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

le 16 Décembre 2020

Par : Anne-Emmanuelle ADOU
Titre : Modélisation du risque hospitalisation sur les données *open source* de la Sécurité sociale

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Marine HABART
Yann QUERE
Renaud CAILLET
Signature :

Membres présents du jury de l'EURIA :
Pierre AILLIOT
Signature :

Entreprise :
Sia Partners
Signature :

Directeur de mémoire en entreprise :
Nicolas SERVAN
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Le faible taux de prise en charge des soins hospitaliers par les complémentaires santé pose des limites à l'appréhension du risque hospitalisation. Pourtant, ce risque reste stratégique dans le cadre du pilotage des régimes frais de santé puisqu'il représente un peu plus d'un tiers des prestations servies par les organismes complémentaires.

Dans la perspective d'améliorer la connaissance de ce risque, ce mémoire propose d'analyser la sinistralité observée à l'échelle nationale en s'appuyant sur les données *open source* de la Sécurité sociale, retraçant la consommation de soins hospitaliers par profil d'individu.

Une première partie s'attachera à décrire le contexte de prise en charge de l'hospitalisation en France, ainsi que les tendances de consommation des soins hospitaliers issues de l'analyse des données. Ensuite, afin d'évaluer l'apport du *machine learning* dans le cadre de la fiabilisation des tarifs, trois stratégies tarifaires basées sur les modèles linéaires généralisés (GLM) seront mises en place :

1. Une tarification globale du risque sans segmentation des actes hospitaliers qui servira de *benchmark* pour l'évaluation des performances des tarifs concurrents ;
2. Une tarification selon un regroupement des actes en sous-postes de soins usuels ;
3. Une tarification selon une segmentation des actes découlant d'un algorithme de classification non supervisé : HDBSCAN.

Ainsi, cette dernière tarification a vocation à challenger le schéma actuel de modélisation du risque hospitalisation pour une meilleure adéquation risque-tarif. En effet, celle-ci vise à proposer une nouvelle segmentation de modélisation optimisant l'homogénéité des groupes d'actes formés en matière de fréquence de consommation et de coût moyen.

Mots clefs: hospitalisation, *open source*, tarification, GLM, classification non supervisée, HDBSCAN.

Abstract

Because supplementary health insurance organizations only cover a small percentage of the costs associated with hospital care, their understanding of the risk of hospitalization is limited. However, understanding this risk is important in the context of healthcare cost schemes because hospitalization costs accounts for slightly more than a third of the benefits provided by supplementary health insurance organizations.

In order to improve knowledge of this risk, this paper analyses the claims experience observed on a national scale by using Sécurité sociale data, tracing the consumption of hospital care by individual profile.

The first part will describe hospital care coverage context in France, as well as the trends in the consumption of hospital care resulting from the analysis of the data. Then, in order to evaluate the use of machine learning in predicting hospital care pricing, three pricing strategies based on Generalised Linear Models (GLM) will be put in place :

1. Global risk pricing without segmentation of hospital procedures, which will serve as a benchmark for evaluating the performance of competing rates ;
2. Pricing according to a classification of the acts into sub-categories of usual care ;
3. Pricing according to a segmentation of acts resulting from an unsupervised classification algorithm : HDBSCAN.

The third pricing strategy is intended to challenge the current hospitalization risk modeling scheme to ensure a better risk/price match. Indeed, it aims to propose a new modeling segmentation that optimizes the homogeneity of the categories of care formed in terms of frequency of consumption and average cost.

Keywords: hospitalization, open source, pricing, GLM, unsupervised classification, HDBSCAN.

Note de synthèse

En France, l'hospitalisation représente 37%[1] des prestations servies par les complémentaires santé. Ce poids important renferme toutefois une réalité contrastée : les complémentaires santé ne connaissent que 5,2% du risque réel car il est particulièrement encadré et pris en charge par la Sécurité sociale.

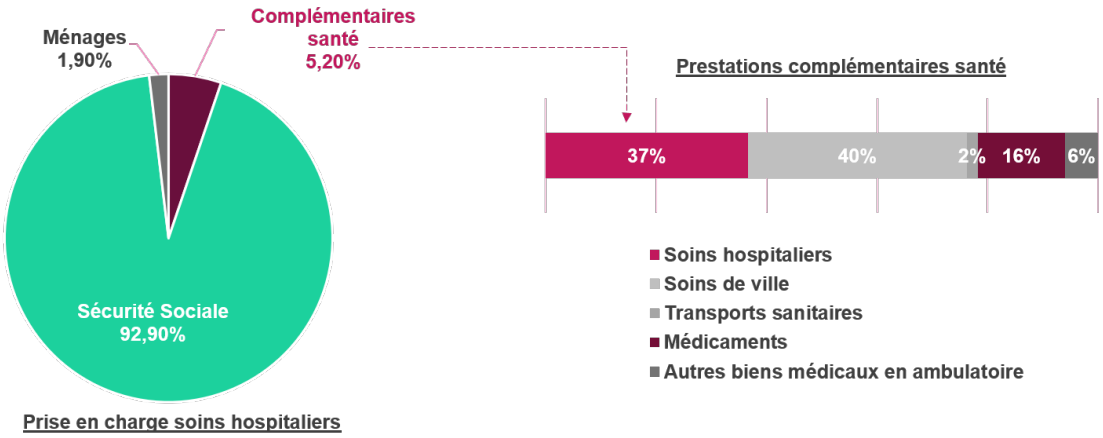


FIGURE 1 – Place des complémentaires santé dans la prise en charge des soins hospitaliers

Cet état de fait pose ainsi la question de la correcte adéquation risque-tarification sur cette garantie, une interrogation devenant primordiale en cas de désengagement de la Sécurité sociale. En effet, cette possibilité est à réellement envisager et anticiper face au déficit important de l'Assurance Maladie et aux mutations fréquentes de législation que connaît le système de santé francais.

Dans ce contexte, la mise à disposition par l'Assurance Maladie de données *open-source* sur l'hospitalisation, tous régimes confondus, représente une opportunité pour une meilleure compréhension du risque et l'évaluation de l'optimalité des pratiques actuelles de tarification. Ainsi, ce mémoire a exploité ces données afin de répondre à trois objectifs :

— analyser les tendances de sinistralité au regard des profils des assurés ;

— évaluer l'efficacité d'une segmentation des actes hospitaliers en sous-postes de soins usuels dans le cadre d'une tarification ;

— exploiter un algorithme de *machine learning* afin de proposer une segmentation en groupes d'actes de distributions plus homogènes, permettant d'aboutir à une meilleure tarification.

1. DRESS 2018

Analyse de l'impact des facteurs de risque

Afin de préserver l'anonymat des populations, les bases d'hospitalisation mises à disposition par l'Assurance Maladie sont des bases agrégées dont les lignes représentent un profil donné d'individus selon diverses variables caractéristiques.

L'analyse préliminaire de ces données a révélé que seules les données du secteur privé étaient disponibles. Les conclusions de ce mémoire ne concernent donc que le domaine privé. De plus, pour des questions de volumétrie, l'étude s'est cantonnée aux données de 2019.

Quatres facteurs de risque ont ainsi pu être analysés, à savoir : le sexe de l'individu, sa région de résidence, sa tranche d'âge et son affiliation à la CMU [2]. Ici, le sexe n'a vocation qu'à permettre de mieux comprendre le risque mais ne pourrait en aucun cas être utilisé en pratique comme variable de tarification comme le prévoit la règlementation. Quant à l'affiliation à la CMU, elle est considérée comme un indicateur de niveau de vie des individus.

Face à ces données de sinistralité, une base d'assurés a été construite à partir des données démographiques de l'Institut National de la Statistique et des Etudes Economiques (INSEE), afin d'obtenir l'exposition correspondante à chaque profil de la base sinistres. Les données de l'INSEE étant ventilées uniquement par région, sexe et âge, une segmentation supplémentaire selon l'affiliation à la CMU a été construite sur la base des caractéristiques de la population CMU contenues dans le rapport d'activité 2018 du Fonds CMU.

Les principales conclusions des analyses sont les suivantes :

— La fréquence de consommation de soins hospitaliers croît avec l'âge. Toutefois, les soins sollicités par les populations les plus jeunes ont un coût moyen plus élevé que ceux des plus âgés. Ces constats sont cohérents avec une détérioration attendue de l'état de santé des assurés avec l'âge. De plus, les actes exploratoires de détection de maladies, particulièrement coûteux, sont majoritairement réalisés chez les jeunes tandis que les personnes âgées sollicitent en quantité des soins curatifs ou palliatifs de maladies déjà diagnostiquées ; d'où un coût moyen plus faible.

— Les départements d'Outre-mer ont des coûts moyens très élévés par rapport aux autres régions et une fréquence particulièrement faible. Un constat cohérent avec la cherté de la vie enregistrée dans ces régions et la jeunesse de leur population.
L'Occitanie se démarque également avec une fréquence de consommation paticulièrement élevée, découlant d'une population viellissante traduit par un indice de vieillessement [3] de 101,13 pour une moyenne nationale de 83,23.

— Les hommes affichent un coût moyen 20% plus élevé que celui des femmes mais ont une fréquence plus faible (-14%) s'expliquant entre autres par une population masculine plus jeune.

— Enfin, l'affiliation à la CMU n'influence pas le niveau des coûts des soins hospitaliers. Toutefois, étant sur un périmètre restreint au secteur privé, la population CMU a moins recours aux soins hospitaliers, du fait de la faiblesse de ses revenus et de l'absence de couverture des dépassements d'honoraires par le Fonds CMU.

2. Désormais Complémentaire santé solidaire depuis le 1er novembre 2019
3. Nombre d'adultes de 65 ans et plus pour un jeune de moins de 20 ans - données de population INSEE 2019

Tarification selon une segmentation par sous-poste de soins usuels

Usuellement, les actes hospitaliers sont regroupés en trois sous-postes de soins. Ils concernent les frais de séjour (FSEJ), les frais liés aux chambres particulières (CHBR) et divers autres frais (AUTR) n'entrant pas dans les deux premières catégories. Ainsi, la majorité des actes hospitaliers sollicités concerne le sous-poste de soins FSEJ qui représente alors 98,78% du coût des sinistres et 86,4% du volume de sinistres.

Sous-poste de soins	Nombre d'actes	Nombre d'actes (%)	Dépense engagée	Dépense engagée (%)	Nombre de sinistres	Nombre de sinistres (%)	Coût moyen	Fréquence moyenne
FSEJ	239	87%	32 459 483 209	99%	234 505 031	86,4%	138	3,50
CHBR	3	1%	55 640 598	0%	1 558 493	0,6%	36	0,02
AUTR	32	12%	345 710 845	1%	35 369 637	13,0%	10	0,53
Total	274		32 860 838 973		271 433 161			

FIGURE 2 – Bilan de la consommation des actes par sous-postes de soins

Une tarification du risque selon cette classification a été réalisée. Pour ce faire, la base de données a été segmentée en base d'apprentissage et base de test. Les travaux étant effectués sur une seule année, une sélection aléatoire des données de 10 mois de soins parmi les 12 a été réalisée pour constituer la base d'apprentissage et les 2 mois restants la base de test. Pour éviter tout biais, à chaque modélisation, la stabilité des coefficients a été vérifiée sur toutes les combinaisons possibles de base d'apprentissage/base de test possibles. Une modélisation classique coût-fréquence par les modèles linéaires généralisés (GLM) a ensuite été appliquée. Les résultats sont les suivants :

	FSEJ	CHBR	AUTR	Tarification 2	Tarification 1
Prime pure moyenne observée	460,94	0,85	5,03	466,82	466,82
Prime pure moyenne prédite	490,13	0,82	5,25	496,20	496,17
Ecart	6,334%	-2,886%	4,222%	6,294%	6,287%

FIGURE 3 – Résultat modèle de coût-fréquence - tarification 1 et 2

Si elle répond à une logique médicale ou opérationnelle, cette segmentation n'améliore pas la modélisation tarifaire du risque. En effet, la tarification par sous-poste de soins (notée "tarification 2") ne permet pas une meilleure prédiction de la dépense engagée comparativement à une tarification sans segmentation (notée "tarification 1"). Cette absence de gain fait pourtant suite à la réalisation de six modèles contre deux modèles dans le cas de la tarification globale.

Ce constat vient confirmer la nécessité de reconsidérer la segmentation par sous-poste de soins usuellement admise, ce que nous proposons de faire par la suite.

Proposition d'une segmentation alternative

Les résultats prédécents ont motivé la construction d'une nouvelle segmentation plus performante. Pour ce faire, un algorithme de classification non supervisée, HDBSCAN, a été exploité afin de construire des groupes d'actes homogènes. Les variables de segmentation retenues ont été la moyenne et l'écart-type des distributions de coût moyen et de fréquence. La méthodologie suivie a ensuite été la suivante :

— segmentation des actes selon leur distribution de coût moyen, d'une part, et leur leur distribution de fréquence, de l'autre ;

— intersection des deux segmentations afin d'obtenir des *clusters* homogènes en termes de coût-moyen et de fréquence ;

— ajustement des *clusters* bruts obtenus pour créer des cluters plus robustes : 4 groupes ont ainsi été constitués.

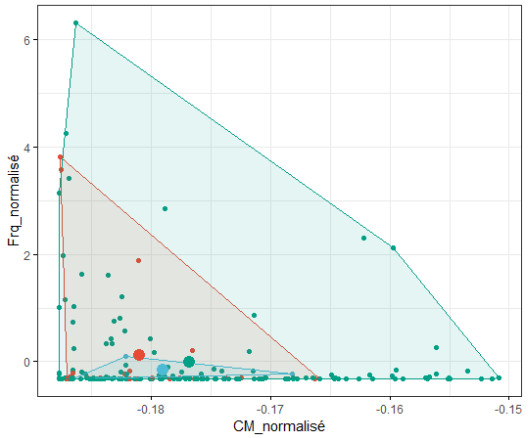


FIGURE 4 – Segmentation par sous-poste

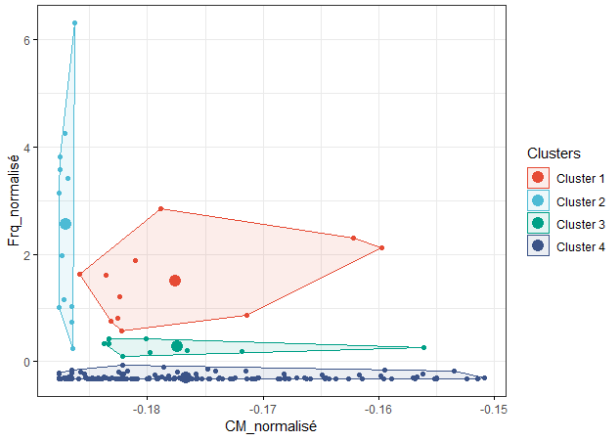


FIGURE 5 – Segmentation HDBSCAN

Comparativement à la segmentation par sous-poste, cette segmentation (notée "tarification 3") met en exergue quatre profils d'actes : les actes avec un coût très élevé et une fréquence élevée (*cluster* 1), les actes avec un coût moyen faible mais une fréquence élevée (*cluster* 2), les actes avec un coût moyen peu élevé mais une fréquence faible (*cluster* 3), et les actes avec un coût moyen élevé mais une fréquence faible (*cluster* 4).

Clusters	Nombre d'actes	Nombre d'actes (%)	Dépense engagée	Dépense engagée (%)	Nombre de sinistres	Nombre de sinistres (%)	Coût moyen	Fréquence moyenne
1	18	7%	28 966 914 563	88,2%	129 921 135	47,9%	222,96	1,94
2	12	4%	280 294 730	0,9%	107 932 535	39,8%	2,60	1,61
3	10	4%	1 175 355 269	3,6%	18 246 874	6,7%	64,41	0,27
4	234	85%	2 438 274 411	7,4%	15 332 617	5,6%	159,03	0,23
Total	274		32 860 838 973		271 433 161			

FIGURE 6 – Caractéristiques des *clusters* formés

Tarification selon la nouvelle segmentation proposée

La réalisation d'une tarification par GLM (contrainte de l'étude) sur ces nouveaux *clusters* permet d'obtenir une dépense totale prédite supérieure de 6% à la dépense observée.

Si ce niveau est proche de celui de la tarification par sous-poste, le gain majeur réside dans l'amélioration des prédictions de sinistralité individuelle. En effet, tandis que la tarification par sous-poste conduit à des ratios sinistres à primes (S/P [4]) s'écartant en moyenne [5] de 12 points du niveau d'équilibre (100%), les nouveaux *clusters* réduisent cet écart à 9,6 points. Une concentration plus importante de ces ratios autour de 100% est alors observée avec des niveaux de dépassements moindres que dans les deux précédentes tarifications comme le montre le graphique 7.

