma(1)
2	SARIMA(0,1,1)(1,1,1)[12]	-111.63	-105.18	ma(1), sar(1),sma(1)	
3	SARIMA(0,1,1)(0,1,1)[12]	-113.1	-108.11	ma(1)	sma(1)
4	SARIMA(0,1,1)(0,1,0)[12]	-114.35	-110.92	ma(1)	

TABLE 3.11 – Test de Wald de significativité des coefficients

Parmi les modèles candidats, seuls les modèles 2 et 4 ont tous leurs paramètres qui sont significatifs.
Nous conserverons le modèle 4 car ayant tous ses paramètres significatifs c'est celui qui minimise l'AIC et le BIC.
Soit le modèle SARIMA (0,1,1)(0,1,0)[12].

Sélection automatique du modèle ARIMA
Nous allons comparer le meilleur modèle choisi manuellement soit le SARIMA (0,1,1)(0,1,0)[12], avec celui renvoyé par l'algorithme de Hyndman-Khandakar qui maximise le critère AIC dans le choix de modèle et développé sur R via la fonction *auto.arima()*.

La fonction *auto.arima()* nous renvoie exactement le même modèle que celui choisi manuellement soit le SARIMA (0,1,1)(0,1,0)[12].

3.4.4 Validation du modèle retenu

Afin de valider le modèle retenu nous allons tester que les résidus obtenus sont un bruit-blanc et sont normalement distribués.

Analyse visuelle des résidus

Le graphique ci-dessous (Figure 3.40) montre 3 représentations graphiques différentes des résidus :

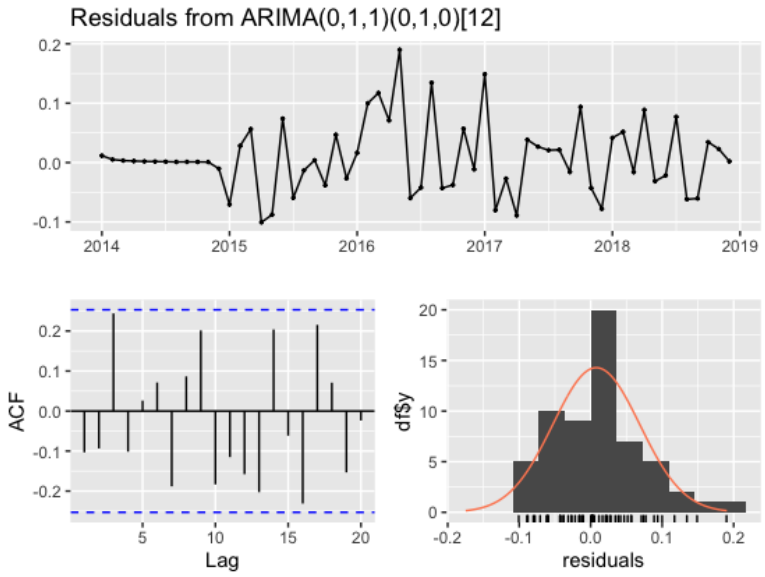

FIGURE 3.40 – Analyse des résidus SARIMA (0,1,1)(0,1,0)[12]

— Le chronogramme des résidus :
Cette représentation des résidus du modèle (partie non modélisable) au cours du temps est centré autour de 0 et ne présente pas de schéma particulier (tendance ou périodicité notable), ce qui est une caractéristique visuelle attendue pour un bruit blanc.
— L'autocorrélogramme empirique :
L'autocorrélogramme des résidus montre qu'il n'y a pas de corrélation entre les résidus pour tout décalages h choisi, en effet, tout les $\hat{\rho}$(h) sont compris dans l'intervalle de confiance à 95% de nullité des coefficients.
— L'ajustement des résidus à une loi normale :
Il est visuellement difficile de conclure à une normalité des résidus en raison de la sur-representation des résidus sur la gauche du graphique.
Un autre test sera nécessaire pour évaluer graphiquement la normalité des résidus.

Comme on peut l'observer sur le QQ-plot ci-dessous, la plupart des quantiles théoriques de la loi normale et ceux de notre distribution s'alignent sur la droite de Henry.
Certains points s'éloignent cependant de la droite pour des niveaux élevés.

Si l'analyse visuelle nous permet de conclure à l'absence d'autocorrélation au niveau des résidus, elle

ne nous permet pas de conclure sur la normalité de ces derniers.
Nous allons analyser des tests statistiques.

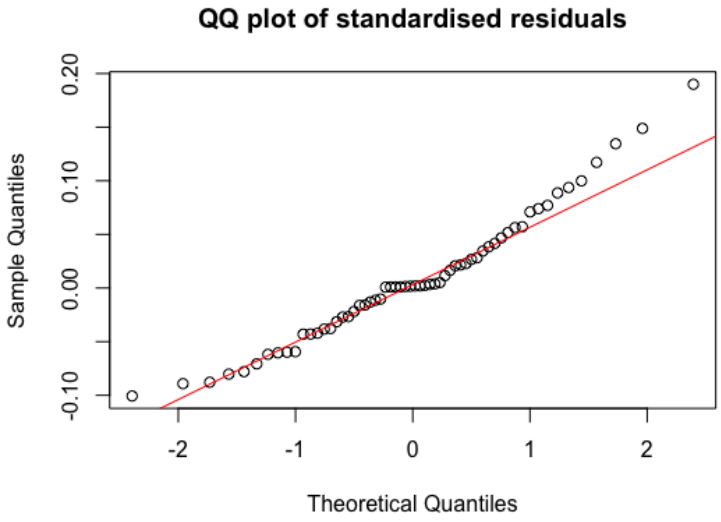

FIGURE 3.41 – Analyse des résidus SARIMA (0,1,1)(0,1,0)[12]

Tests statistiques de non corrélation et normalité des résidus
Nous allons utiliser la statistique Ljung-Box Q (LBQ) au décalage 12. Cette statistique permet de tester l'hypothèse nulle selon laquelle les autocorrélations jusqu'au décalage 12 sont égales à zéro soit que les valeurs sont aléatoires et indépendantes jusqu'au décalage 12.
Afin de tester l'hypothèse de normalité des résidus, nous utiliserons le test d'adéquation à la loi normale de Shapiro–Wilk qui teste l'hypothèse nulle selon laquelle les résidus sont normalement distribués.