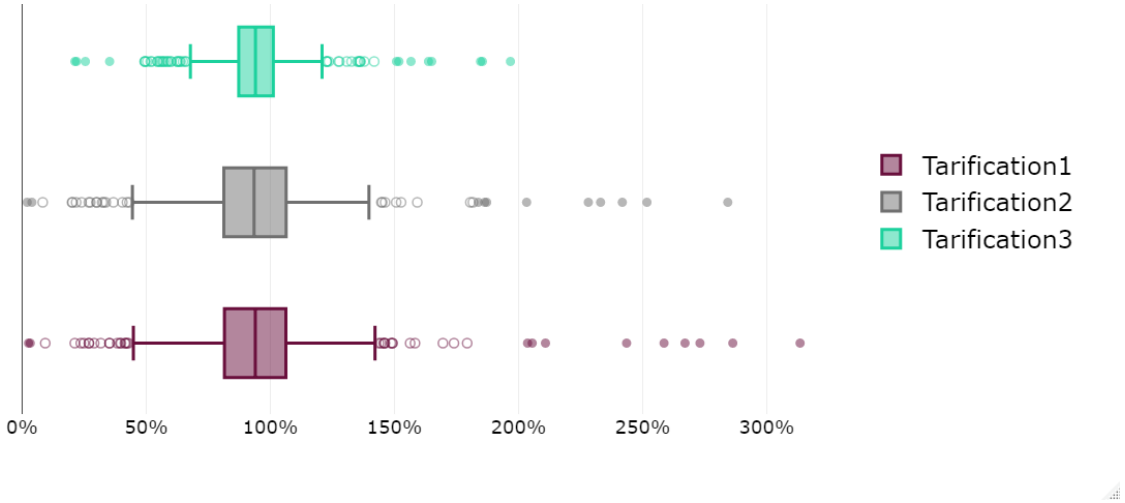


FIGURE 7 – Dispersion des ratios sinistres à primes selon tarification

Par ailleurs, plus que 28% des profils ont une sinistralité sous-estimée (S/P>100%) contre 37% dans la tarification 2. Une erreur de tarification représentant plus de 30% de la dépense engagée, chutant à 17% après application de la nouvelle segmentation proposée, soit une réduction de 45% de la perte enregistrée.

Ainsi, cette dernière tarification vient améliorer l'adéquation risque-tarification et la rentabilité du portefeuille.

4. Dépense engagée observée / dépense engagée prédite
5. Moyenne pondérée par l'exposition de la valeur absolue des écarts à 100%

	Part profils avec S/P > 100%	Dépense engagée pour les profils avec S/P >100%		Ecart moyen des S/P à 100%	S/P maximal
		Total (en euro)	Part de la dépense engagée totale (%)		
Tarification 1	37,7%	1 847 894 409	35,46%	12,74%	313,44%
Tarification 2	37,0%	1 652 856 821	31,72%	12,18%	284,31%
Tarification 3	28,1%	908 104 601	17,43%	9,60%	196,74%

FIGURE 8 – Comparaison des résultats des prédictions selon tarification

Conclusion

L'étude menée cherchait à observer la possibilité d'utiliser un algorithme de *machine learning*, HDBSCAN, afin de mettre à l'épreuve une modélisation de la sinistralité hospitalisation à une maille sous-postes.

S'appuyant sur les données *open source* fournies par l'Assurance Maladie, l'absence de segmentation des actes a servi de référentiel de comparaison.

La ventilation test est obtenue après le *clustering* des actes par moyenne et écart-type des coûts moyens et fréquences. Sur ces axes, les *clusters* créés affichent des caractéristiques plus homogènes. L'adéquation tarifaire par profil perd ainsi en volatilité (-25%) après avoir remanié en 4 groupes les actes usuellement séparés en 3 et permet, par ricochet, une réduction de 45% des pertes dues à une sous-tarification du risque.

Les résultats présentés sont tenus par les hypothèses prises et moyens à disposition pour l'étude, à savoir l'utilisation d'une base agrégée, impliquant une perte d'information individuelle, ainsi que l'utilisation des GLM, modèles usuellement utilisés opérationnellement. De plus, les résultats dépendent des paramètres de *clustering* testés, notamment de la distance initiale de répartition des points et des variables de segmentation. Sur ce dernier point, de nouvelles variables jugeant de la pertinence d'adéquation des densités des actes à des lois adaptées aux GLM, pourraient compléter l'analyse.

L'apport de cette méthodologie, sur l'exemple développé, laisse ainsi présager un gain significatif si elle devait être déclinée sur les segmentations de sinistralité opérationnellement réalisées par les assureurs. Certaines garanties présentant des distributions similaires pourraient ainsi être modélisées ensemble tandis que des garanties dont les distributions seraient inadaptées pourraient être re-segmentées.

Executive Summary

In France, hospitalisation accounts for 37% [6] of the benefits provided by supplementary health insurance. However, this significant weighting contains a contrasting reality : supplementary health insurance only covers 5.2% of the real risk because it is particularly supervised and paid for by Sécurité Sociale.

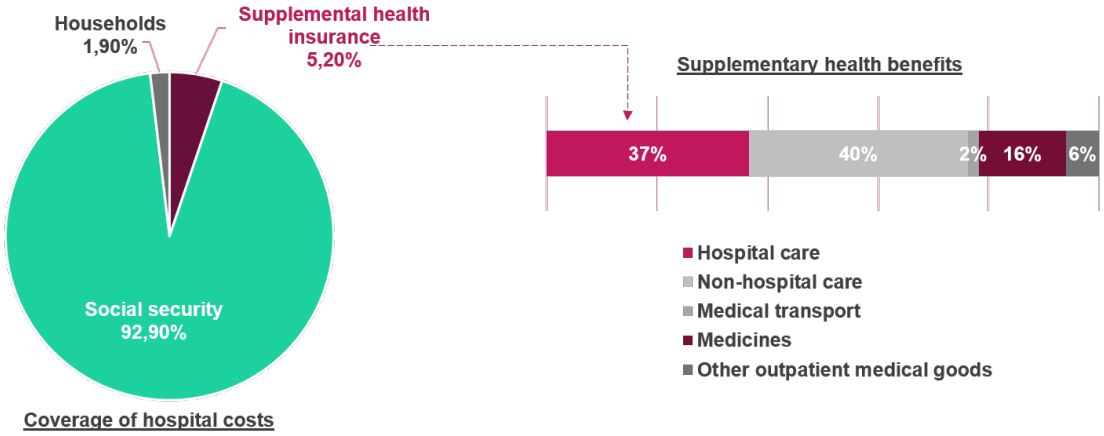


FIGURE 9 – The contribution of complementary health insurance to hospital care coverage

This raises the question of the correct risk/pricing adequacy of this guarantee, a question that becomes crucial in the event of withdrawal from the Sécurité Sociale system. Indeed, this possibility should be strongly considered and anticipated in view of the significant deficit of the Assurance Maladie and the frequent changes in legislation that the French health system is undergoing.

In this context, the provision of open-source data on hospitalisation, all schemes included, by the Assurance Maladie, represents an opportunity for a better understanding of risk and the evaluation of the optimality of current pricing practices. Thus, this paper has exploited this data to meet three objectives :

— analysing claims trends in relation to the profiles of policyholders ;

— assess the effectiveness of segmenting hospital procedures into the usual subcategories of care in the context of a pricing scheme ;

— use a machine learning algorithm to propose a segmentation into groups of acts with more homogeneous distributions, resulting in better pricing.

6. DRESS 2018

Analysis of the impact of risk factors

In order to preserve the anonymity of populations, the hospitalisation databases made available by the Assurance Maladie are aggregated databases whose lines represent a given profile of individuals according to various characteristic variables.

Preliminary analysis of these data revealed that only private sector data were available. The conclusions of this brief therefore relate only to the private sector. Furthermore, for volumetric reasons, the study was conducted only with data from 2019.

Four risk factors could thus be analysed, namely : the individual's gender, region of residence, age group and CMU affiliation [7]. Here, gender is only intended to provide a better understanding of the risk but could in no way be used in practice as a pricing variable as provided for in the regulations. As for CMU membership, it is considered as an indicator of the standard of living of individuals.

In view of these claims data, a database of policyholders was built up using demographic data from the French National Institute for Statistics and Economic Studies (INSEE), in order to have the corresponding exposure for each profile in the claims database. As INSEE data is only broken down by region, gender and age, an additional segmentation according to CMU membership was constructed on the basis of the characteristics of the CMU population contained in the CMU Fund's 2018 activity report.

The main conclusions of the analyses are as follows :

— The frequency of hospital care consumption increases with age. However, the average cost of the care requested by the youngest populations is higher than that of the oldest.
These findings are consistent with an expected deterioration in the state of health of insured persons with age. In addition, exploratory disease detection procedures, which are particularly costly, are mostly carried out among young people, while many older people seek curative or palliative care for illnesses that have already been diagnosed ; hence the average cost is lower.

— Overseas departments have very high average costs compared to other regions and a particularly low frequency. This observation is consistent with the high cost of living recorded in these regions and the youthfulness of their populations.
Occitanie also stands out with a particularly high frequency of consumption, resulting from an ageing population translated by an ageing index [8] of 101.13 for a national average of 83.23.

— Men have an average cost 20% higher than that of women but have a lower frequency (-14%) due to a younger male population, among other reasons.

— Finally, CMU membership does not influence the level of hospital care costs. However, as it is limited to the private sector, the CMU population has less recourse to hospital care, due to its low income and the lack of coverage of fee overruns by the CMU Fund.

7. Complémentaire santé solidaire since 1 November 2019
8. Number of adults aged 65 and over for a young person under 20 years old - INSEE 2019 population data

Performance of segmentation by care sub-category

Usually, hospital acts are grouped into three sub-categories of care. They relate to accommodation costs (FSEJ), costs related to private rooms (CHBR) and various other costs (AUTR) that do not fall under the first two categories. Thus, the majority of the hospital acts requested concern the FSEJ care sub-item, which represents 98.78% of the cost of claims and 86.4% of the volume of claims.

Care sub-category	Number of cares	Numbers of cares (%)	Expenditure incurred	Expenditure incurred (%)	Numbers of claims	Numbers of claims (%)	Average cost	Average frequency
FSEJ	239	87%	32 459 483 209	99%	234 505 031	86,4%	138	3,50
CHBR	3	1%	55 640 598	0%	1 558 493	0,6%	36	0,02
AUTR	32	12%	345 710 845	1%	35 369 637	13,0%	10	0,53
Total	274		32 860 838 973		271 433 161			

FIGURE 10 – Assessment of the consumption of procedures by care sub-items

A risk rating according to this classification has been carried out. For this purpose, the database was segmented into a learning database and a test database. As the work was carried out over a single year, a random selection of 10 months of care from the 12 months was carried out to form the learning base and the remaining 2 months the test base. In order to avoid any bias, the stability of the coefficients was checked for all possible combinations of learning base and test base at each modelling session. A classical cost-frequency modelling by Generalised Linear Models (GLM) was then applied. The results were as follows :

	FSEJ	CHBR	AUTR	Pricing 2	Pricing 1
Average premium observed	460,94	0,85	5,03	466,82	466,82
Predicted average pure premium	490,13	0,82	5,25	496,20	496,17
gap	6,334%	-2,886%	4,222%	6,294%	6,287%

FIGURE 11 – Result Frequency-cost model - pricing 1 and 2

If it corresponds to a medical or operational logic, this segmentation has a marginal contribution in the modelling of the risk. In fact, when considering a classic cost-frequency modelling using generalised linear models (GLM), pricing by sub-item (noted "pricing 2") doesn't improve the prediction of incurred expenditure compared to pricing without segmentation (noted "pricing 1"). This lack of gain, however, follows the implementation of six models compared to two models in the case of global pricing.

This observation confirms the need to reconsider the segmentation by subcategory of care usually accepted, which is what we propose to do next.

Proposal of an alternative segmentation

Previous results have motivated the construction of a new, more efficient segmentation. To do so, an unsupervised classification algorithm, HDBSCAN, was used to build groups of acts with homogeneous distributions. The selected segmentation variables were the mean and standard deviation of the mean cost and frequency distributions. The methodology followed was then as follows :

— segmentation of acts according to their average cost distribution, on the one hand, and their frequency distribution, on the other ;

— intersection of the two segmentations in order to have homogeneous clusters in terms of average cost and frequency ;

— adjustment of the raw clusters obtained to have more robust cluters : 4 clusters were thus obtained.

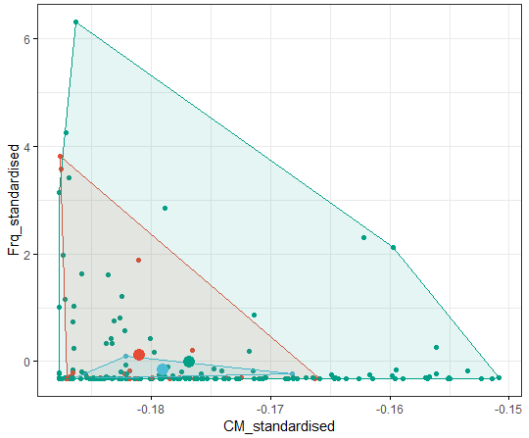


FIGURE 12 – Segmentation by sub-category of care

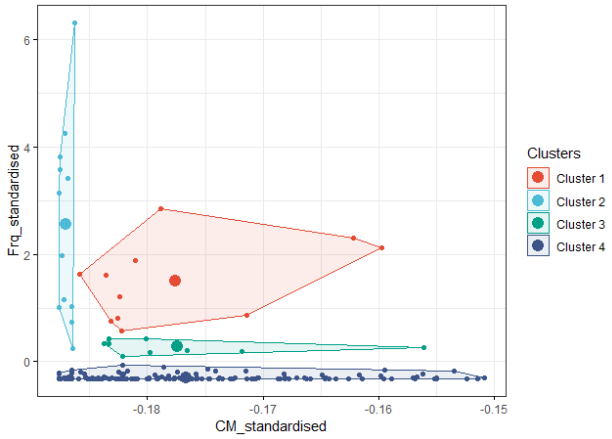


FIGURE 13 – HDBSCAN segmentation

Compared to the segmentation by sub-item, this segmentation (noted "pricing 3") highlights four act profiles : care with a very high cost and high frequency (cluster 1), care with a low average cost but high frequency (cluster 2), care with a low average cost but low frequency (cluster 3) and care with a high average cost but low frequency (cluster 4).

Clusters	Number of cares	Numbers of cares (%)	Expenditure incurred	Expenditure incurred (%)	Numbers of claims	Numbers of claims (%)	Average cost	Average frequency
1	18	7%	28 966 914 563	88,2%	129 921 135	47,9%	222,96	1,94
2	12	4%	280 294 730	0,9%	107 932 535	39,8%	2,60	1,61
3	10	4%	1 175 355 269	3,6%	18 246 874	6,7%	64,41	0,27
4	234	85%	2 438 274 411	7,4%	15 332 617	5,6%	159,03	0,23
Total	274		32 860 838 973		271 433 161			

FIGURE 14 – Caractéristiques des *clusters* formés

Pricing according to the proposed new segmentation

Pricing by GLM (a constraint of the study) on these new clusters makes it possible to obtain a total predicted expenditure that is 6% higher than the observed expenditure.

While this level is close to that of sub-line pricing, the major gain lies in the improvement of individual claims predictions. Indeed, while the sub-item pricing leads to loss ratios (S/P [9]) deviating on average [10] by 12 points from the equilibrium level (100%), the new clusters reduce this deviation to 9.6 points. A higher concentration of these ratios around 100 percent is then observed with lower levels of overshoot than in the two previous pricing as shown in the graph 15.

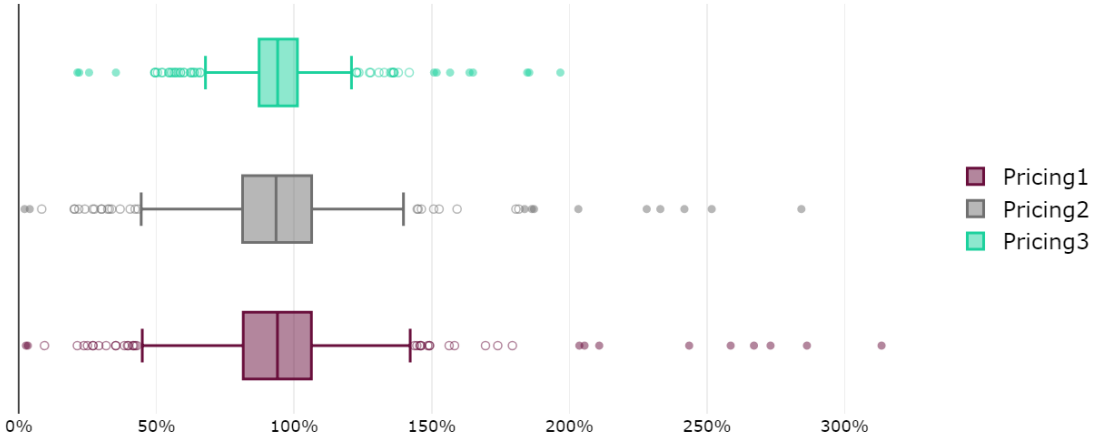


FIGURE 15 – Dispersion of loss ratios according to pricing

In addition, more than 28% of the profiles have an underestimated loss ratio (S/P>100%) against 37% in tariff 2. A pricing error representing more than 30% of the expenses incurred, falling to 17% after application of the new proposed segmentation, i.e. a 45% reduction in the loss recorded.

In this way, the latter pricing improves the risk/pricing adequacy and profitability of the portfolio.

	Profiles (%) with loss ratio > 100%	Cost incurred for profiles with S/P >100%		Average loss ratio deviation at 100%	Max. Loss ratio
		Euro	Share of total expenditure incurred (%)		
Pricing 1	37,7%	1 847 894 409	35,46%	12,74%	313,44%
Pricing 2	37,0%	1 652 856 821	31,72%	12,18%	284,31%
Pricing 3	28,1%	908 104 601	17,43%	9,60%	196,74%

FIGURE 16 – Comparison of Prediction Results According to Pricing

9. Observed incurred expenditure / predicted incurred expenditure
10. Weighted average by exposure of the absolute value of the deviations at 100%

Conclusion

The study carried out sought to observe the possibility of using a machine learning algorithm, HDBSCAN, in order to test a modelling of the hospitalisation sinistrality at a sub-station grid.

Based on open source data provided by the Assurance Maladie, the absence of segmentation of acts has served as a reference for comparison.

The test breakdown is obtained after clustering the procedures by mean and standard deviation of average costs and frequencies. On these axes, the clusters created display more homogeneous characteristics. Tariff adequacy by profile thus loses volatility (-25%) after having redesigned into 4 groups the acts usually separated into 3.

The results presented are based on the assumptions made and the resources available for the study, namely the use of an aggregated basis, implying a loss of individual information, as well as the use of GLMs, which are models commonly used operationally. In addition, the results depend on the clustering parameters tested, in particular the initial distance of distribution of points and the segmentation variables. On this last point, new variables judging the relevance of the adequacy of care to laws adapted to GLM, could complete the analysis.

The contribution of this methodology to the example developed thus suggests a significant gain if it were to be applied to the claims segmentation operationally carried out by insurers. Certain guarantees with similar distributions could thus be modelled together while guarantees with unsuitable distributions could be re-segmented.

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais tout d'abord adresser toute ma reconnaissance à mon équipe d'encadrants *Sia Partners* composée de Nicolas SERVAN, Baptiste ANDRIEU et Jordan MARIE-ROSE, pour leur disponibilité et leurs judicieux conseils qui ont contribué à alimenter ma réflexion.

Je désire aussi remercier le corps professoral de l'EURIA qui m'a fourni les outils nécessaires à la réussite de mes études et de ce mémoire.

Un grand merci à ma tutrice universitaire, Annaëlle LE BERRE, pour ses conseils qui ont grandement participé à améliorer mon travail.

Je tiens à remercier également Sarah DIDO, Andréa EHUI et Sophie NAVARRO de l'équipe Actuariat *Sia Partners*, pour leurs divers apports tout au long de ces travaux.

Enfin, je tiens à témoigner toute ma gratitude à ma famille et à tous mes proches pour leur soutien moral inestimable.

Table des matières

Introduction

En France, la santé est un secteur en constante mutation du fait des fréquentes évolutions de la législation la régissant. En effet, de l'exigence de contrats complémentaires solidaires et responsables, à l'obligation pour les entreprises de faire adhérer leurs employés à des complémentaires collectives, et plus récemment, l'instauration d'une offre de soins à reste à charge nul pour les assurés sur certaines branches [11] , les organismes proposant des complémentaires santé se retrouvent de plus en plus sollicités dans la prise en charge du risque santé.

Face à tous ces bouleversements du cadre de remboursement imposé par l'exécutif, les complémentaires santé cherchent à se préparer à absorber le risque grandissant en s'efforçant à acquérir une meilleure compréhension de la consommation de soins des assurés. Cette meilleure connaissance du risque permettrait ainsi d'améliorer la tarification des produits proposés et constituerait un levier de performance intéressant dans le contexte fortement concurrentiel du marché de l'assurance maladie complémentaire.

Toutefois, l'amélioration des capacités de quantification du risque des complémentaires est entravée par certaines mesures, notamment sur le périmètre des contrats collectifs et/ou responsables. En effet, l'absence d'examens médicaux à la souscription amplifie l'asymétrie d'informations entre assureur et assuré. Cela se traduit donc par un risque accru d'inadéquation entre le risque supporté et la tarification qui en découle.

Sur la garantie hospitalisation, cette probable inadéquation risque-tarification est d'autant plus forte que les complémentaires santé ne connaissent qu'une infirme partie du risque réel couvert du fait d'une prise en charge importante de la Sécurité sociale sur cette branche. Elles font donc face à un risque peu maitrisé mais qui représente 37% de leur portefeuille.

Ainsi, ce mémoire vise, à partir des bases de données *open source* d'hospitalisation de la Sécurité sociale, à fournir une analyse du risque puis à questionner les méthodes de tarification actuelles pratiquées sur cette garantie.

Pour ce faire, la consommation de soins hospitaliers sera étudiée puis modélisée à l'aide des modèles tarifaires usuels (GLM). Une première tarification sera alors réalisée sans distinction des actes, suivi d'une autre tarification basée sur un regroupement desdits actes en sous-postes de soins conformément aux pratiques des assureurs. Enfin, un algorithme de classification non supervisé sera utilisé afin de challenger la segmentation de marché utilisée pour la seconde tarification.

11. Optique, dentaire et prothèse auditive

Chapitre 1

La prise en charge de la santé en France

En France, la santé fait partie des risques sociaux particulièrement encadrés dont la prise en charge s'inscrit dans un dispositif particulier. Avant toute étude, il est donc indispensable d'en comprendre le fonctionnement.

Ainsi, ce premier chapitre vise, dans ses trois premières sections, à décrire le cadre global (acteurs, financement, règlementation) de prise en charge des dépenses de santé en France en prêtant une attention particulière à la couverture des coûts générés par une hospitalisation. Deux autres sous-parties viendront ensuite expliciter notre problématique ainsi que l'issue offerte par les données *open source* de la Sécurité sociale.

1.1 La prise en charge de la Sécurité sociale

Depuis 1945, la France est dotée d'un organisme, la Sécurité sociale (SS), chargée de protéger les individus face aux risques sociaux qui pourraient impacter à la baisse leurs revenus ou augmenter leurs dépenses. Les risques concernés sont par exemple la maladie, l'invalidité ou le chômage, et sont regroupés en 4 branches :

— la branche maladie comprenant la maladie, la maternité, l'invalidité et le décès ;

— la branche accident du travail et maladie professionnelle ;

— la branche famille (handicap, logement, aides contre la précarité...) ;

— la branche retraite (vieillesse et veuvage).

La branche vieillesse et la branche maladie sont les deux principales branches de la Sécurité sociale et représentent respectivement 49% et 42% des prestations nettes versées en 2018 [1].

1. régime général - Source : Commission des Comptes de la Sécurité sociale

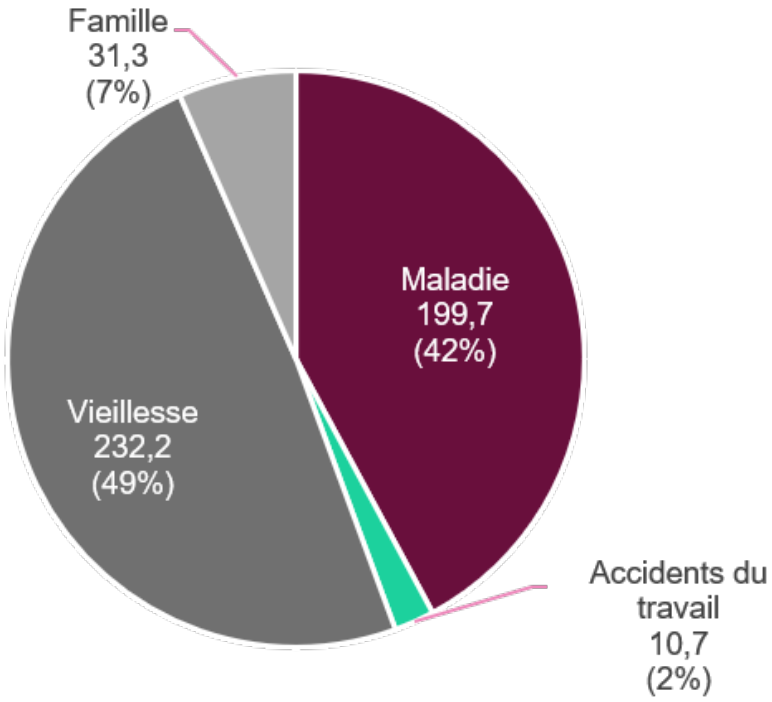


FIGURE 1.1 – Répartition par branche des prestations nettes de la Sécurité sociale en 2018 (en Md€)

La Sécurité sociale, par l'intermédiaire de sa branche maladie, assure la prise en charge des dépenses de santé des assurés et leur garantit l'accès aux soins. Pour ce faire, la Sécurité sociale est structurée autour de différents régimes basés sur la segmentation socio-professionnelle de la population. Il existe ainsi :

— le régime général qui couvre les travailleurs salariés et les travailleurs indépendants (depuis le 1er janvier 2018). Il se charge également des étudiants et de toutes autres personnes n'appartenant pas aux catégories suscitées dont les plus démunis à travers des dispositifs tels que la Complémentaire santé solidaire, anciennement Couverture Maladie Universelle (CMU) ;

— le régime agricole regroupant les exploitants et les salariés agricoles ;

— de nombreux régimes spéciaux pour des corporations professionnelles particulières (agents SNCF, agents RATP, Assemblée nationale, Sénat, . . .).

Ces différents régimes, dits régimes de bases, sont obligatoires et assurent le remboursement partiel ou total des frais médicaux réels (tarifs appliqués par les praticiens). En effet, pour un acte donné, le niveau de remboursement dépend du tarif de référence fixé par la Sécurité sociale en lien avec ledit acte, pouvant différer du tarif réel. Sur ce tarif de référence, dit tarif conventionné ou base de remboursement, est appliqué un taux de remboursement pour obtenir le niveau de prise en charge de la Sécurité sociale (cela peut également être un montant forfaitaire).

Dans le cas d'une prise en charge partielle, le surplus à la charge des assurés sociaux peut être segmenté en 3 parties :

— une participation forfaitaire de 1 euro (certains profils d'assurés en sont exemptés [2]) ;

— le Ticket Modérateur (TM) qui représente l'écart entre le remboursement de la Sécurité sociale et le tarif conventionné déduction faite de la participation forfaitaire ;

— le dépassement d'honoraires qui correspond à la différence entre le tarif réel et le tarif conventionné.

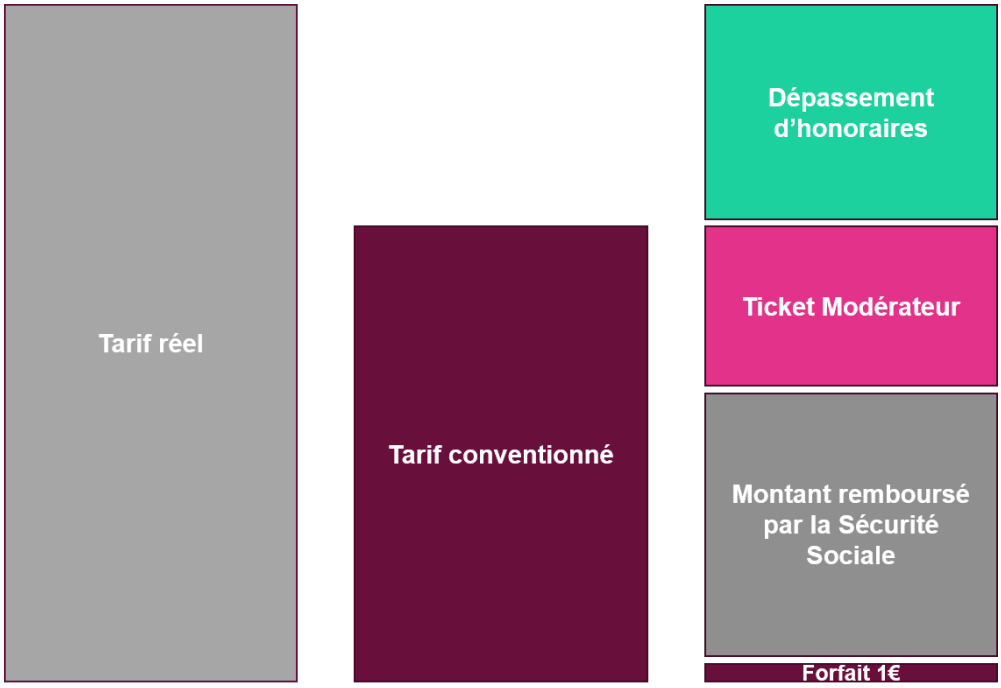


FIGURE 1.2 – Reste à charge (en % de la dépense totale) des ménages par poste de soins en 2018

Ce principe de remboursement est conditionné par deux facteurs :

— Le respect d'un parcours de soins imposé par la Sécurité sociale depuis 2007 pour une optimisation de la consommation de soins médicaux : chaque assuré doit désigner un médecin généraliste référent, dit médecin traitant, dont il devra avoir l'aval avant de consulter des spécialistes. En cas de non-respect de ce dispositif, l'assuré devra s'attendre à un niveau de remboursement réduit.

— Le statut du médecin consulté qui peut appartenir à trois catégories : les conventionnés secteur 1 dont les tarifs sont ceux fixés par la Sécurité sociale en contrepartie d'une réduction de leurs charges sociales, les conventionnés secteur 2 qui pratiquent des dépassements d'horaires encadrés, et ceux non signataires d'un des dispositifs de pratique tarifaire maîtrisée. Tandis que la Sécurité sociale appliquera un remboursement similaire basé sur son tarif de remboursement pour les deux premières catégories, une minoration importante sera appliquée pour les médecins non conventionnés.

2. Enfants et jeunes de moins de 18 ans, femmes enceintes à partir du 1er jour du sixième mois de grossesse et jusqu'au 12e jour suivant la date d'accouchement et les bénéficiaires de la Complémentaire santé solidaire (ex CMU-C et ACS) ou de l'aide médicale de l'État (AME).

1.2 La place des complémentaires santé

Comme vu *supra*, la prise en charge des soins de santé par la Sécurité sociale n'est pas totale et occasionne un reste à charge pour les patients : en 2018, le reste à charge toutes branches confondues représentait 20% de la consommation de soins. Toutefois, ce niveau global renferme de fortes disparités car si sur certains postes de soins la participation des ménages est faible, sur d'autres ils doivent financer plus de la moitié des coûts réels comme le montre le graphique ci-dessous :

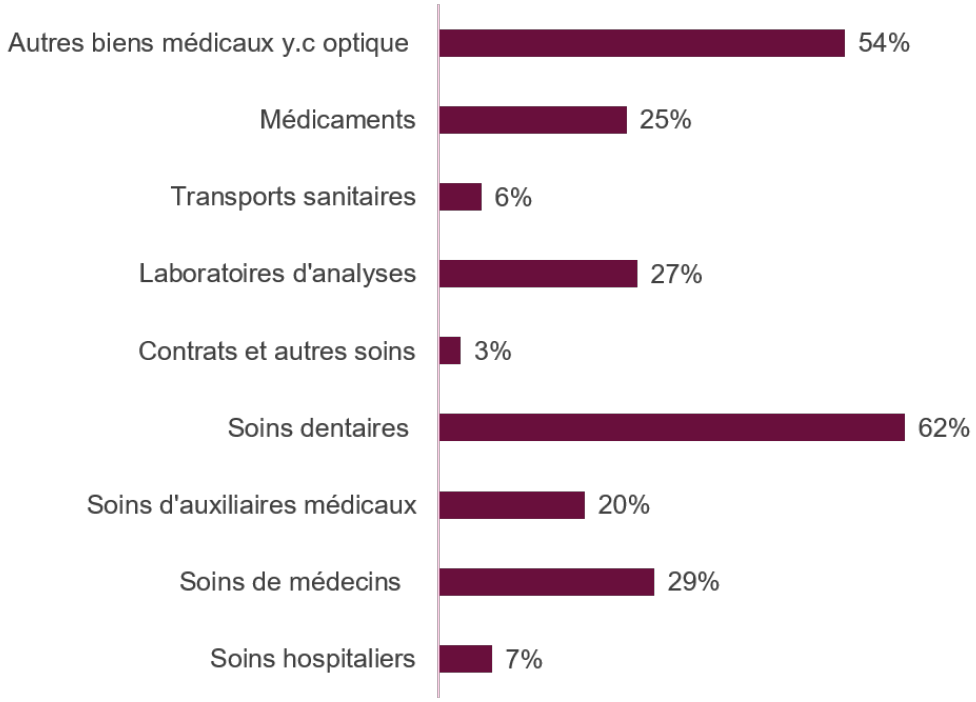


FIGURE 1.3 – Reste à charge ménage par poste de soins avant prise en charge des complémentaires santé

De ce fait, les complémentaires santé sont particulièrement importantes pour éviter un renoncement aux soins du fait de restes à charge trop importants. Elles viennent ainsi constituer une seconde couche de remboursement total ou partiel de ces restes à charge. De manière obligatoire ou facultative, collectives ou individuelles, les complémentaires santé ont couvert 66% du reste à charge des ménages, soit 13,43% des dépenses globales de santé en 2018 [3], ce, à travers trois catégories d'organismes : les sociétés d'assurance, les institutions de prévoyance et les mutuelles.