Type de test	Statistique	P-Value
Ljung-Box Q (Blancheur)	17.617	0.0909
Shapiro-Wilk (Normalité)	10.96681	0.1017

TABLE 3.12 – Test du portemanteau et de normalité

Pour le modèle retenu SARIMA(0,1,1)(0,1,0)[12], la p-value = 0.0909 du test de blancheur est supérieure au seuil α =0.05. On ne rejette pas l'hypothèse nulle, les résidus ne sont pas correlés.
De même le test de normalité de Shapiro-Wilk conclut à ne pas rejeter l'hypothèse nulle car la P-value est supérieure au seuil α =0.05.

3.4.5 Analyse de la prévision

Sur notre échantillon de test (données 2019), nous allons réaliser une prévision à partir du modèle retenu et la comparer avec les autres modèles qui ont été éliminés au cours de cette démarche de modélisation.
L'analyse a postériori permet de quantifier les écarts entre les prévisions et les réalisations. Nous allons utiliser les modèles les plus significatifs retenus pour chacune des méthodes de modélisation pour estimer le RAC de l'année 2019.

Les performances ont été évaluées sur la base de trois mesures : l'erreur absolue moyenne (MAE), l'erreur absolue moyenne en pourcentage (MAPE) et l'erreur quadratique moyenne (MSE).

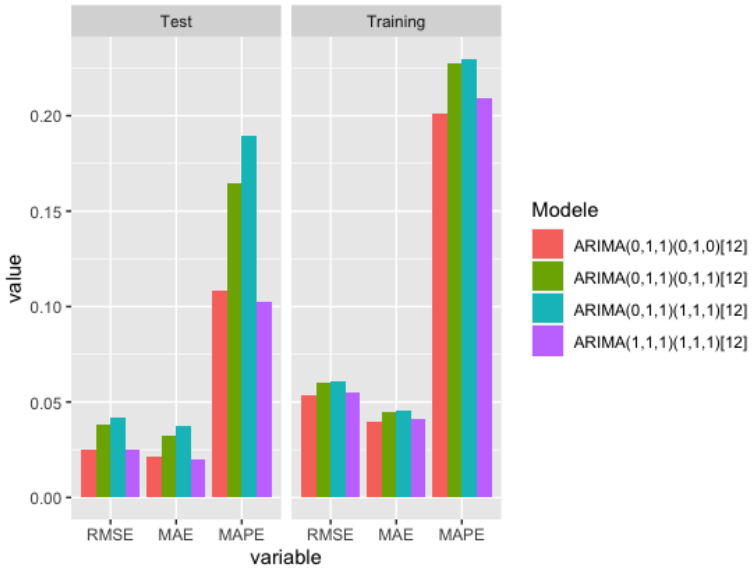

FIGURE 3.42 – Comparaison des RMSE, MAE et MSE des modèles SARIMA.

Le modèle SARIMA(1,1,1)(1,1,1)[12] est celui qui minimise les critères de RMSE, MAE et MSE sur le modèle de prédiction.

Dans le tableau ci-dessous, nous avons à partir des estimations des logarithmes, des montants de reste à charge de pharmacie donnés par le modèle SARIMA(1,1,1)(1,1,1)[12] et SARIMA (0,1,1)(0,1,0)[12], calculé les montants de dérive annuelle glissante sur 1 an.
Le modèle SARIMA(1,1,1)(1,1,1)[12] s'ajuste mieux que le modèle SARIMA (0,1,1)(0,1,0)[12].

Date	Est.Sarima	Ln(RAC 2019)	Ln(RAC 2018)	Dérive.réelle	Dérive.estimée
*Jan*2019	19,706	19,715	19,774	-0,300 %	-0,344 %
*Feb*2019	19,606	19,616	19,66	-0,226 %	-0,273 %
*Mar*2019	19,656	19,63	19,714	-0,426 %	-0,293 %
*Apr*2019	19,598	19,652	19,648	0,017 %	-0,256 %
*May*2019	19,591	19,588	19,663	-0,380 %	-0,364 %
*Jun*2019	19,601	19,566	19,651	-0,435 %	-0,255 %
*Jul*2019	19,603	19,621	19,66	-0,249 %	-0,337 %
*Aug*2019	19,426	19,442	19,493	-0,263 %	-0,340 %
*Sep*2019	19,529	19,55	19,555	-0,025 %	-0,132 %
*Oct*2019	19,673	19,684	19,747	-0,316 %	-0,371 %
*Nov*2019	19,577	19,545	19,647	-0,520 %	-0,357 %
*Dec*2019	19,615	19,638	19,649	-0,054 %	-0,172 %
TotalAnnuel	235,183	235,246	235,87	-0,265 %	-0,291 %

TABLE 3.13 – Dérive estimée par SARIMA (1,1,1)(1,1,1)[12]

Cette dérive autour de - 0,3 % est proche de celle observée par l'assureur sur son portefeuille de contrat (Figure 1.29)

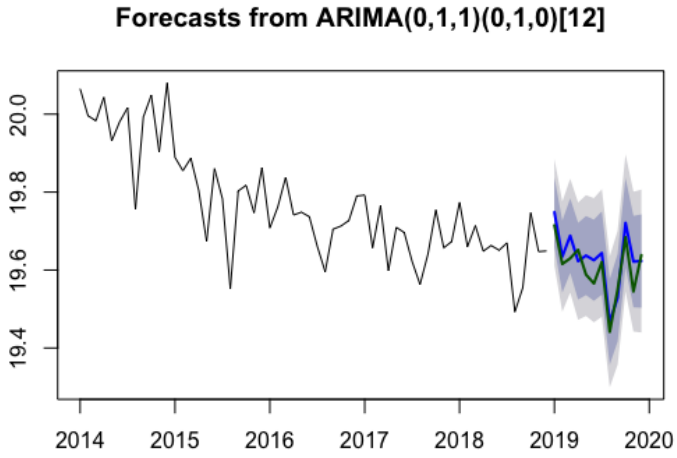

FIGURE 3.43 – Prévisions 2019 et valeurs réelles des restes à charge en pharmacie.

Comme le montre la figure 3.43 les vraies valeurs de la dérive se retrouvent toujours dans l'intervalle de confiance à 95% donné par le modèle.

3.4.6 Conclusion sur la modélisation SARIMA

Nous venons de voir que le modèle SARIMA était particulièrement bien adapté pour estimer le montant de reste à charge Sécurité Sociale pour les dépenses en pharmacie. En effet, parmi les 4 modèles étudiés, les modèles SARIMA (1,1,1)(1,1,1)[12] et SARIMA (0,1,1)(0,1,0)[12] nous donnaient un intervalle de confiance de prédiction à 5% qui contenait les valeurs réelles observées sur les bases 2019.
Le modèle avec toutes les variables significatives au sens de Student, n'a pas été celui qui nous a donné la meilleure prédiction.
La difficulté principale a été de choisir le meilleur des modèles, c'est à dire celui qui optimise les critères biais-variance tout en étant meilleur sur la prédiction.