3. DREES, la Sécurité sociale contient également l'État et la CMU-C org. de base

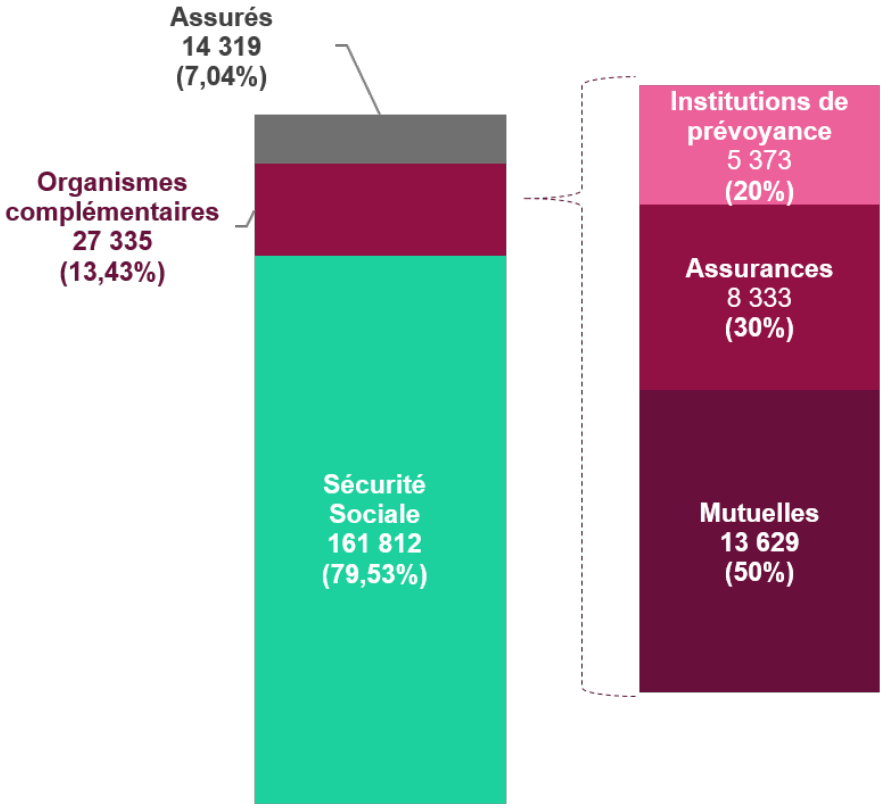


FIGURE 1.4 – Structure du financement des prestations médicales en 2018 (en Md€)

Dans le cadre des complémentaires, les remboursements peuvent prendre différentes formes :

— en pourcentage du Ticket Modérateur ;
— remboursement forfaitaire ;
— en pourcentage des frais réels ;
— en pourcentage de la base de remboursement de la Sécurité sociale...

Tous les remboursements sont effectués dans la limite des frais réels encourus.

1.3 Cas de l'hospitalisation

1.3.1 Etat des lieux des dépenses hospitalières et de leur prise en charge

Les soins hospitaliers regroupent l'ensemble des prestations (soins et hébergement) de court et moyen séjours fournis par les hôpitaux publics et privés. Ils s'articulent autour de 4 champs sanitaires :

— Médecine, Chirurgie, Obstétrique (MCO) qui a concerné 88% des patients en 2018 ;

— Soins de Suite et de Réadaptation (SSR) pour 7,6% des patients en 2018 ;

— Hospitalisation A Domicile (HAD) encore peu sollicitée (moins de 1% des patients en 2018) ;

— Psychiatrie qui a concentré 3,1% des patients en 2018.

Ces soins représentent 46,4% des dépenses globales de santé, et pèsent respectivement 54% et 37% dans les prestations servies par la Sécurité sociale et les complémentaires santé.

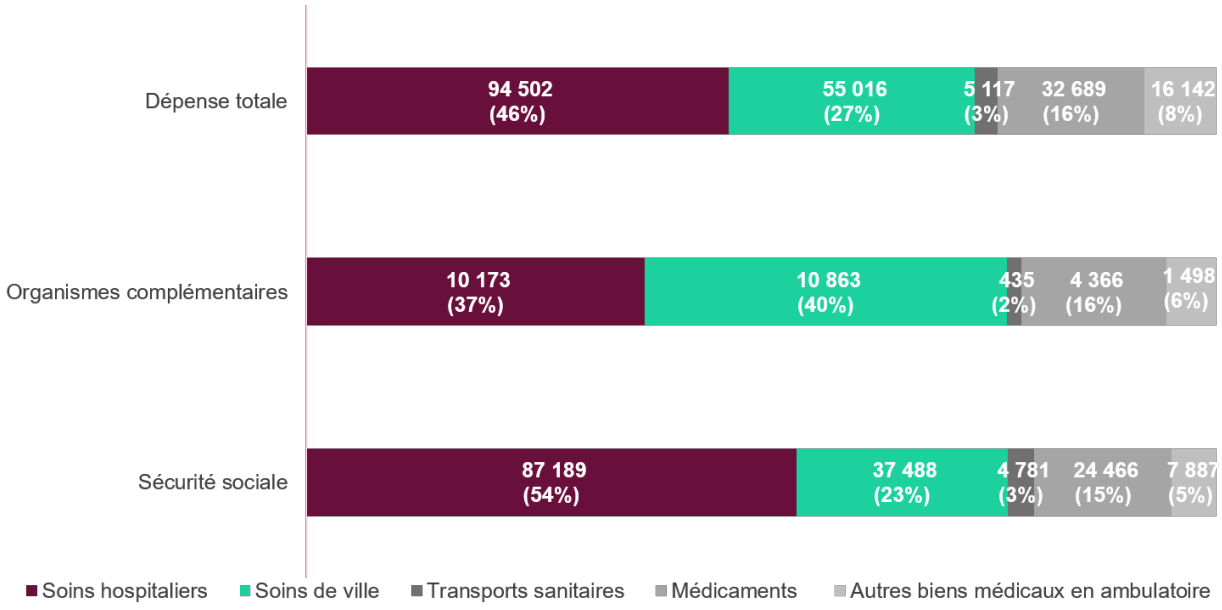


FIGURE 1.5 – Poids des soins hospitaliers dans les prestations de santé en 2018 [4]

Les frais d'hospitalisation remboursés correspondent aux frais directement engendrés par le traitement de la pathologie du patient. Ils concernent donc :

— les frais de séjour ;

— les frais de séjour ;

— les frais de salle d'opération ;

— les honoraires des praticiens et auxiliaires médicaux ;

— les frais d'analyses et d'examens de laboratoire relatifs aux soins dispensés pendant le séjour dans l'établissement hospitalier.

En cas de séjour dans un établissement conventionné (hôpital public ou clinique), les frais d'hospitalisation sont remboursés à hauteur de 80% par la Sécurité sociale, hors dépassements

4. DRESS

d'honoraires. La prise en charge des éventuels suppléments pour confort (chambre individuelle, télévision, etc.), du ticket modérateur, des dépassements d'honoraires et du forfait hospitalier est quant à elle partagée entre le patient et sa complémentaire santé en fonction du niveau de garantie du contrat souscrit.

Il est à noter que le forfait hospitalier correspond à une participation aux frais d'hébergement et d'entretien occasionnés par une hospitalisation. Depuis le 1er janvier 2018, il est de 20€ par jour en hôpital et en clinique tandis qu'il s'élève à 15 par jour en service psychiatrique.

Avec une évolution moyenne de 1,9% depuis 2009, la consommation de soins hospitaliers observée en 2018 s'élève à 94,5 Mds€.

En millions d'euros	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	TCAM
Total	78 356	80 316	82 461	84 567	86 688	89 060	90 430	92 320	93 770	94 502	
Evolution (%)		2,5%	2,7%	2,6%	2,5%	2,7%	1,5%	2,1%	1,6%	0,8%	1,9%

FIGURE 1.6 – Consommation des soins hospitaliers entre 2009 et 2018 [5]

Cette consommation de soins est majoritairement absorbée par la Sécurité sociale à hauteur de 92,9% tandis que les organismes complémentaires en supportent 5,2%. Les ménages quant à eux en assume les 1,9% restants.

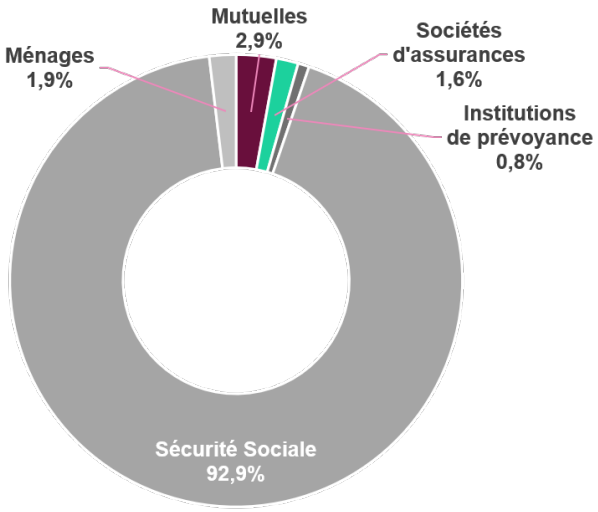


FIGURE 1.7 – Structure de financement des soins hospitaliers en 2018 [6](M€)

5. DRESS
6. DREES, la Sécurité sociale y.c l'État et la CMU-C org. de base

1.3.2 La tarification à l'activité (T2A)

Avant 2005, les hôpitaux fonctionnaient selon un budget forfaitaire global alloué annuellement par l'Assurance Maladie pour couvrir les frais générés par les soins hospitaliers. Ce budget avait la particularité de ne pas être défini selon le volume d'activité ni la nature des soins réalisés. Ainsi, bien qu'un tel mode de financement permît aux hôpitaux d'avoir une lisibilité claire de leurs ressources, il pénalisait, en n'évoluant pas nécessairement dans la même proportion, ceux dont l'activité augmentait, et à l'inverse créait des réserves pour ceux dont l'activité était en baisse continue. Afin d'y remédier et avoir une meilleure adéquation entre le financement reçu et les soins procurés, la tarification à l'activité fut introduite pour les activités de médecine, chirurgie et obstétrique (MCO) dans le cadre d'une vaste réforme du système de santé intitulé la « Nouvelle gouvernance hospitalière ». Cette tarification fut appliquée d'abord au secteur privé puis au secteur public pour parvenir à une harmonisation des ressources provenant de la Sécurité sociale reçues par les deux secteurs. Désormais c'est le nombre et la nature des actes et séjours pris en charge qui déterminent les ressources de l'hôpital.

Le mécanisme de la T2A repose sur un système de classification des séjours de chaque patient selon des groupes homogènes de malades (GHM) prédéfinis et associés à des niveaux de coûts qui serviront de base de remboursement par l'Assurance Maladie. Ces GHM dépendent du diagnostic de la maladie et des actes pratiqués pendant le séjour : en mars 2019, il en existait 2 593.

La classification des GHM est régulièrement actualisée pour s'assurer de l'homogénéité des regroupements. Par ailleurs, chaque GHM est décliné en 4 sous-groupes selon le niveau de sévérité de la maladie : on parle de GHM racine. A ces derniers, sont également associés un ou plusieurs groupes homogènes de séjours (GHS) indiquant l'intervalle « normal » de temps d'hospitalisation nécessaire pour un patient donné. Toute cette segmentation conditionne le niveau de facturation admis par l'Assurance Maladie.

1.3.3 Une différence significative entre le secteur public et le secteur privé

Sur la période 2009 – 2018, la part des dépenses d'hospitalisation inhérentes au secteur public s'élève à 77% :

En millions d'euros	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Ventilation 2009-2018
Total	78 356	80 316	82 461	84 567	86 688	89 060	90 430	92 320	93 770	94 502	
Secteur public	60 211	61 701	63 294	64 952	66 779	68 603	69 781	71 182	72 373	72 959	77%
Secteur privé	18 145	18 615	19 166	19 615	19 909	20 457	20 649	21 138	21 397	21 543	23%

FIGURE 1.8 – Consommation des soins hospitaliers entre 2009 et 2018 [8]selon le type d'établissements (M€)

Cet écart s'explique notamment par une capacité d'accueil des patients largement supérieure dans les hôpitaux publics : 76 [9]% des lits et places se trouvent dans le secteur public. Par ailleurs, ces deux types d'établissements ont des profils très différents sur de nombreux aspects :

— **Le mode de financement**
Une différence fondamentale entre les hôpitaux publics et les cliniques réside dans la source de financement qui conditionne leur objectif final. En effet, une clinique est mise en place sur fonds privés avec des objectifs de rentabilité fixés par les actionnaires tandis que l'hôpital public est financé par des fonds publics avec une mission d'utilité publique à but non lucratif.

Ainsi, la tarification pratiquée, le profil des malades pris en charge ou encore les maladies soignées se verront fortement impactés par le mode de financement de l'établissement considéré.

— **Les dépassements d'honoraires**
Dans les hôpitaux publics, les tarifs des consultations, examens médicaux et de tout autre acte/service respectent la grille secteur 1 de l'Assurance Maladie sans dépassement d'honoraires. En revanche, dans le secteur privé, la facturation est fortement dépendante des praticiens, qui, contrairement aux médecins salariés des hôpitaux publics exercent leur activité de manière libérale (cf. Figure 1.9). Ce qui occasionne des dépassements d'honoraires fréquents et parfois très importants.

8. DREES
9. Etablissements publics et privés à but non lucratif – données 2018 DRESS
10. Données 2017-DREES

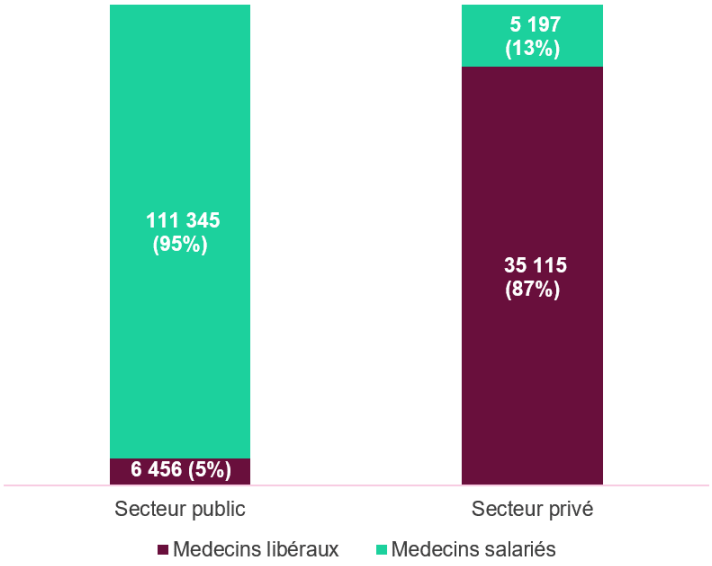


FIGURE 1.9 – Part des médecins salariés et libéraux par type d'établissement [10]

En 2018, les dépassements d'honoraires s'élevaient à 1 milliards d'euros avec des niveaux plus ou moins élevés selon les actes concernés :

Familles d'actes	% séjours avec dépassement	Dépassement moyen
Actes thérapeutiques sur les articulations du membre inférieur	72,50%	547,3
Actes thérapeutiques sur les nerfs crâniens et les nerfs spinaux	9,80%	126,1
Actes thérapeutiques, de radiothérapie externe, curiethérapie, ou administration de radio-isotope	16,20%	541,8
Actes thérapeutiques sur la trachée, les bronches, les poumons, et la plèvre	27,10%	861,6
Actes thérapeutiques sur les sourcils et les paupières	63,40%	283,7

FIGURE 1.10 – Exemple de dépassements d'honoraires sur certaines familles d'actes en 2018[11]

— **Les soins pris en charge**
Comme susmentionné, les cliniques n'existent que par l'investissement de fonds personnels de tiers et n'ont pas vocation à perdre de l'argent. Ainsi, elles auront tendances à se positionner sur des affections légères et/ou « rentables » et à laisser le secteur public absorber tout le reste. Le fait est que les GHM à faible rentabilité ne sont pas très attractifs pour le secteur privé. Quant au secteur public, il ne peut se cantonner aux activités rentables au risque de faillir à sa mission de service public et prendra davantage en charge les pathologies les plus lourdes, et accessoirement, les moins rentables.

Il en découle un fossé entre les maladies prises en charge par les cliniques et les hôpitaux publics. En effet, en 2017, par exemple, les affections chroniques de longues durées (séances

11. ATIH

de chimiothérapie, dialyses...) n'ont représenté que 9% des causes de soins hospitaliers en clinique contre 33% pour les hôpitaux publics.

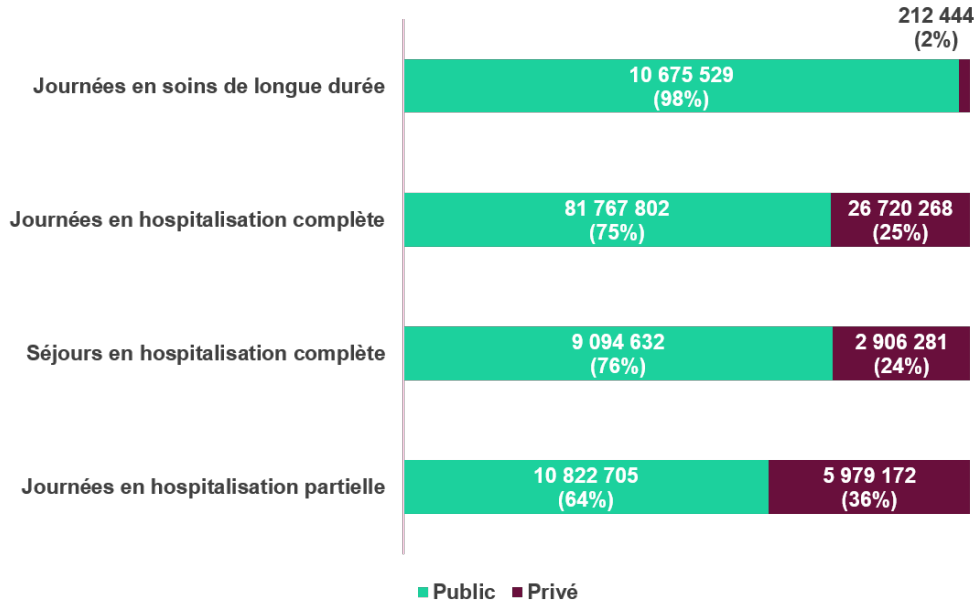


FIGURE 1.11 – Répartition des séjours hospitaliers selon le type d'établissement en 2017 [11]

— **Le refus implicite de soins**
Lors de la sollicitation d'un acte médical, un patient peut être amené à avancer la totalité des frais médicaux ou bénéficier du système de tiers payant dispensant de régler immédiatement le praticien de la part remboursée la Sécurité sociale et/ou par les complémentaires. L'objectif de ce dispositif est d'éviter le renoncement aux soins par les ménages faute de ressources financières. Toutefois, n'étant pas obligatoire [12] , de nombreux praticiens ne le proposent pas comme c'est souvent le cas dans les établissements privés.

En plus de cet absence de tiers payant, des dépassements d'honoraires fréquents conduisent à une faible représentation des populations les plus précaires parmi les patients du secteur privé. En effet, bien qu'en théorie toute personne peut s'y faire soigner, certaines sont freinées par la nécessité de mobiliser des sommes d'argent conséquentes. De plus, selon les directives de la Sécurité sociale, les bénéficiaires de la PUMa ne peuvent avoir recours à des praticiens appliquant des dépassements d'honoraires. Ces patients iront de fait très rarement dans les cliniques.

Il y a donc une différence entre les profils des patients accueillis en secteur privé et ceux de l'hôpital public qui accueille et prend en charge tous les patients quelle que soit leur situation sociale.

11. DRESS
12. Sauf maternité ou d'une affection de longue durée

1.3.4 Quelle évolution de la prise en charge des complémentaires ?

Les complémentaires santé ne couvrent qu'une infirme partie des soins hospitaliers contrairement à d'autres segments tels que l'optique ou encore le dentaire : Ainsi, l'Assurance Maladie absorbe une

Catégories de soins	Dépenses des organismes complémentaires	
	En millions d'euros	Part des dépenses par catégorie prise en charge par les complémentaires
Soins hospitaliers	4 941	5,2%
Soins de ville	11 877	21,6%
Médicaments	4 072	12,5%
Optique	4 746	74,2%
Soins dentaires (y compris prothèses)	4 823	42,2%

FIGURE 1.12 – Financement des organismes complémentaires par poste de soins en 2018[14]

partie importante des dépenses mais se retrouve confrontée à des déficits importants des hôpitaux publics depuis plusieurs années.

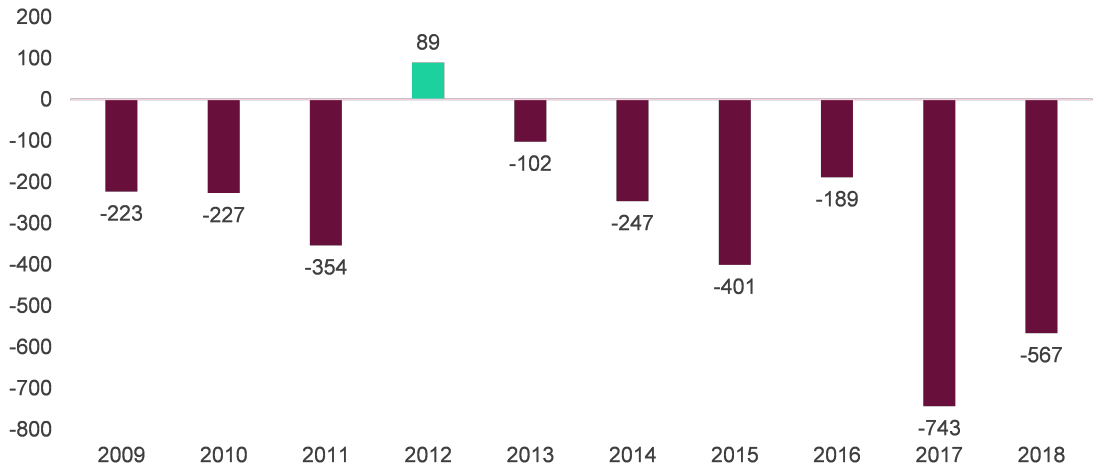


FIGURE 1.13 – Evolution du résultat net des hôpitaux publics entre 2009 et 2018[14] (M€)

14. DRESS

Il est donc légitime de s'interroger sur la voie de résolution que pourrait adopter l'exécutif. Sachant que les complémentaires santé couvrent aujourd'hui 95% de la population, l'un des moyens probables serait d'augmenter leur prise en charge des soins pour réduire celle supportée par la Sécurité sociale.

Un tel changement impacterait considérablement l'activité des complémentaires santé car, comme vu plus haut, elles ne couvrent que 5,2% du risque tandis que celui-ci représente 37% de leurs portefeuilles. Elles ont de fait une sensibilité accrue à tout désengagement de la Sécurité sociale sur ce segment sur lequel elles naviguent pourtant à l'aveugle.

Il est donc primordial d'étudier ce poste de soins dans son entièreté à travers les données *open source* mise à disposition par la Sécurité sociale afin de mieux connaitre le risque couvert.

1.3.5 Quels apports des données *open source* dans le domaine de l'assurance santé ?

L'informatisation complète du secteur médical effective depuis de nombreuses années permet à la Sécurité sociale de disposer de nombreuses données de santé qu'elle met de plus en plus à la disposition du public. Afin d'être en adéquation avec RGPD, ces données sont toujours anonymisées et présentées sous forme d'indicateurs synthétiques ou sous forme de bases de données brutes nécessitant une exploration détaillée et des retraitements.

Si l'objectif premier est de permettre d'alimenter diverses études médicales, ces données constituent une mine d'or pour les complémentaires santé qui ont aujourd'hui une vision parfois opaque de certains postes de soins. En effet, leur fiabilité et leur exhaustivité à l'échelle de la population nationale sont des atouts majeurs par rapport aux informations pouvant être recueillies sur des portefeuilles spécifiques. Ainsi, à partir de ces données, un suivi plus fin et global des comportements peut être effectué pour pouvoir mieux évaluer et gérer les risques couverts. Elles offrent également plusieurs avantages en matière d'innovation et de tarification :

- conception de nouveaux produits plus facilement particulièrement pour les acteurs ne disposant pas d'un portefeuille interne conséquent pour faire les modélisations ;
- amélioration des travaux sur les portefeuilles existant en y intégrant ces données externes ;
- identification plus robuste des lois adaptées au risque avant de faire les modélisations sur un portefeuille réduit donné ;
- conception de produits plus adaptés à la réalité ;
- ajustement de la prime d'assurance au plus près de la réalité du risque ;
- établissement de programmes de prévention ciblés.

Globalement, ces données feront émerger de nouveaux indicateurs d'analyse du risque pour trouver les leviers d'une meilleure adéquation technique et flexibilité des garanties offertes, d'où l'intérêt qui leur est porté dans la présente étude.

Chapitre 2

Construction des bases de modélisation et analyses préliminaires des données

Avant toute modélisation tarifaire, il est indispensable de disposer ou de construire une base de données contenant l'information à modéliser et répondant aux exigences des modèles utilisés.

Ce chapitre, articulé en trois sections, vise donc à présenter les étapes suivies pour parvenir à des bases de modélisation adaptées à la présente étude. Les deux premières sections développent ainsi les méthodologies de construction des bases « sinistres » et « assurés ». Les bases obtenues font ensuite l'objet, dans la dernière section, d'analyses statistiques afin d'établir les caractéristiques empiriques du risque. Les conclusions de ces analyses serviront de référentiels pour évaluer la cohérence des modèles à réaliser.

2.1 Construction de la base « sinistres »

Partant de la base Open DAMIR mise à disposition par la Sécurité sociale, plusieurs traitements ont été nécessaires pour aboutir à une base de modélisation répondant aux besoins de l'étude. Ils sont présentés ci-après.

2.1.1 Présentation de la base brute

Les données utilisées dans le cadre de la présente étude correspondent aux données *open source* mises à disposition par l'Assurance Maladie. Celles-ci découlent d'une extraction du Système National Inter Régimes d'Assurance Maladie (SNIIRAM) couvrant l'ensemble des prestations facturées à l'Assurance Maladie y compris les prestations hospitalières, et ce, tous régimes confondus.

A date, l'historique disponible couvre les exercices d'assurance 2009 à 2019, soit 11 ans de données. Du fait de l'importante volumétrie de données, chaque année est associée à un ensemble de 12 fichiers dont chacun retrace la consommation de soins observée sur un mois donnée. Il est à noter que ces fichiers mensuels font état des prestations versées selon les dates de remboursement (année et mois précisés dans l'intitulé de chaque fichier) et non les dates de soins. Ils sont disponibles sous la forme de fichiers CSV comportant à minima 30 millions de lignes pour un total de 55 variables : exprimé en octet, chaque base mensuelle représente environ 9Go, soit 108Go pour une base annuelle.

Par ailleurs, contrairement aux bases assureurs traditionnelles, afin d'éviter toute identification des personnes et entités concernées, les données ont été anonymisées et traduisent le risque par profil homogène d'individus. Ces profils dépendent à la fois des caractéristiques personnelles des

patients, de la nature de la prestation, du type de prescripteurs et du type d'exécutant médical. Ils sont ainsi décrits par 55 variables pouvant être regroupées en 6 volets :

Axes d'analyse	Variables
Période	▪ Année de soins ▪ Mois de soins
Bénéficiaires	▪ Sexe ▪ Tranche d'âge au moment des soins ▪ Qualité (souscripteur principal, conjoint, enfant…) ▪ ZEAT (Zone d'Etudes et d'Aménagement du Territoire) de résidence ▪ Région de résidence ▪ Modulation du ticket modérateur ▪ Appartenance au CMU-C
Prestations	▪ Nature de prestation (actes médicaux) ▪ Nature d'assurance (maladie, maternité, ATMP…) ▪ Nature de l'accident du travail ▪ Type d'enveloppe (régime de prise en charge) ▪ Complément d'acte (majorations diverses) ▪ 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 du prescripteur ▪ ZEAT du prescripteur ▪ Région du prescripteur ▪ Statut juridique du prescripteur ▪ ZEAT d'implantation de l'établissement ▪ Région d'implantation de l'établissement ▪ Catégorie de l'établissement
Exécutant	▪ Catégorie de l'exécutant ▪ Spécialité médicale de l'exécutant ▪ Nature d'activité de l'exécutant ▪ ZEAT de l'exécutant ▪ Région de l'exécutant ▪ Statut juridique de l'exécutant ▪ Mode de fixation des tarifs de l'établissement ▪ ZEAT d'implantation de l'établissement ▪ Région d'implantation de l'établissement ▪ Catégorie de l'établissement ▪ Discipline de prestation de l'établissement ▪ Mode de traitement de l'établissement
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 2.1 – Bilan des variables de modélisation disponibles

2.1.2 Extraction des données sur l'hospitalisation

Comme susmentionné, la base brute disponible contient toutes les prestations versées par l'Assurance Maladie. Ainsi, un premier travail d'isolement des dépenses entrant dans le périmètre du risque hospitalisation a dû être réalisé avec pour objectif l'extraction des lignes de prestations relatives aux :

- frais de séjour ;
- frais de salle d'opération ;
- honoraires des praticiens et auxiliaires médicaux ;
- frais d'analyses et d'examens de laboratoire en lien avec les soins dispensés pendant un séjour en établissement hospitalier.

Période d'observation retenue

Pour des questions de volumétrie de données et de capacités de traitement, l'étude portera sur les données les plus récentes, soient celles l'année de 2019. Toutefois, en cas, de travaux sur plusieurs années, une harmonisation des niveaux de dépenses est nécessaire afin de retranscrire l'effet de l'inflation. Deux solutions peuvent ainsi être adoptée pour traiter cette inflation :

- intégrer l'année de survenance du sinistre comme variable explicative qualitative. Ainsi, il sera possible, par l'intermédiaire des coefficients de régression obtenus pour cette variable, de capter l'évolution du coût des sinistres imputable à l'inflation. Une extrapolation de ces coefficients pourra ensuite être réalisée afin d'obtenir le montant de sinistres futur ;
- Simuler l'impact de l'inflation en réalisant un ajustement du coût des sinistres sur la base d'un indicateur reflet de l'inflation des dépenses de santé, à l'image du CMT (ce dernier est un exemple).

Actes médicaux retenus

Le périmètre d'étude exclut les actes de maternité et de pharmacie en lien avec les soins hospitaliers. Quant à la sélection des autres actes médicaux entrant dans le champ de l'étude s'est faite en deux étapes :

- **Etape 1 : extraction des actes selon la classification de la Sécurité sociale**
  Les filtres nécessaires à l'extraction des données relatives au risque hospitalisation ont été déterminé à la suite de divers échanges avec la gestionnaire de la base Open DAMIR. Ainsi, il est important de noter que :
  - La base contient les données complètes d'hospitalisation des cliniques privées : frais de séjour, frais de salle d'opération, honoraires et examens de biologie, autres produits et prestations délivrées en cliniques.
  - Les données relatives au secteur public sont incomplètes pour l'hospitalisation et concernent uniquement les honoraires des médecins.

  Cette première étape conduit à restreindre l'étude au secteur privé uniquement. En effet, les divergences entre les deux secteurs, comme expliqué dans la section 1.3.3, ne permettent pas d'extrapoler les résultats obtenus pour le secteur public

- **Etape 2 : exclusion ligne à ligne de certains actes**
  Les filtres suggérés par la gestionnaire de la base Open DAMIR permettent de réduire le périmètre d'étude mais restent insuffisants puisqu'ils conduisent à considérer des actes dentaires

ou d'optique réalisés en cliniques, de la pharmacie mais également diverses aides/subventions versées par la Sécurité sociale. Ainsi, par souci de cohérence avec la segmentation des actes utilisée par les assureurs, un second filtre manuel a été appliqué. Ce dernier réduit de 146 le nombre d'actes pris en compte puisqu'il passe de 420 à 274. Les actes exclus représentent 16% du volume de sinistres et 14% du coût total des prestations en lien avec les actes de la sélection initiale.

La liste des actes sélectionnés est présentée en annexe A.

	Nombre d'actes	Coût total (euro)	Sinistres (volume)
Tous actes	420	38 433 041 966	322 962 924
Actes retenus	274	32 860 834 652	271 433 161
%	*65%*	*86%*	*84%*
Actes écartés	146	5 572 207 314	51 529 763
%	*35%*	*14%*	*16%*

FIGURE 2.2 – Bilan de la sélection des actes

2.1.3 Test de cohérence des données

Afin de s'assurer que la base Open DAMIR retranscrit bien les dépenses enregistrées par l'Assurance Maladie, les données comptables du régime général (principal régime avec 88% de la population couverte) ont été comparées aux niveaux de dépenses de la base Open DAMIR relatives à ce régime. Ce test de cohérence a d'abord été réalisée sans aucune distinction quant à la nature des soins puis s'est vu restreinte dans un second temps au segment des soins hospitaliers du secteur privé. Cette comparaison se justifie par le fait que les prestations dans la base Open DAMIR sont enregistrées en date de remboursement s'inscrivent donc bien dans le même référentiel que les données comptables.