En conclusion, le modèle SARIMA présente des avantages dans ses propriétés statistiques, bien connues et son processus de modélisation efficace. Il peut être facilement réalisé à l'aide de logiciels statistiques courants, tels que SAS et R. Il peut être utilisé lorsque les séries chronologiques saisonnières sont stationnaires et ne comportent aucune donnée manquante. L'inconvénient du modèle SARIMA est qu'il ne permet que la modélisation des relations linéaires dans les données de la série chronologique.

3.5 Conclusion sur la modélisation par séries temporelles

Dans cette partie, nous allons comparer les modèles de séries temporelles utilisés pour projeter les restes à charges sur l'année 2019.

La table 3.14 est le reflet des montants de restes à charges estimés.
La table 3.16 est le reflet des dérives estimées par les différents modèles. Dans ces tables, le modèle SARIMA est le SARIMA (0,1,1)(0,1,0)[12].

On constate que le modèle de lissage exponentiel est celui qui a le meilleur pouvoir de prédiction, avec une dérive annuelle de -5% qui est proche de la dérive réelle sur les données brutes (-5.09%).
Certaines dérives mensuelles sont plus marquées dans le modèle lissage exponentiel par rapport au réel. Pour exemple, en janvier, le modèle de lissage exponentiel prévoit une baisse très importante (-11.3%) contre une baisse réelle (-5.75%).
L'erreur moyenne (ME) et l'erreur quadratique (RSME) du modèle de lissage exponentiel sont les plus faibles.

	RAC 2018	RAC 2019	RAC 2019	RAC 2019	RAC 2019
Mois	Réel	LissExpo.	Décompose	SARIMA	Réel
Janvier	387 026 948	343 194 482	348 366 400	377 197 695	364 754 694
Février	345 355 096	328 935 678	331 506 482	336 584 570	330 334 539
Mars	364 467 583	344 631 897	354 640 292	355 213 650	335 132 572
Avril	341 361 178	324 194 842	322 811 684	332 692 766	342 512 666
Mai	346 336 764	319 591 209	324 060 213	337 541 829	321 412 702
Juin	342 224 362	332 540 020	335 291 371	333 535 545	314 202 066
Juillet	348 614 971	320 529 223	320 954 582	339 763 382	331 973 860
Aout	292 139 350	280 686 990	282 249 821	284 720 911	277 566 188
Septembre	310 846 081	322 135 568	328 655 710	302 953 534	309 301 984
Octobre	376 612 259	345 616 749	343 970 647	367 046 829	353 822 867
Novembre	340 996 881	323 636 830	323 108 605	332 336 975	307 873 728
Décembre	341 496 437	344 570 728	351 032 402	332 825 870	337 871 383
Total	4 137 477 910	3 930 264 216	3 966 648 209	4 032 413 556	3 926 759 249

Table 3.14 – Tableau comparatif des projections de reste à charges des différentes méthodes

Mesure	LissaExpo.	Décompose	SARIMA
ME	-292 081	-3 324 080	-8 804 526
RMSE	12 618 041	18 488 362	33 468 005

Table 3.15 – Mesure de l'erreur de prédiction des RAC

Mois	DER 2019 (LissExpo.)	DER 2019 (Décompose)	DER 2019 (SARIMA)	DER 2019 (réelle)
Janvier	-11.3%	-9.99%	-2.5397%	-5.75%
Fevrier	-4.8%	-4.01%	-2.5396%	-4.35%
Mars	-5.4%	-2.70%	-2.5390%	-8.05%
Avril	-5.0%	-5.43%	-2.5394%	0.34%
Mai	-7.7%	-6.43%	-2.5394%	-7.2%
Juin	-2.8%	-2.03%	-2.5389%	-8.19%
Juillet	-8.1%	-7.93%	-2.5391%	-4.77%
Aout	-3.9%	-3.39%	-2.5393%	-4.99%
Septembre	+3.6%	+5.73%	-2.5391%	-0.5%
Octobre	-8.2%	-8.67%	-2.5399%	-6.05%
Novembre	-5.1%	-5.25%	-2.5396%	-9.71%
Decembre	+0.9%	+2.79%	-2.5390%	-1.06%
Annuelle	-5%	-4.13%	-2.5393%	-5.09%

TABLE 3.16 – Tableau comparatif des projections de dérives des différentes méthodes

3.6 Estimation par GLM

Les méthodes d'estimation par lissage exponentiel et ARIMA présentées sont univariées, elles s'intéressent uniquement à l'étude de l'évolution de la variable RAC dans le temps. Elles n'étudient pas l'influence des variables exogènes sur la variable d'intérêt.

Nous avons vu au chapitre 1 que la consommation médicale était influencée par plusieurs facteurs. En particulier que celle-ci augmentait avec l'âge, était plus importante pour les individus de sexe féminin, et que les niveaux de consommation variaient en fonction des régions en raison de la disponibilité de l'offre de soin.

On sait que l'un des facteurs discriminants de la consommation médicale est également le niveau de garanties proposé dans les contrats. Cette variable n'a pas été croisée avec les bases assureurs pour la ressortir dans les bases DAMIR car l'objet de notre étude est la modélisation des soins en pharmacie. Pour la pharmacie ce critère est moins significatif.
Au delà d'avoir une estimation du reste à charge, l'assureur souhaite également comprendre comment la structure démographique de son portefeuille, les garanties des assurés, leur ancienneté... peuvent influencer la dérive.

L'assureur peut par exemple, s'intéresser à savoir quel est le niveau de dérive relatif à une classe d'âge, à une CSP ... d'où l'importance d'avoir plusieurs méthodes d'estimation de cette dérive et notamment des méthodes ayant un pouvoir explicatif plus important que celles proposées par les séries temporelles univariées.

Dans cette partie, nous nous proposons de modéliser la consommation médicale à partir des variables exogènes mises à notre disposition. Notre choix s'est porté sur un modèle de type GLM, pour sa relative simplicité de mise en oeuvre, il est souvent utilisé dans la définition du tarif, possède un fort pouvoir explicatif et est bien documenté dans la littérature.

3.6.1 Retraitement des données

Les données que nous avons sont celles issues des différents retraitements effectués au chapitre 1.