En considérant toutes les prestations remboursées par le régime général, il existe un écart de 54 milliards d'euros entre la base Open DAMIR et les comptes dudit régime. Cet écart correspond bien (à 1% près du fait de retraitements comptables sur les données brutes) aux dépenses d'hospitalisation relatives au secteur public et confirme ainsi l'absence de ces données dans la base Open DAMIR.

Mois	Open DAMIR	Compte Ameli	Ecart	Dépenses d'hospitalisation des hôpitaux publics (maladie)
Janvier	10 646 891 200	15 896 021 811	5 249 130 611	5 310 034 087
Février	9 757 331 153	14 531 373 767	4 774 042 614	4 633 177 979
Mars	10 322 378 560	14 975 599 404	4 653 220 845	4 829 732 146
Avril	11 549 052 684	15 854 641 298	4 305 588 614	4 521 361 582
Mai	10 682 856 011	15 028 167 440	4 345 311 429	4 438 830 773
Juin	10 419 467 720	14 476 580 570	4 057 112 850	4 193 612 660
Juillet	11 073 758 946	15 204 287 410	4 130 528 464	4 243 636 588
Août	9 156 135 831	13 796 349 212	4 640 213 381	4 428 528 678
Septembre	10 228 619 238	14 454 495 772	4 225 876 533	4 257 950 191
Octobre	11 551 787 298	15 831 719 843	4 279 932 545	4 511 468 436
Novembre	10 402 246 739	14 593 389 399	4 191 142 660	4 163 319 221
Décembre	11 166 577 029	16 564 557 505	5 397 980 476	5 234 924 532
Total	**126 957 102 408**	**181 207 183 431**	**54 250 081 024**	**54 766 576 874**

FIGURE 2.3 – Prestations mensuelles 2019 du régime général selon la base Open DAMIR et les comptes AMELI

L'analyse de la sinistralité inhérente aux soins hospitaliers du secteur privé montre que la base Open DAMIR présente une sinistralité supérieure à celle affichée par les comptes AMELI. Les écarts observés sont détaillés dans la figure suivante et découlent de retraitements comptables réalisés lors de l'établissement des comptes. Ainsi, les résultats de cette étude auront tendance à être particulièrement prudents en retranscrivant un risque plus important que ce qui est observé dans la comptabilité de la Sécurité sociale.

Mois	Open DAMIR	Données comptables	Ecarts
Janvier	1 651 651 173	1 435 881 585	15,0%
Février	1 554 436 236	1 350 036 981	15,1%
Mars	1 422 052 709	1 189 612 148	19,5%
Avril	1 796 736 578	1 611 387 011	11,5%
Mai	1 654 078 102	1 430 895 459	15,6%
Juin	1 583 248 024	1 380 790 993	14,7%
Juillet	1 665 118 210	1 442 674 203	15,4%
Août	1 287 425 253	1 151 874 342	11,8%
Septembre	1 413 835 432	1 279 288 827	10,5%
Octobre	1 715 560 260	1 519 738 672	12,9%
Novembre	1 530 506 038	1 345 479 728	13,8%
Décembre	1 716 967 010	1 516 425 011	13,2%
Total	18 991 615 025	16 654 084 960	14,0%

FIGURE 2.4 – Prestations mensuelles 2019 d'hospitalisation privée du régime général selon la base Open DAMIR [15] et les comptes AMELI

15. Après sélection des actes hospitaliers

2.1.4 Traitements des données

La revue des données d'hospitalisation du secteur privé extraites de la base Open DAMIR a mis en exergue deux principales anomalies : l'existence de valeurs manquantes concernant les caractéristiques des bénéficiaires et la présence de valeurs négatives de coûts et nombres de sinistres. Des traitements ont donc été réalisés afin d'y remédier :

Traitements des valeurs négatives

Cette anomalie est le résultat de régularisations opérées par la Sécurité sociale sur les prestations remboursées. Elle concerne moins de 1% des lignes (correspondant à moins de 1% du volume de sinistres également) qui, en matière de dépenses, représentent 0,6% des prestations.

Ainsi, les lignes concernées ont été exclues pour réaliser tous les travaux de modélisation. Un abattement pourrait toutefois être réalisé sur les résultats pour coïncider avec le niveau global des dépenses obtenu en considérant lesdites régularisations.

Traitement des valeurs manquantes

Les données manquantes à corriger portaient sur les caractéristiques des bénéficiaires à savoir l'âge, le sexe et la région de résidence. 7% des lignes sont ainsi concernées avec, pour une ligne donnée, une ou plusieurs des variables susmentionnées absentes à la fois. La variable région est celle avec le plus de données manquantes suivi de la variable âge puis de la variable sexe. Les statistiques sur la répartition de ces anomalies selon les variables et les lignes concernées se présentent comme suit :

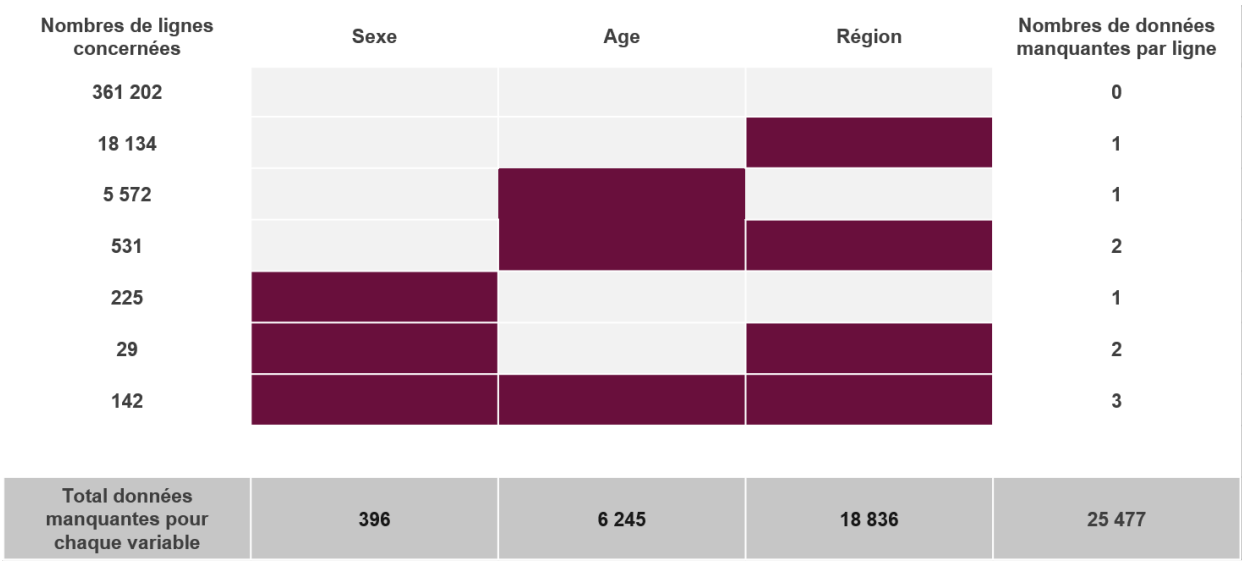


FIGURE 2.5 – Ventilation des données manquantes par variables et par lignes

Lecture : sur 361 202 lignes, il n'y a aucune donnée manquante. Sur 531 lignes, les variables âge et région ne sont pas précisées à la fois. Au total, la variable région n'est pas mentionnée pour 18 836 lignes.

Les lignes concernées par ces données manquantes correspondent à 3,4% du volume de sinistres et à 1% du coût total des prestations.

Par ailleurs, une analyse plus approfondie de ces données manquantes montre qu'elles concernent toutes les régions, tous les âges et les deux sexes qu'ils soient affiliés à la CMU ou non. Elles sont donc absentes de manière aléatoire. Cette conclusion a été déterminante dans le choix de la méthode de reconstitution des données afin d'éviter de perdre des informations en écartant certains profils d'individus de la base.

Pour ce faire, le principe d'imputation utilisé est celui proposé par le package MICE (*Multivariate Imputation by Chained Equations*[1]) du logiciel R. Supposant leur caractère aléatoire, cet algorithme prédit les données manquantes d'une variable donnée à partir des autres variables disponibles de la base. En fonction de la nature des données à prédire, différents modèles statistiques peuvent être choisis dans lesquels les variables disponibles constituent les variables explicatives et la variable manquante celle à expliquer. Après calibrage du modèle, seules les lignes manquantes de la variable incomplète sont imputées créant ainsi un nouveau jeu de données.

Dans cette étude, les variables présentant des données manquantes sont des variables qualitatives. L'algorithme a donc réalisé des régressions logistiques (binaire ou multinomiale).

Ainsi, 5 jeux d'imputations différents ont été réalisés pour chacune des variables concernées afin de tester la stabilité et la robustesse du modèle utilisé. La comparaison des distributions de coût moyen et de fréquence de sinistres avant et après imputation permet de conclure quant à la stabilité de la méthode employée. De plus, les résultats obtenus sont cohérents et montrent ainsi qu'il n'est pas nécessaire de recourir à des algorithmes plus complexes (arbres de décision, random forest, KNN,...) afin de reconstituer les données manquantes.

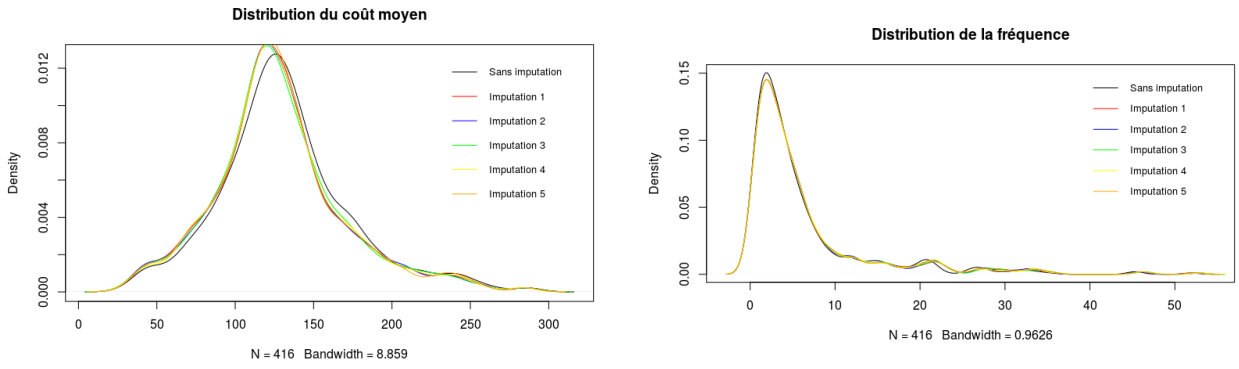


FIGURE 2.6 – Comparaison des distributions avant et après imputation

1. Imputation multivariée par équations chaînées

Afin de choisir l'imputation qui convient le mieux, l'écart quadratique moyen entre les distributions découlant des imputations et les distributions initiales a été évalué pour chaque jeu d'imputations : le choix s'est porté sur le jeu d'imputation 4 qui minimise cet écart tant sur les coûts moyens que sur les fréquences.

	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
Coût moyen	20,66	22,67	18,01	16,32	44,25
Fréquence	5,96	6,47	4,56	4,07	4,90

FIGURE 2.7 – Ecart quadratique moyen des distributions de coût moyen et de fréquence selon le jeu d'imputation

2.1.5 La base « sinistres » de modélisation

Les étapes présentées précédemment ont permis d'obtenir une base sans anomalies retraçant la consommation de soins hospitaliers en 2019, agrégée selon les caractéristiques des bénéficiaires, les actes médicaux et la période de remboursement des soins. Celle-ci comporte 385 835 lignes décrites par 8 variables :

Variables	Libellés	Modalités
FLX_ANN_MOI	Année et mois de remboursement	- 201901 à 201912
AGE_BEN_SNDS	Age du bénéficiaire	- 0 : 0 :19 ans - 20 : 20 à 29 ans - 30 : 30 à 39 ans - 40 : 40 à 49 ans - 50 : 50 à 59 ans - 60 : 60 à 69 ans - 70 : 70 à 79 ans - 80 : 80 ans et plus
BEN_CMU_TOP	Appartenance ou non au CMU	- 0 : non CMU - 1 : CMU
BEN_SEX_COD	Sexe du bénéficiaire	- 1 : masculin - 2 : féminin
BEN_RES_REG	Région de résidence du bénéficiaire	- 5 : Régions et Départements d'outre-mer - 11 : Ile-de-France - 24 : Centre-Val de Loire - 27 : Bourgogne-Franche-Comté - 28 : Normandie - 32 : Hauts-de-France - Nord-Pas-de-Calais-Picardie - 44 : Grand Est - 52 : Pays de la Loire - 53 : Bretagne - 75 : Aquitaine-Limousin-Poitou-Charentes - 76 : Languedoc-Roussillon-Midi-Pyrénées - 84 : Auvergne-Rhône-Alpes - 93 : Provence-Alpes-Côte d'Azur et Corse
PRS_NAT	Nature de la prestation (acte médical)	
PRS_PAI_MNT	Coût total de la prestation	
PRS_ACT_QTE	Nombre d'actes concernés	

FIGURE 2.8 – Variables de la base « sinistres »

Ainsi, la consommation de soins est décrite par la dépense engagée totale (part remboursée par la Sécurité sociale, Ticket Modérateur et dépassement d'honoraires) et le nombre d'actes concernés. La sinistralité à l'étude sera donc basée sur la quantité d'actes consommés.

2.2 Construction de la base « assurés »

La base Open DAMIR ne présentant que la consommation d'actes médicaux, il est nécessaire de construire une base d'assurés à laquelle sera adossée cette sinistralité. Les étapes pour y parvenir sont décrites ci-après.

2.2.1 Exploitation des données démographiques disponibles

En tant qu'assuré social, l'Assurance Maladie couvre toute la population présente sur le territoire français qui représente de fait la population d'assurés de cette étude. Les données démographiques mise à disposition par L'Institut National de la Statistique et des Etudes Economiques (INSEE) ont donc été exploitées pour construire une base ligne à ligne d'assurés respectant les profils de bénéficiaires présents dans la base « sinistres » à savoir ceux découlant du croisement des modalités des variables âge, sexe, région et CMU.

Pour y parvenir, la base INSEE 2019 présentant la répartition de la population nationale par sexe, départements et tranches d'âges quinquennales fut exploitée et adaptée selon les modalités des variables de segmentation susmentionnées. Les traitements réalisés sont explicités ci-après :

— Regroupement des départements en région selon la segmentation de la Sécurité sociale :

INSEE	Sécurité Sociale	Différence
Auvergne-Rhône-Alpes	Auvergne-Rhône-Alpes	Non
Bourgogne-Franche-Comté	Bourgogne-Franche-Comté	Non
Bretagne	Bretagne	Non
Centre-Val-de-Loire	Centre-Val de Loire	Non
Corse	Provence-Alpes-Côte d'Azur et Corse	Oui
Grand Est	Grand Est	Non
Hauts-de-France	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Non
Île-de-France	Ile-de-France	Non
Normandie	Normandie	Non
Nouvelle-Aquitaine	Aquitaine-Limousin-Poitou-Charentes	Non
Occitanie	Languedoc-Roussillon-Midi-Pyrénées	Non
Pays de la Loire	Pays de la Loire	Non
Provence-Alpes-Côte d'Azur	Provence-Alpes-Côte d'Azur et Corse	Oui
Guadeloupe	Régions et Départements d'outre-mer	Oui
Martinique	Régions et Départements d'outre-mer	Oui
Guyane	Régions et Départements d'outre-mer	Oui
La Réunion	Régions et Départements d'outre-mer	Oui
Mayotte	Régions et Départements d'outre-mer	Oui

FIGURE 2.9 – Comparaison des régions INSEE – Sécurité sociale

— Création de nouvelles tranches d'âges :

INSEE	Sécurité Sociale
0 à 4 ans	0-19 ANS
5 à 9 ans	0-19 ANS
10 à 14 ans	0-19 ANS
15 à 19 ans	0-19 ANS
20 à 24 ans	20 - 29 ANS
25 à 29 ans	20 - 29 ANS
30 à 34 ans	30 - 39 ANS
35 à 39 ans	30 - 39 ANS
40 à 44 ans	40 - 49 ANS
45 à 49 ans	40 - 49 ANS
50 à 54 ans	50 - 59 ANS
55 à 59 ans	50 - 59 ANS
60 à 64 ans	60 - 69 ANS
65 à 69 ans	60 - 69 ANS
70 à 74 ans	70 - 79 ANS
75 à 79 ans	70 - 79 ANS
80 à 84 ans	80 ANS ET +
85 à 89 ans	80 ANS ET +
90 à 94 ans	80 ANS ET +
95 ans et plus	80 ANS ET +

FIGURE 2.10 – Conversion des tranches d'âges INSEE en tranche d'âges considérées par la Sécurité sociale

2.2.2 Segmentation supplémentaire des données

Suite aux traitements précédents, la base ne comportait que les variables âge, région et sexe. Des traitements supplémentaies ont été nécessaires afin d'intégrer le critère d'affiliation à la CMU. Pour ce faire, les étapes suivantes ont été suivies :

— **Etape 1 : calcul du nombre d'assurés CMU 2019 par région**
Selon les données INSEE, le nombre total d'assurés à la CMU s'élevait à 5 176 157 d'habitants en 2019, soit 8% de la population totale. La répartition de cette population par département était disponible. Ces données ont été agrégées à la maille région pour calculer la part d'affiliés CMU dans chaque territoire.