Pour les années de 2014 à 2019, nous avons les données sur l'année, le mois de prestation, l'âge du bénéficiaire, la nature de la prestation, le sexe du bénéficiaire, la région du bénéficiaire.
Dans la base et pour chaque année nous avons l'information si les individus ont eu une prise en charge. Nous sommes donc en présence de données longitudinales répétées. Les informations relatives aux groupes d'individus ne changeant pas d'une année sur l'autre, nous aurons donc des individus corrélés entre-eux ce qui pourra générer des résidus.
Dans notre situation, étant donnée que seule la variable RAC apporte de l'information supplémentaire dans l'année, nous allons transposer les informations afin de déplacer la corrélation sur les individus au niveau des variables.

Création de la variable dérive par année
Nous introduisons de nouvelles variables explicatives que sont les restes à charges pour les années 2014, 2015, 2016, 2017,2018, 2019.
Puis nous créerons notre variable d'intérêt qui est la dérive :

$$Der_{annee} = (RAC_{annee}/RAC_{annee-1}) - 1$$

par exemple :

$$Der_{2018} = (RAC_{2018}/RAC_{2017}) - 1$$

pour un individu.

Retraitement de la variable Région
Les modalités de la variable "Région" et "Age" ne sont pas toutes bien représentées. Certaines modalités ont donc été regroupées afin d'être significatives dans le modèle.

Afin d'appliquer notre modèle de regression, nous avons fait le choix d'agréger les individus par caractéristiques identiques. Ainsi, tous les individus de sexe féminin, résidant en région parisienne dans la tranche d'âge de 0 à 19 ans, ont été regroupés en un seul individu présentant ces caractéristiques. Le montant global de dépenses et de remboursements ont été sommés par année de soin.
Nous remarquerons que pour cette étude, la maille d'analyse a changé : on est plus dans l'étude de la projection/prévision mensuelle du RAC, mais dans l'étude d'une prévision annuelle.
Nous obtenons ainsi une nouvelle base de données avec 180 observations et 9 variables.

3.6.2 Choix de la fonction de lien

La variable réponse est ici le taux de dérive observé en 2018 (Der2018).
Nous avons centré et réduit les variables Der2015, Der2016, Der2017, Der2018 afin de ramener les variables explicatives sur la même échelle et d'éviter les inégalités de traitements entre les coefficients.

Ajustement à une loi Normale
Les données étant centrées sur 0, notre première loi candidate est la loi normale $N(0, 1)$.
Ci-dessous l'histogramme des variables centrées et réduites :

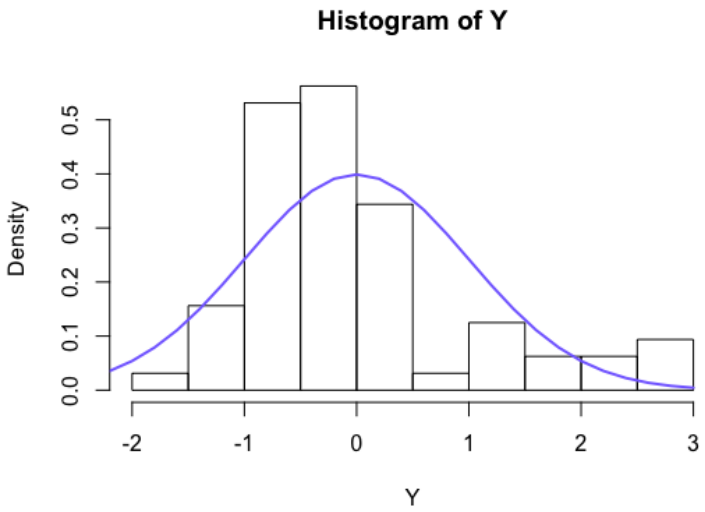

FIGURE 3.44 – Prévisions 2019 et valeurs réelles des restes à charge en pharmacie.

L'observation des Figure 3.44 et Figure 3.45 nous fait comprendre que le jeu de données ne suit pas une loi normale.
On constate une forte représentation du taux de dérive entre -2 et 0.

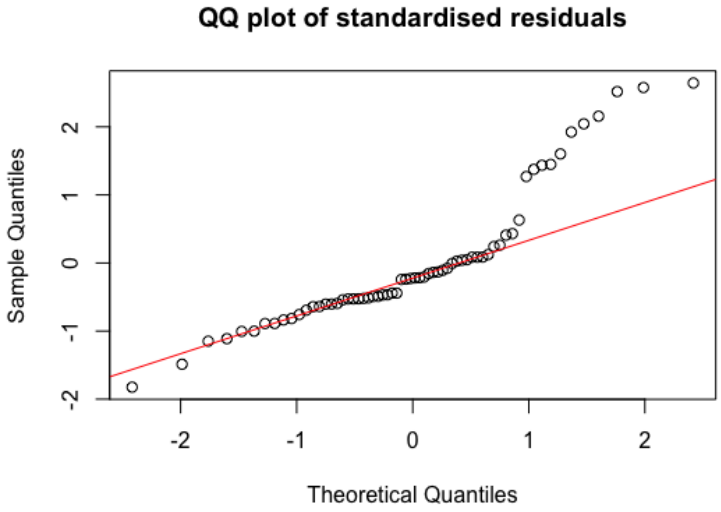

FIGURE 3.45 – QQ-Plot dérive centrée réduite.

Test de normalité
Nous réalisons néanmoins un test d'adéquation de Kolmogorov et Smirnov. la P-Value renvoyée est de 0.009644 qui est inférieur à 0,5% et donc on rejette l'hypothèse de normalité de cette distribution.

Ajustement à une loi Gamma
La seconde loi candidate est la loi Gamma.

Rappel sur la loi Gamma :
On dit que X suit la loi gamma de paramètres k $>$ 0 et $\lambda > 0$, notée , si elle admet pour densité

$$f(x) = \frac{\lambda}{\Gamma(k)} e^{-\lambda x} (\lambda x)^{k-1} \mathbf{1}_{\mathbb{R}_+}(x).$$

Avec k, le paramètre de forme et $\lambda = \frac{1}{\theta}$, le paramètre d'intensité.
- L'espérance définit par : E(X)=$k\theta$
- La variance définit par : V(X)=$k\theta^2$
- Le Skewness définit par : Skewness = $\frac{2}{\sqrt{k}}$

Avec la moyenne empirique de l'échantillon $\hat{\mu} = 2$, sa variance empirique de $\hat{\sigma} = 1$, son skewness égal à 1.153534.
On identifie $\hat{k} = 3.006070709$ et $\hat{\theta} = 1.503035355$.

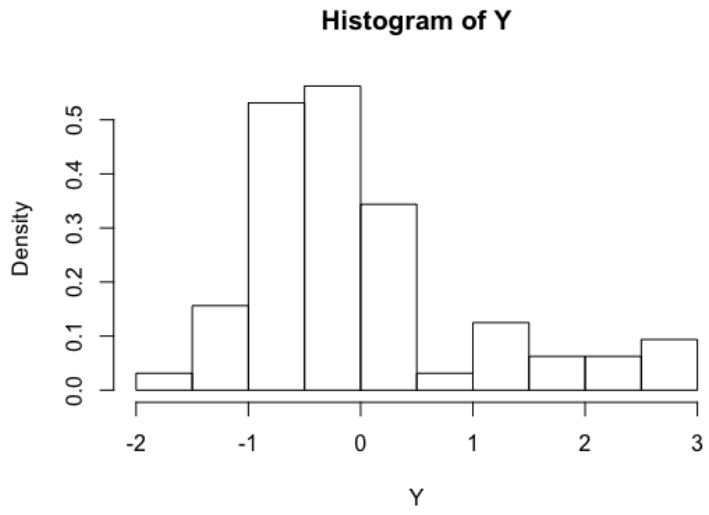

FIGURE 3.46 – Histogramme du taux de dérive 2018 réduite.

Le Skewness associé à ce jeu de données est de 1.153534, soit une distribution décalée à gauche de la médiane, et donc une queue de distribution étalée vers la droite.
Nous déduisons de même, le coefficient d'aplatissement : le Kurtosis qui est de 1.153534.
Afin de tester si nous sommes en présence d'une loi gamma-normale, nous réalisons un changement d'échelle en décalant la moyenne de 2 unités.
La nouvelle distribution ainsi obtenue est la suivante :

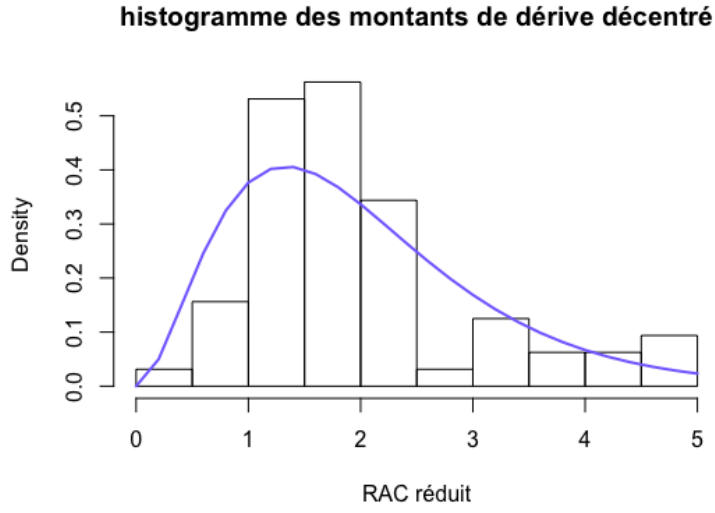

FIGURE 3.47 – Histogramme de la dérive ajusté de la loi gamma.

Visuellement, cette densité s'ajuste mieux que celle de la loi normale ci-dessus.
On réalise un test d'adéquation de la loi $\Gamma(3, 0.7)$ avec $\lambda = \frac{1}{1.503}$ =0.7.

Test d'adéquation à la loi Gamma

Le résultat du test d'adéquation de notre jeu de données à la $\Gamma(3, 0.7)$, par Kolmogorov et Smirnov est le suivant : la P-Value du test est égale à 0.1197, on ne rejette pas l'hypothèse que la loi de notre jeu de données soit une loi $\Gamma(3, 0.7)$.

On peut utiliser la loi Gamma comme fonction de lien pour notre modèle GLM.

On va dans un second temps observer l'ajustement des quantiles empiriques et des quantiles théoriques de la loi $\Gamma(3, 0.7)$.
Sur le QQ-Plot ci-dessous, nous constatons que les quantiles s'ajustent bien à la droite de Henry.

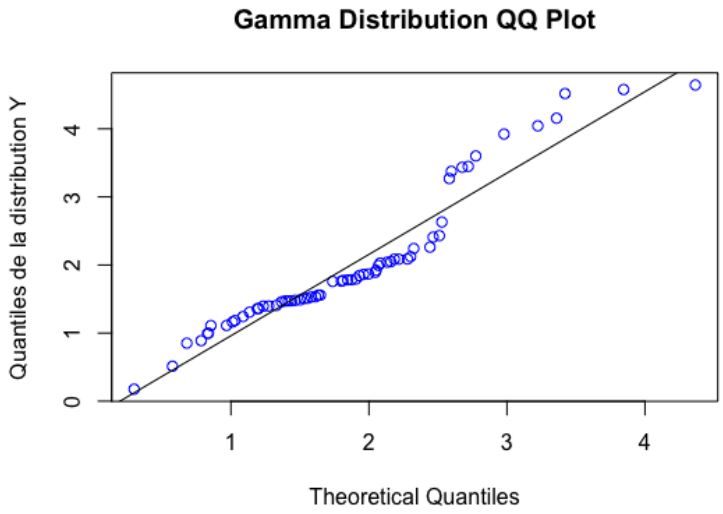

FIGURE 3.48 – QQ-Plot loi Gamma.

3.6.3 Estimation des paramètres

Nous avons lancé notre modèle GLM avec la fonction de lien Gamma.
Ci-dessous les coefficients estimés par la fonction *glm()* du logiciel R :

Coefficients	Estimate Std.	Error	t value	Pr(>|t|)
BEN-RES-REGCentre-Est	-0.0009414	0.0019945	-0.472	0.639011
BEN-RES-REGDOMTOM	0.0053537	0.0046363	1.155	0.253795
BEN-RES-REGEst	-0.0023651	0.0019607	-1.206	0.233523
BEN-RES-REGMediterranné	0.0018381	0.0031506	0.583	0.562289
BEN-RES-REGNord	-0.0021171	0.0015528	-1.363	0.178997
BEN-RES-REGOuest	-0.0001753	0.0014790	-0.119	0.906153
BEN-RES-REGIdF	-0.0007653	0.0020702	-0.370	0.713205
BEN-SEX-CODMasculin	0.0037263	0.0007916	4.707	2.10e-05***
AGE-BEN-SNDS20-59ans	0.0107281	0.0014876	7.212	3.12e-09***
AGE-BEN-SNDS60-79ans	0.0148635	0.0018469	8.048	1.62e-10***
AGE-BEN-SNDS80ansET+	0.0168286	0.0019101	8.810	1.13e-11***
Der2015	0.0052513	0.0050385	1.042	0.302419
Der2016	-0.0920179	0.0339008	-2.714	0.009143**
Der2017	-0.1654649	0.0452037	-3.660	0.000616***

TABLE 3.17 – Coefficients estimés de la regression

Test de significativité des paramètres
Le test de Student de significativité des coefficients renvoyé par le tableau montre que seules les variables : $Der2016$, $Der2017$, $BEN-SEX-CODMasculin$, $AGE-BEN-SNDS20-59ans$, $AGE-BEN-SNDS60-79ans$, $AGE-BEN-SNDS80ANSET+$, sont significatifs au seuil 5%.
La région n'a pas d'effet sur la dérive 2018.
Les paramètres estimés ont les mêmes ordres de grandeur.