— **Etape 2 : ventilation de la population CMU par tranche d'âge et sexe**
Les données du rapport d'activité 2018 du Fonds de financement de la Couverture maladie universelle ont été exploitées pour déterminer la ventilation des assurés CMU par tranche d'âge et par sexe. Les critères à respecter étaient les suivants :

Tranche d'âge	Poids dans la population totale	Poids femmes	Poids hommes
<10 ans	24,20%	48,80%	51,20%
<20 ans	19,40%	48,80%	51,20%
<40 ans	28,50%	52,50%	47,50%
<60 ans	22,10%	52,50%	47,50%
60 ans et plus	5,80%	52,50%	47,50%
Poids total		50,9%	49,1%
Nombre total d'assurés		2 864 927	2 765 073

FIGURE 2.11 – Répartition par âge et par sexe de la population CMU 2019

— **Etape 3 : création de la base finale**
Toutes les conditions précédentes ont été traitées par un solveur pour obtenir la base cible.

2.2.3 La base « assurés » de modélisation

La base de modélisation ainsi obtenue représente la population nationale 2019, soit 66 977 703 individus, ventilée selon 416 profils correspondant aux croisements des modalités des variables caractérisant les bénéficiaires de la base « sinistres ».

Par ailleurs, la couverture de la Sécurité sociale étant permanente, l'exposition annuelle considérée pour un profil d'individus donné correspondra au nombre d'individus concernés par ce profil.

Ainsi, la base « assurés » finale contient les variables suivantes :

— sexe ;
— région ;
— tranches d'âge ;
— affiliation CMU ;
— exposition.

2.3 Analyses univariées et bivariées du risque

Avant toute modélisation, une analyse des données a été réalisée afin d'évaluer de manière empirique l'impact de chaque variable tarifaire sur la consommation des soins hospitaliers. Ces analyses sont réalisées tous actes confondus et serviront de référentiel pour l'évaluation des coefficients tarifaires obtenus par modélisation statistique. Toutefois, des statistiques plus fines seront fournies *infra* selon la granularité des modélisations réalisées.

2.3.1 Vue d'ensemble

Au cours de l'année 2019, la Sécurité sociale a enregistré 271 433 161 actes hospitaliers consommés pour un coût total de 32 860 834 652€. Il en découle un coût moyen (CM) 121€ pour une fréquence (Frq) de 4,05 actes consommés en moyenne sur l'année.

Nombre de sinistres	Dépense engagée	Coût moyen	Fréquence	Exposition
271 433 161	32 860 834 652	121	4,05	66 977 703

FIGURE 2.12 – Bilan global de sinistralité 2019

2.3.2 Analyse univariée du coût moyen

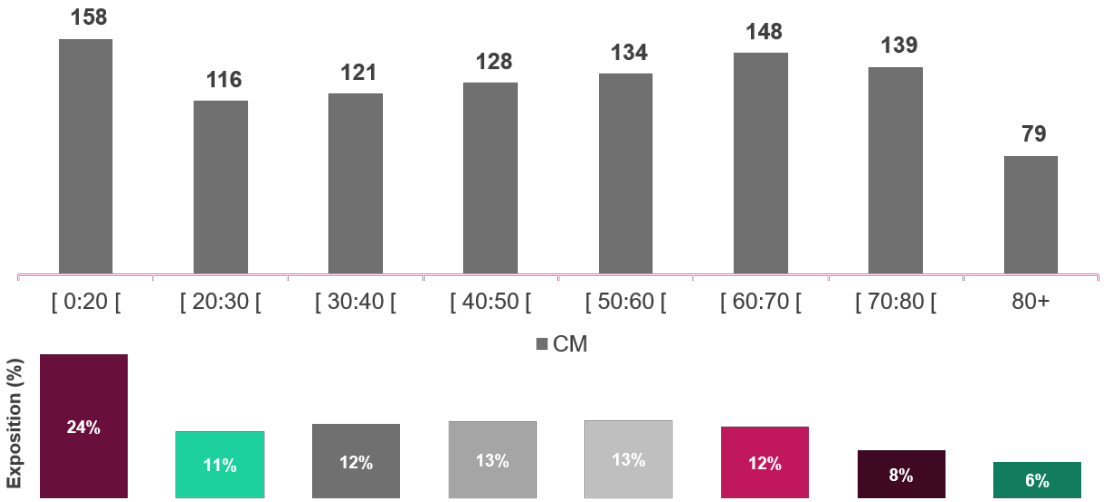
Influence de la variable âge

Au niveau des coûts moyens des actes médicaux consommés, les populations les plus jeunes et les plus âgées représentent les niveaux extrêmes : les jeunes de moins de 20 ans enregistrent le coût moyenle plus élevé tandis que les personnes âgées de 80 ans et plus, coûtent en moyenne moins chers que le reste de la population. Quant aux autres tranches d'âge, elles présentent des coût moyen croissants avec l'âge jusqu'à 69 ans avec un pic entre 60 et 69 ans. Ce pic est suivi d'une légère baisse des coûts (-6%) sur les âges compris entre 70 ans et 79 ans.

Ces constats sont cohérents avec deux phénomènes connus :

— les actes exploratoires de détection de maladies, particulièrement coûteux, sont majoritairement réalisés chez les jeunes ;

— les personnes âgées sollicitent en quantité des soins curatifs ou palliatifs de maladies déjà diagnostiquées d'où un coût moyen plus faible.

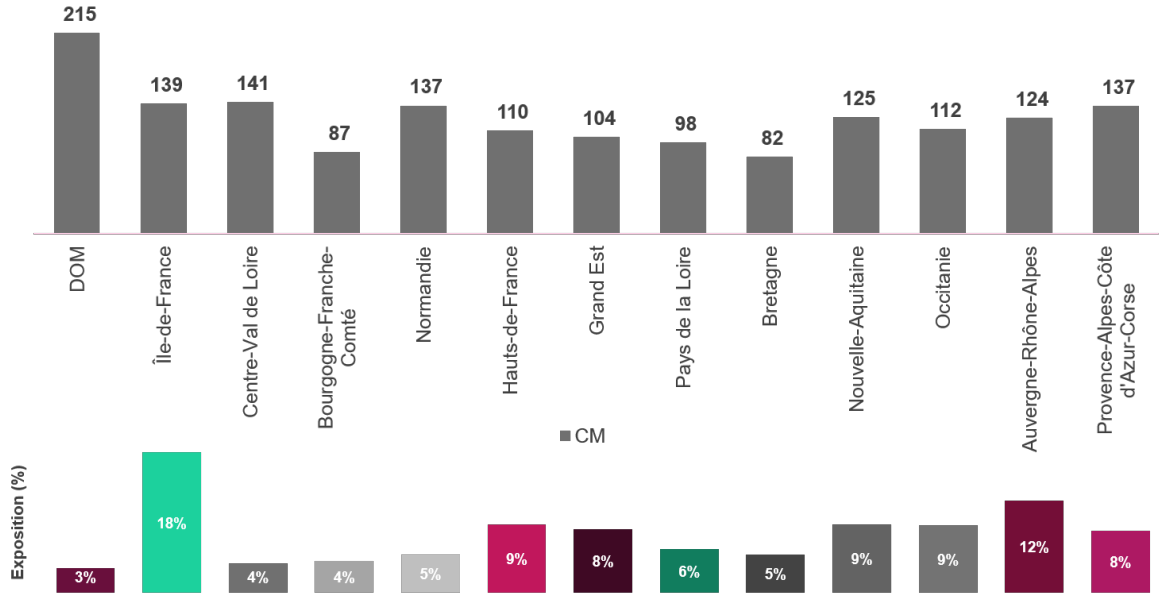


Tranche d'âge	Nombre de sinistres	Dépense engagée	Coût moyen	Exposition
[0:20[	13 393 407	2 109 717 434	158	16 175 671
[20:30[	13 763 590	1 597 692 927	116	7 522 501
[30:40[	17 554 813	2 125 204 144	121	8 294 289
[40:50[	23 970 528	3 074 441 406	128	8 658 984
[50:60[	35 306 025	4 741 221 200	134	8 783 646
[60:70[	44 183 163	6 550 651 157	148	7 995 019
[70:80[	48 375 361	6 712 316 470	139	5 443 092
80+	74 886 274	5 949 589 915	79	4 104 501
Total	**271 433 161**	**32 860 834 652**		**66 977 703**

FIGURE 2.13 – Coût moyen par tranche d'âge

Influence du lieu de résidence

L'analyse du coût moyen par région montre que les départements d'outre mer se distinguent par des coûts de prise en charge élevés pouvant aller de 1,5 à 2,6 fois les niveaux observés sur le reste du territoire. Cette observation est d'ailleurs en ligne avec la cherté de la vie enregistrée dans ces régions. Enfin, les coûts moyen les plus faibles sont quant à eux enregistrés en Bretagne et Bourgogne -Franche-Comté.

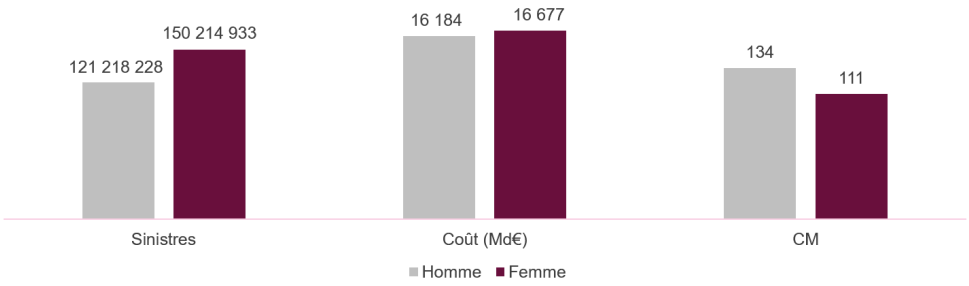


Région	Code région	Nombre de sinistres	Dépense engagée	Coût moyen	Exposition
DOM	5	5 824 540	1 250 606 065	215	2 155 749
Île-de-France	11	42 027 822	5 844 851 550	139	12 244 807
Centre-Val de Loire	24	8 757 655	1 237 779 923	141	2 565 258
Bourgogne-Franche-Comté	27	13 508 152	1 180 624 925	87	2 793 333
Normandie	28	11 733 126	1 610 295 266	137	3 313 340
Hauts-de-France	32	23 181 690	2 553 960 984	110	5 977 437
Grand Est	44	20 742 712	2 152 477 600	104	5 525 291
Pays de la Loire	52	15 661 063	1 535 543 664	98	3 787 400
Bretagne	53	14 035 031	1 156 712 202	82	3 333 720
Nouvelle-Aquitaine	75	22 120 733	2 764 161 924	125	5 986 318
Occitanie	76	36 540 417	4 100 417 839	112	5 899 209
Auvergne-Rhône-Alpes	84	28 486 310	3 524 915 484	124	8 005 882
Provence-Alpes-Côte d'Azur-Corse	93	28 813 910	3 948 487 227	137	5 389 959
Total		**271 433 161**	**32 860 834 652**		**66 977 703**

FIGURE 2.14 – Coût moyen par région

Influence du Sexe de l'assuré

Bien que règlementairement le sexe ne puisse être retenu comme variable tarifaire, l'analyse de son influence sur la sinistralité revêt une importance capitale dans l'appréhension du risque hospitalisation. Ainsi, les hommes affichent un coût moyen 20% plus élevé que celui de femmes. En effet, les niveaux de dépenses observés chez chacune des deux populations sont sensiblement égaux tandis que le volume de sinistres inhérent aux hommes est beaucoup plus faible que celui des femmes (-19%).

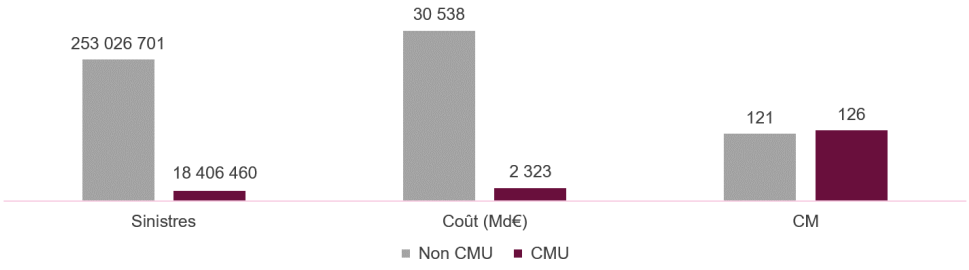


Sexe	Code sexe	Nombre de sinistres	Dépense engagée	Coût moyen	Exposition
Homme	1	121 218 228	16 184 325 989	134	32 369 504
Femme	2	150 214 933	16 676 508 662	111	34 608 199
Total		271 433 161	32 860 834 652		66 977 703

FIGURE 2.15 – Coût moyen par sexe

Influence du statut CMU

En moyenne, les affiliés à la CMU ont des coûts supérieurs de seulement 5 euros au reste de la population. L'affiliation à la CMU ne semble donc pas être un facteur déterminant du coût moyen des soins hospitaliers.



Statut CMU	Code CMU	Nombre de sinistres	Dépense engagée	Coût moyen	Exposition
Non CMU	0	253 026 701	30 537 583 922	121	61 801 545
CMU	1	18 406 460	2 323 250 730	126	5 176 158
Total		**271 433 161**	**32 860 834 652**		**66 977 703**

FIGURE 2.16 – Coût moyen selon l'affiliation CMU

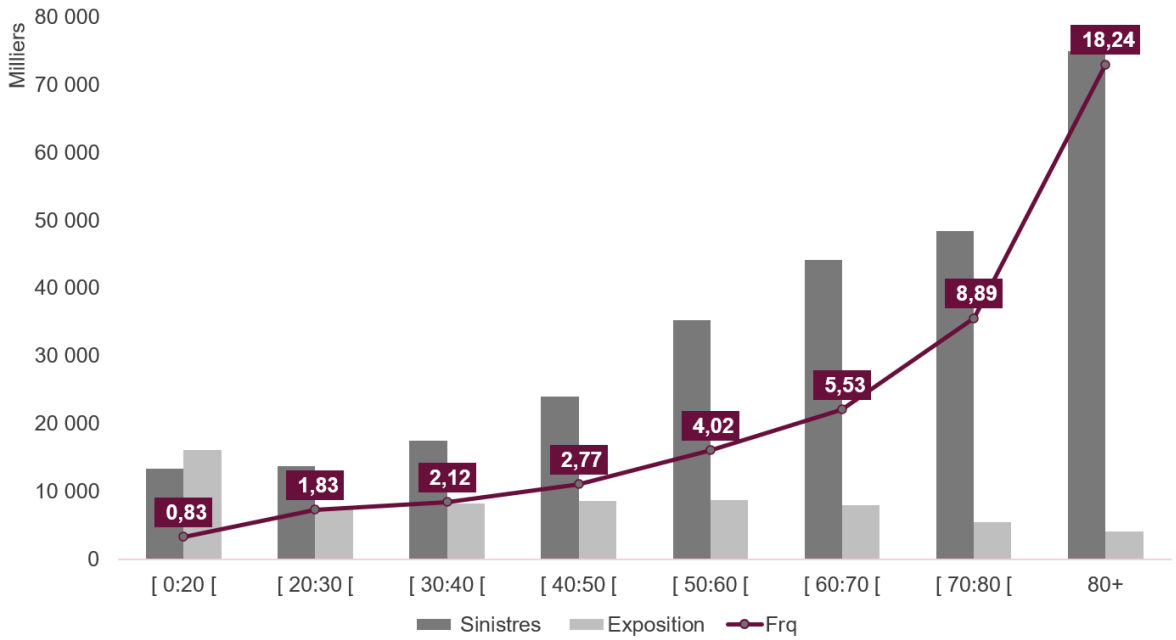
2.3.3 Analyse univariée de la fréquence

Dans toute la suite de cette étude, la fréquence analysée ne traduit pas le nombre moyen d'hospitalisation mais plutôt le rythme moyen de consommation des actes hospitaliers.