3.6.4 Analyse de la multicolinéarité des variables explicatives

Pour pouvoir estimer correctement notre modèle, nous devons nous assurer de l'absence de colinéarité parfaite entre les variables explicatives. Par construction, il existe une forte correlation entre les variables explicatives liées à la dérive $Der2015$, $Der2016$, $Der2017$.
Cependant, des variables correlées ne sont pas forcément colinéaires même si l'inverse est vrai.
Nous allons nous assurer de l'absence de colinéarité dans notre modèle en examinant les facteurs d'inflation de la variance (FIV).
L'indicateur FIV renseigne sur l'augmentation de la variance d'un coefficient en raison de la présence d'une relation de linéarité avec un autre prédicteur.

Variables	GFIV	Dégré de liberté	$GFIV^{1/(2\times Df)}$
BEN-RES-REG	24.817990	7	1.257842
BEN-SEX-COD	1.221378	1	1.105160
AGE-BEN-SNDS	5.908394	3	1.344554
Der2015	17.642028	1	4.200241
Der2016	2.403716	1	1.550392
Der2017	4.068605	1	2.017078

TABLE 3.18 – facteurs d'inflation de la variance modèle complet

La variable d'intérêt ici est $GFIV^{1/(2\times Df)}$ correspond à la valeur du FIV corrigé des degrés de liberté présents dans les variables qualitatives. Dans ce tableau tous les FIV sont supérieures à 1. Il faut donc analyser la multicolinéarité.

Il n'y a pas de consensus sur la valeur du FIV au-dessus de laquelle on peut considérer qu'il y a colinéarité. Certains auteurs comme Allison 2012 disent qu'il y a colinéarité ou qu'il faut commencer à investiguer pour une valeur de FIV supérieure à 2,5.
Avec une valeur à 4.2 pour la variable $Der2015$, nous concluons donc à la présence de multicolinéarité dans notre modèle.

Méthode d'élimination de la multicolinéarité

Il existe, dans la littérature, plusieurs actions possibles pour éliminer la multicolinéarité d'un modèle telles que : la sélection de variables selon un critère, la mise en oeuvre d'une regression pénalisée de type regression ridge ou lasso.
Nous testerons uniquement la sélection de variables, car elle permettra d'éliminer notre multicolinéarité et nous permettra d'ajuster notre modèle.

3.6.5 Sélection et validation du modèle

La méthode choisie est la sélection pas à pas de variables qui maximisent l'AIC.
Nous utiliserons la méthode stepwise car elle est plus optimale pour notre jeu de données.
Son principe est le suivant : partant d'un modèle avec uniquement l'intercept, on rajoute la variable la plus significative dans le modèle, puis réexamine les tests de Student pour chaque variable explicative déjà présente dans le modèle. Si certaines variables ne sont plus significatives, elles sont retirées du modèle.
L'algorithme fonctionne ainsi par itération jusqu'à ce qu'aucune variable ne puisse être ajoutée ou retirée sans dégrader l'AIC.
Le modèle ainsi obtenu est le suivant :
$Der2018 = "BEN - SEX - COD" + "AGE - BEN - SNDS" + "Der2016" + "Der2017"$
avec des paramètres estimés suivants :

Min	1Q	Median	3Q	Max
-0.0145376	-0.0036056	0.0000368	0.0036239	0.0127798

Table 3.19 – Coefficients estimés de la regression

Coefficients	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	0.072625	0.321019	0.226	0.8218
$BEN - SEX \times CODMasculin$	-0.015349	0.003244	-4.732	1.51e-05***
$AGE - BEN - SNDS20 - 59ans$	-0.043572	0.005760	-7.565	3.64e-10***
$AGE - BEN - SNDS60 - 79ans$	-0.060416	0.006812	-8.869	2.53e-12***
$AGE - BEN - SNDS80ansET+$	-0.068407	0.006929	-9.872	6.02e-14***
$Der2016$	0.331791	0.124901	2.656	0.0102*
$Der2017$	0.675721	0.160387	4.213	9.07e-05***

Table 3.20 – Coefficients estimés de la regression

Le test de Student de significativité des coefficients nous permet de conclure que toutes les variables et modalités retenues sont significatives.
On constate également que les coefficients des variables Der2017 et Der2016 qui représentent la dérive 2017 et 2016, ont un poids important dans la modélisation.
Ce qui n'est pas étonnant et montre bien que la dérive d'une année est liée à celle des autres années.

Mesure de l'information apportée par les variables
Par construction la procédure *Stepwise* nous a sélectionné un modèle avec un AIC plus faible que le modèle.

Robustesse du modèle
Indépendance des covariables
Afin de s'assurer qu'il n'y a plus de multicolinéarité dans le modèle, nous examinons de nouveau le facteur d'inflation de la variance (FIV) sur le modèle sélectionné.

Variables	GFIV	Dégré de liberté	$GFIV^{1/(2*Df)}$
$BEN - SEX - COD$	1.144416	1	1.069774
$AGE - BEN - SNDS$	4.127070	3	1.266505
$Der2016$	1.809141	1	1.345043
$Der2017$	2.895268	1	1.701549

TABLE 3.21 – Facteurs d'inflation de la variance dans le modèle sélectionné

On constate que tous les VIF sont inférieures à 2,5.
Nous pouvons conclure à une indépendance des variables dans le modèle.

Déviance du modèle sélectionné
La déviance résiduelle du modèle sélectionné est plus faible et nous indique dans quelle mesure la variable de réponse peut être prédite par le modèle spécifique que nous adaptons avec p variables prédictives. Ici la déviance résiduelle est plus faible que la déviance nulle.

Résidus du modèle
Plus de 95% des résidus sont dans la bande d'acceptation entre -2 et 2, ce qui permet de dire que peu de valeurs sont aberrantes. Ceci laisse penser que le modèle est robuste.

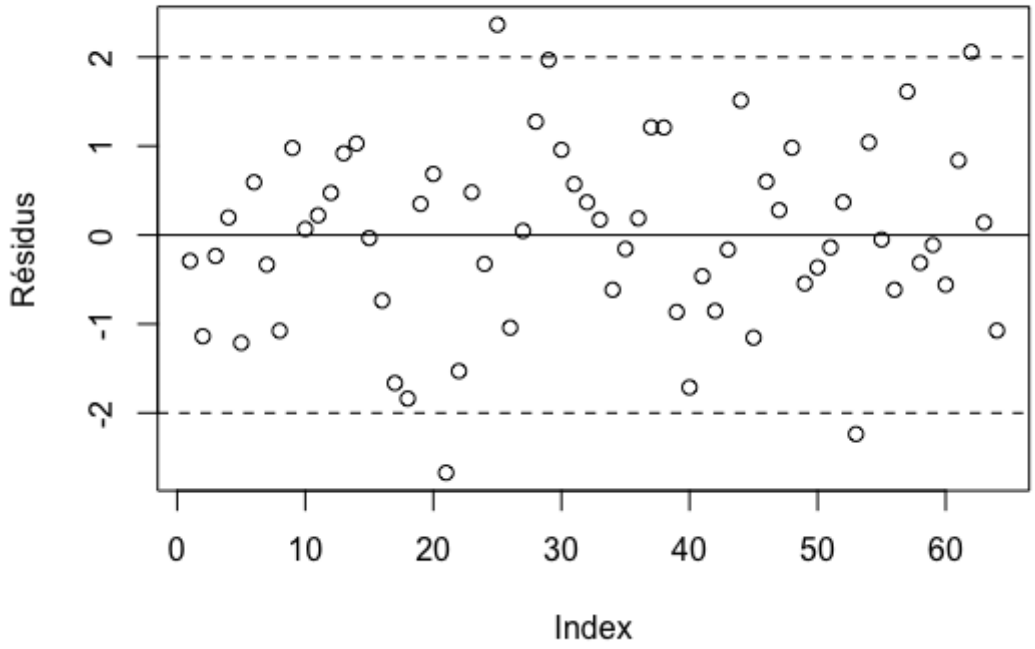

FIGURE 3.49 – Résidus du modèle GLM

Prédiction

Afin de tester le modèle, nous avons estimé sur la base 2019, la dérive attendue :

$Der2019_{estime} = -0.015349 \times (Masculin) - 0.043572 \times (Ag20-59) - 0.060416 \times (Ag60-79) - 0.068407 \times (Ag80) + 0.331791 \times (Der2017) + 0.675721 \times (Der2018)$.

Avec :
Masculin : une variable binaire qui vaut 0 lorsque l'individu est de sexe masculin.
$Ag20-59$: une variable binaire qui vaut 0 lorsque l'individu est âgé de 20 à 59 ans.
$Ag60-79$: une variable binaire qui vaut 0 lorsque l'individu est âgé de 60 à 79 ans.
$Ag80$: une variable binaire qui vaut 0 lorsque l'individu est âgé de plus de 80 ans.

Ci-dessous les résultats obtenus :
Nous obtenons une erreur moyenne de ME = -0,019 et RMSE =0,06.
La dérive obtenue sur tout le portefeuille est de -0,289%.

3.6.6 Conclusion sur le modèle GLM

Le modèle GLM construit nous a permis de regarder l'influence des variables exogènes sur la dérive en pharmacie.
On a pu ainsi constater que la dérive d'une année pouvait être exprimée en fonction de celle des 2 années précédentes mais également du sexe et de l'âge des individus dans la population générale.
La contribution au modèle de la dérive (n) et de la dérive (n-1) est positive et les critères démogra-

phiques tels que le sexe et l'âge viennent ajuster à la baisse cette contribution.
Si sur l'année 2019, le pouvoir prédictif est satisfaisant, sur une année exceptionnelle comme l'année 2020, le modèle ne tient pas la route.
Dans ce genre de modèle, l'impact des années exceptionnelles (avec une forte dépense à la hausse ou à la baisse) ne sera pas capté puisque le modèle lissera les données. Dans ce type de situation la qualité de préparation et de corrections des données en amont sera encore plus forte pour proposer un modèle de qualité.

Un des risques principaux observés lors des modélisations de type GLM sont les risques de modèles dû à :

— L'incapacité à modéliser les relations non linéaires entre les covariables :
  Les modèles GLM vont s'attacher à modéliser les relations linéaires entre les différentes covariables. Les effets non linéaires des covariables qui peuvent apporter de l'information ne seront donc pas utilisés.
  Par ailleurs, le fait de supposer systématiquement un lien linéaire entre les variables peut mener à des modèles fallacieux.
— Modèles avec un apriori sur la fonction de lien,
— L'absence de contrôles sur les résidus pour valider le modèle.

Afin de s'affranchir des risques de paramétrages, un modèle de type Générale Additive Modèle (GAM) qui ne fait pas d'apriori sur la loi ni sur le type de lien entre variable peut être utilisé en complément.
Afin de s'affranchir des risques de paramétrages, un modèle de type Générale Additive Modèle (GAM) qui ne fait pas d'a priori sur la loi ni sur le type de lien entre variables peut être utilisé en complément.

Conclusion

La mesure du risque est le coeur du métier d'un actuaire. En santé, les primes appelées sont obtenues par l'espérance des prestations futures ceci montre l'importance pour les organismes complémentaires d'assurance santé de prévoir la sinistralité de leur portefeuille. Cette sinistralité autrement dit la consommation de soins et biens médicaux a beaucoup évoluée ces dernières années. On observe une dérive de sinistralité causée par l'évolution des récentes réglementations et de facteurs environnementaux.

L'objectif de ce mémoire est de proposer différents outils de modélisation pour estimer cette dérive de sinistralité.

La première étape a consisté en la définition du périmètre d'étude. Bien que nous disposions de complètes bases de données assureur, celles ci se déclinaient seulement sur 2 exercices de survenance.
La possibilité donnée de pouvoir utiliser les Open Data telles que la DAMIR, a été très utile pour mettre en oeuvre notre étude. La première difficulté que l'on a rencontrée est qu'il s'agit d'une base nationale. Il a donc fallu la calibrer et bien ajuster ses données au portefeuille étudié.
Nous avons fait le choix réfléchi de la pharmacie qui est l'un des domaines de la santé à n'avoir subi aucune ou très peu d'évolution règlementaire susceptible d'affecter la série chronologique des coûts de santé car nous voulions comparer les méthodes sans se préoccuper des retraitements lourds et des ajustements de modèles qu'il faut mettre en oeuvre lorsqu'on modélise un domaine très volatil tel que le dentaire ou l'optique.

La phase de modélisation fut assez longue car notre objectif était de proposer un large éventail et d'explorer le champs des possibles en matière de modélisation du risque.
Les modèles retenus dans cette étude sont :
— La modélisation par décomposition simple
— La modélisation par lissage exponentiel
— La modélisation par Box-Jenkins
— La modélisation par GLM

Les conclusions sur ces modèles sont les suivantes :

Sur la modélisation par décomposition simple :
La décomposition simple peut être fastidieuse et ne garantit pas toujours un résidu de type bruit blanc à la fin quand elle est réalisée manuellement car pour chaque composante il faut de tester plusieurs paramètres pour estimer précisément la tendance et la saisonnalité exacte de la série. Le risque de modèle est important puisqu'on choisit une loi à priori pour caractériser les différentes composantes de la série.
La décomposition simple implémentée par les logiciels notamment les fonctions *decompose()* et *stl()* du logiciel R sont plus facile à mettre en œuvre et donnent des résultats satisfaisants mais ne permettent pas de caractériser précisément la loi sous-jacente de la série, ce qui dans le cadre d'une prédiction

n'est pas nécessaire.

Sur la modélisation par lissage exponentiel :
Les modèles de lissage exponentiel sont basés sur la description de la nature de la tendance et de la saisonnalité. Il s'agit de modèles simples à mettre en œuvre mais très peu testés. L'inconvénient majeur de cette méthode est qu'elle donne plus d'importance aux observations récentes.

Sur la modélisation par Box-Jenkins :
Les modèles de type ARMA vont plus loin et proposent de modéliser l'autocorrélation dans les données. Le modèle SARIMA présente des avantages dans ses propriétés statistiques bien connues et son processus de modélisation est efficace. Il peut être utilisé lorsque les séries chronologiques saisonnières sont stationnaires et ne comportent aucune donnée manquante. L'inconvénient du modèle SARIMA est qu'il ne permet que la modélisation des relations linéaires dans les données de la série chronologique et ne permet pas de rajouter des variables exogènes qui peuvent permettre de comprendre la chronologie de la série. Une alternative aux modèles ARMA sont les modèles de type ARMAX qui admettent des variables explicatives.

Sur la modélisation par GLM :
Le GLM est facilement mise en œuvre car largement utilisé en assurance santé. Dans ce type de modélisation où l'on souhaite prédire une valeur future à partir des données passées il y a une attention particulière à avoir sur le calibrage et la validation des modèles car le risque que les résidus soient autocorrelés est important.
Une alternative aux GLM serait d'utiliser le GAM (Modèle Additif Généralisé) qui relâche l'hypothèse de linéarité entre la variable à expliquer et les variables explicatives. Il est de même possible d'utiliser des modèles de régression non paramétriques. Une telle démarche permettrait de faire diminuer le risque de modèle en prenant des hypothèses moins restrictives.

Parmi les modèles de séries temporelles testés sur nos données, il en ressort que celui qui à le meilleur pouvoir prédictif est le modèle de lissage exponentiel. Ces modèles nous ont permis d'identifier un profil de dérive infra-annuelle. Pour l'assureur, ce type de modèles a du sens lors d'exercices prévisionnels infra-annuels. Par exemple, la projection annuelle à partir de données semestrielles.
Le modèle GLM reste intéressant pour l'assureur lors des exercices de tarifications ou encore pour mesurer l'impact de phénomènes exogènes sur la dérive (exemple : anticipation d'inflation, réformes réglementaires...). Dans ce cas, le modèle GLM est le modèle le plus efficient.

Enfin, le choix de la bonne dérive à appliquer au portefeuille dépend non seulement des modèles qui viennent éclairer la décision de l'actuaire mais également de tous les acteurs : les experts, le marché, l'anticipation de réforme, la volonté d'équilibrer certains comptes... De même, nous devons nous montrer agiles et avoir des modèles suffisamment flexibles, pour permettre de s'ajuster rapidement.

Notre mémoire a porté sur l'étude de la dérive en pharmacie, mais elle peut se généraliser à d'autres postes en faisant attention à bien retraiter les données notamment lorsque les montants des frais passés ont fluctué avec des évolutions réglementaires ou ont eu des phénomènes de rupture. Nous pensons notamment au dentaire et à l'optique qui sont 2 postes inflatés.

Bibliographie

Bibliographie

Allison, Paul (2012). https://statisticalhorizons.com/multicollinearity/.

AMELI (2020). *Site internet de l'Assurance Maladie*. https://assurance-maladie.ameli.fr/etudes-et-donnees/open-damir-depenses-sante-interregimes-2020. (Page consultée le 30 septembre 2021).

Charpentier, Arthur (2013). "ACT2040 Partie - Modèles linéaires généralisés." In : Université de Québec à Montréal, p. 30. URL : http://freakonometrics.free.fr/slides-2040-4.pdf.

DRESS (2020). *Rapport 2020 sur la situation financière des organismes complémentaires assurant une couverture santé*. URL : https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-12/rapport-oc-2020.pdf.

DRESS' (2020). *LES DÉPENSES DE SANTÉ EN 2019*. URL : https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-10/infographie-cns2020.pdf.

Eilstein, Daniel (2019). "Séries temporelles et modèles de régression". In : 240, p. 53. URL : https://www.santepubliquefrance.fr/content/download/185725/document_file/30172_1890-d527.pdf.

Hyndman, Rob J et George Athanasopoulos (2018). *Forecasting : Principles and Practice*. URL : https://otexts.com/fpp2/stationarity.html.

IPSOS (2020). *Les préoccupations des Français*. URL : https://www.ipsos.com/fr-fr/fractures-francaises-face-aux-crises-qui-frappent-le-pays-un-besoin-de-protection-plus-fort-que.

ROCHE, Angelina (2018-2019). "Lissage exponentiel." In : 3, p. 26. URL : https://www.ceremade.dauphine.fr/~roche/Enseignement/Series_temps_exMaster/SeriesTemp_Cours3.

ROCHE, Angelina' (2018-2019). "Processus stationnaires – modèles ARMA." In : 2, p. 34. URL : https://www.ceremade.dauphine.fr/~roche/Enseignement/Series_temps_exMaster/SeriesTemp_Cours2.

ROCHE, Angelina" (2018-2019). "Modèles ARIMA et SARIMA, prédiction et choix de modèle." In : 4, p. 25. URL : https://www.ceremade.dauphine.fr/~roche/Enseignement/Series_temps_exMaster/SeriesTemp_Cours4_slides.