Influence de la variable âge

La fréquence de consommation des actes hospitaliers est croissante avec l'âge. En effet, celle-ci se révèle être très faible chez les plus jeunes puis est ensuite multipliée par trois entre le début de la vie active et la retraite. Elle continue ensuite de croitre sur les âges les plus élevés.

Ainsi, ce schéma souligne comme attendu une détérioration de l'état de santé des assurés avec l'âge.



Tranche d'âge	Nombre de sinistres	Exposition	Fréquence
[0:20[	13 393 407	16 175 671	0,83
[20:30[	13 763 590	7 522 501	1,83
[30:40[	17 554 813	8 294 289	2,12
[40:50[	23 970 528	8 658 984	2,77
[50:60[	35 306 025	8 783 646	4,02
[60:70[	44 183 163	7 995 019	5,53
[70:80[	48 375 361	5 443 092	8,89
80+	74 886 274	4 104 501	18,24
Total	**271 433 161**	**66 977 703**	

FIGURE 2.17 – Fréquence par tranche d'âge

Influence du lieu de résidence

Les fréquences de consommation moyennes observées sur le territoire vont de 2,7 à 6,2 actes par an et varient plus ou moins d'une région à l'autre.

Les niveaux les plus faibles sont enregistrés dans les départements d'outre mer et s'expliquent notamment par :

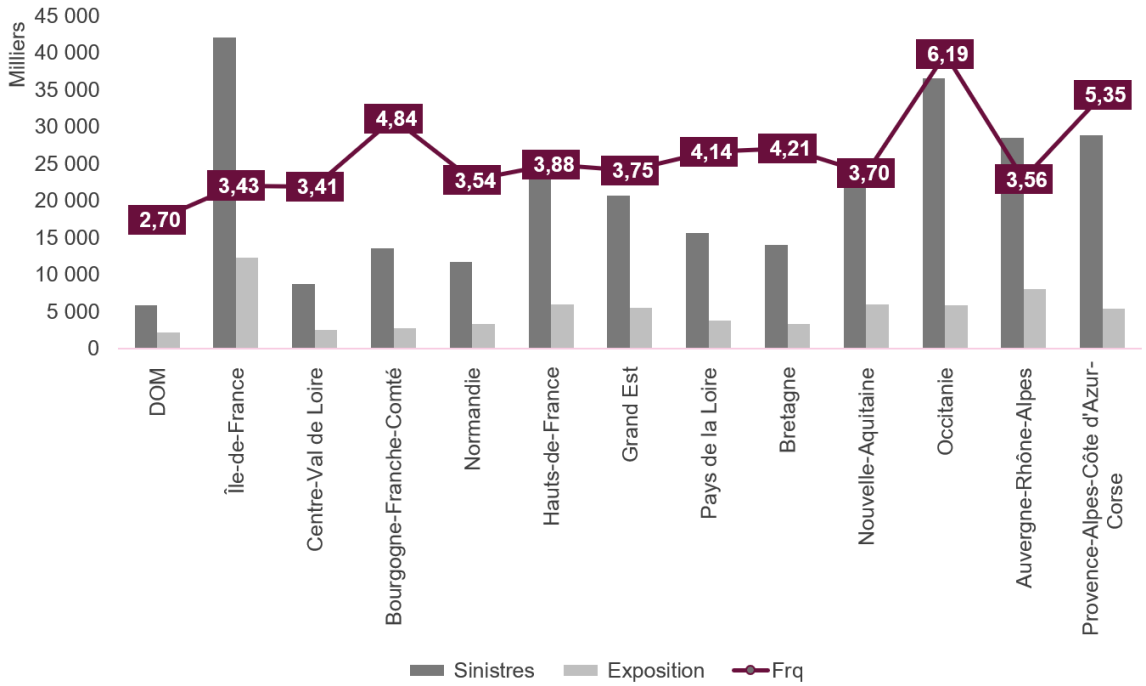
— un niveau de vie faible : cette région a le taux de pauvreté le plus élevé des départements français (36,2 contre 14,26 en moyenne sur les autres régions) [2] ne permettant pas aux populations d'aller dans des cliniques et de s'exposer à des dépassements d'honoraires ;

— une offre de soins réduite : l'offre de soins y est la plus faible des régions de France avec une densité de lits et places d'hospitalisation de 526 [3] pour 100 000 habitants contre 726 en moyenne sur les autres régions ;

— une population plus jeune qu'ailleurs : avec 32% de la population ayant moins de 20 ans contre 24% en moyenne enregistrée sur les autres régions, il est cohérent d'y observer un faible recours à des soins hospitaliers par rapport aux autres régions.

Quant aux niveaux élevés de fréquence, ils sont enregistrés en Bourgogne-Franche-Comté, Occitanie et en Provence-Alpes-Côte d'Azur-Corse. Ces régions ont la particularité d'avoir les populations les plus vieillissantes traduit par des indices de vieillessement [4] de l'ordre de 100 pour une moyenne nationale de 83,23.

2. INSEE 2016, moyenne pondérée sur la base des taux disponibles de la Martinique et de la Réunion
3. ATIH - 2018
4. Nombre d'adultes de 65 ans et plus pour un jeune de moins de 20 ans - données de population INSEE 2019

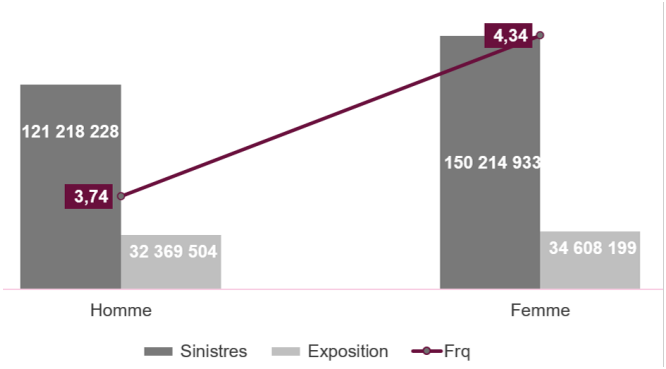


Region	Code région	Nombre de sinistres	Exposition	Fréquence
DOM	5	5 824 540	2 155 749	2,70
Île-de-France	11	42 027 822	12 244 807	3,43
Centre-Val de Loire	24	8 757 655	2 565 258	3,41
Bourgogne-Franche-Comté	27	13 508 152	2 793 333	4,84
Normandie	28	11 733 126	3 313 340	3,54
Hauts-de-France	32	23 181 690	5 977 437	3,88
Grand Est	44	20 742 712	5 525 291	3,75
Pays de la Loire	52	15 661 063	3 787 400	4,14
Bretagne	53	14 035 031	3 333 720	4,21
Nouvelle-Aquitaine	75	22 120 733	5 986 318	3,70
Occitanie	76	36 540 417	5 899 209	6,19
Auvergne-Rhône-Alpes	84	28 486 310	8 005 882	3,56
Provence-Alpes-Côte d'Azur-Corse	93	28 813 910	5 389 959	5,35
Total		**271 433 161**	**66 977 703**	

FIGURE 2.18 – Fréquence par région

Influence du Sexe de l'assuré

Contrairement au constat fait sur le coût moyen, les hommes ont une fréquence plus faible (-14%) que les femmes s'expliquant, entre autres, par une population masculine plus jeune comme susmentionné.

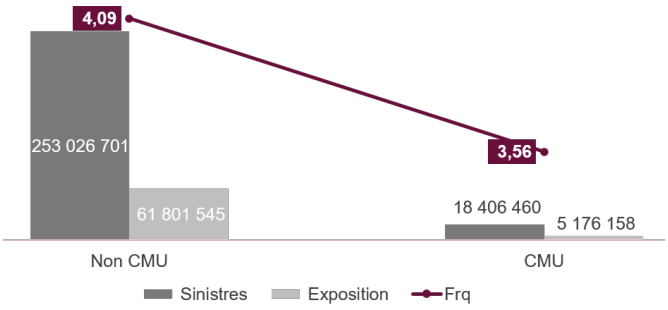


Sexe	Code sexe	Nombre de Sinistres	Exposition	Fréquence
Homme	1	121 218 228	32 369 504	3,74
Femme	2	150 214 933	34 608 199	4,34
Total		271 433 161	66 977 703	

FIGURE 2.19 – Fréquence selon la région

Influence du statut CMU

L'analyse des coûts moyens a montré que l'affiliation à la CMU inluençait peu les coûts des sinistres. Toutefois, sur la fréquence, la variable se révèle impactante puisqu'un différentiel de 15% est observable entre les deux populations (cf. figure **??**). Etant sur un périmètre restreint au secteur privé, ce constat est cohérent avec la faiblesse des revenus de cette population non couverte par le fonds CMU en cas de dépassements d'honoraires.



Statut CMU	Code CMU	Nombre de Sinistres	Exposition	Fréquence
Non CMU	0	253 026 701	61 801 545	4,09
CMU	1	18 406 460	5 176 158	3,56
Total		271 433 161	66 977 703	

FIGURE 2.20 – Fréquence selon le statut CMU

2.3.4 Analyse bivariée du coût moyen

Au delà des tendances globales vues *infra*, la mise en relation des différentes variables explicatives du coût moyen et ainsi de mettre en exergue les tendances suivantes :

— Si la vision univariée par sexe présentait un coût moyen hommes supérieur à celui des femmes, sur la tranche d'âges [30,50 [, cette tendance est inversée. Ce constat découle de coûts plus élevés pour les femmes sur les actes sollicités par les deux sexes à ces âges.

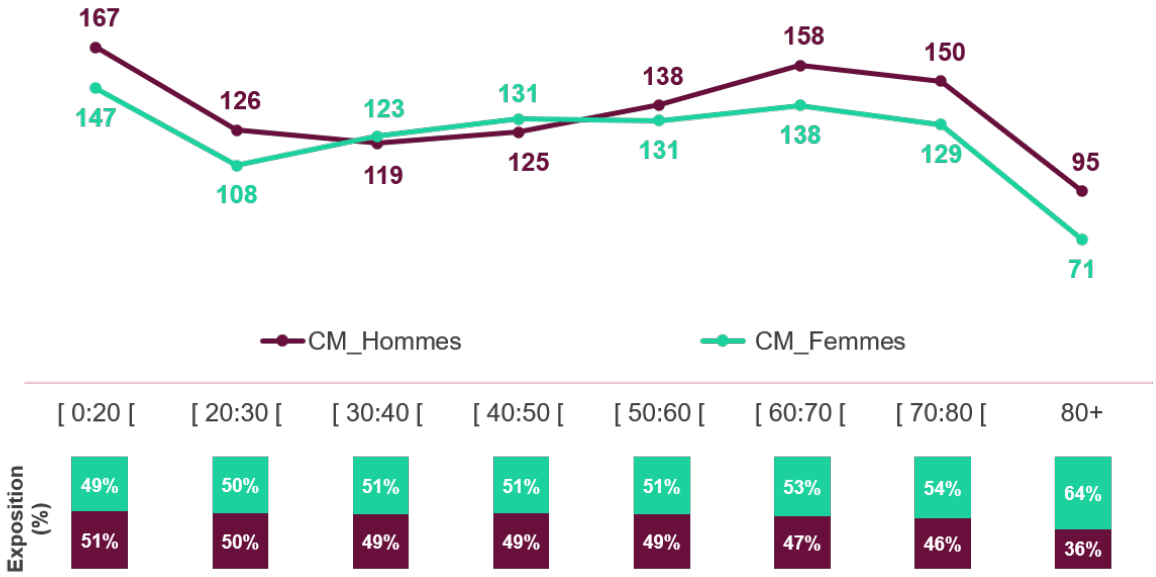


FIGURE 2.21 – Coût moyen par âge et par sexe

— Tous âges confondus, la population non affiliée à la CMU a un coût moyen légèrement moins élevé que celle à la CMU. Toutefois, la ventilation par âge montre que ce fait est tributaire de coûts moyens particulièrement faibles (comparativement à ceux de la population CMU) chez les assurés non CMU de 70 ans et plus.

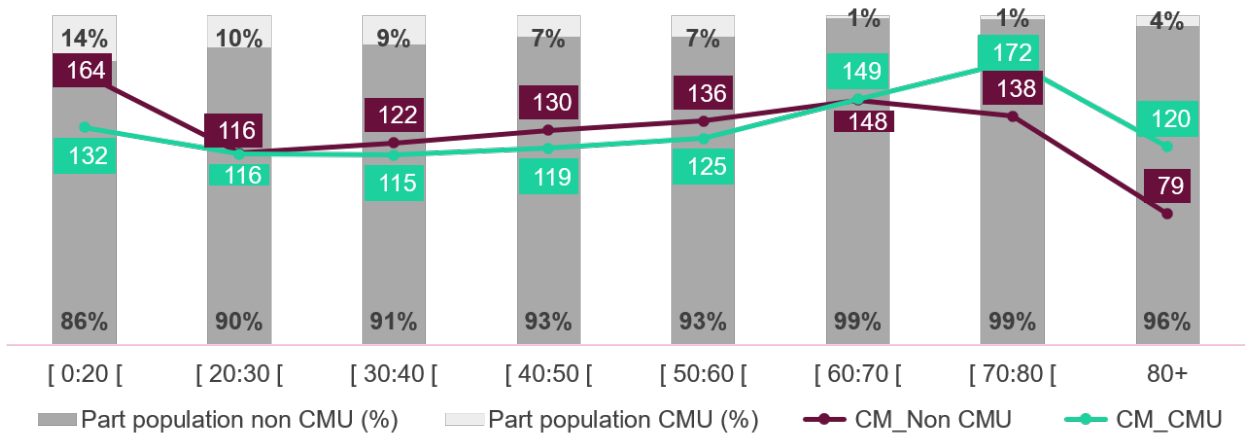


FIGURE 2.22 – Coût moyen par âge et selon le statut CMU

— Par ailleurs, aucune uniformité du coût moyen selon les régions et l'affiliation CMU n'est observable.

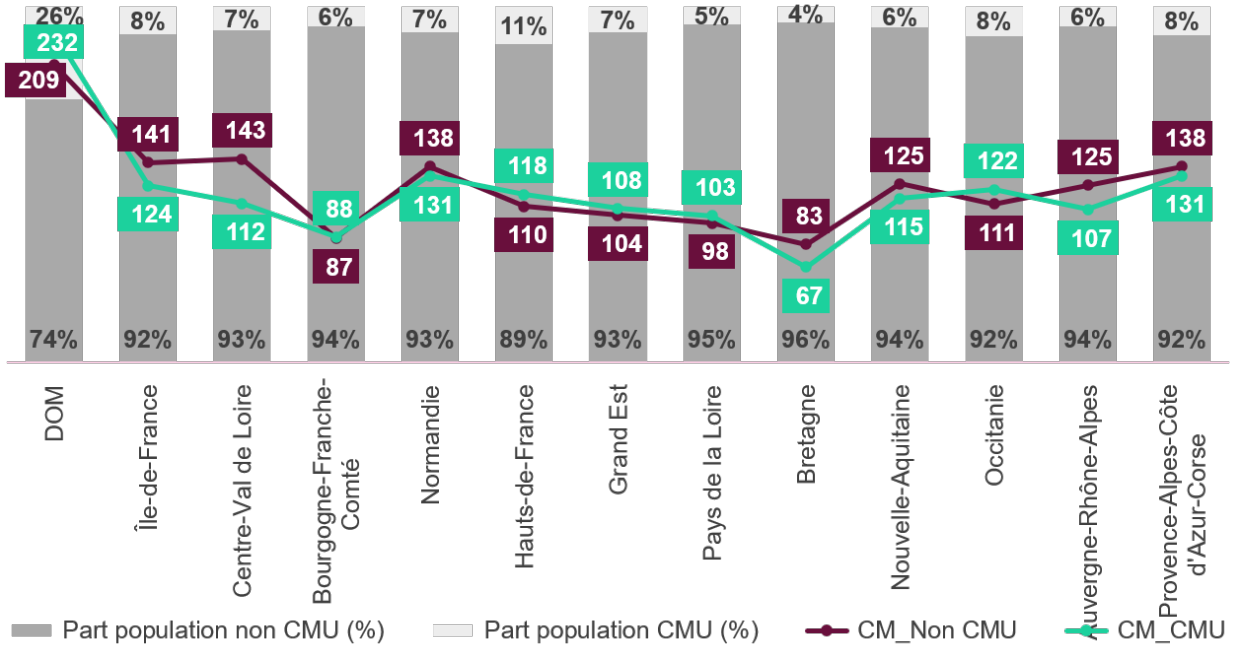


FIGURE 2.23 – Coût moyen par région et selon le statut CMU

— Il a été constaté que la chute particulièrement significative du coût moyen de la population âgée de plus de 80 ans n'était pas observée pour la région Ile-de-France, et ce, bien que la structure de sa population ne soit pas significativement différente des autres. Les autres régions affichent quant à elles des variations de coût moyen en fonction de l'âge identiques à celles observées au niveau portefeuille.

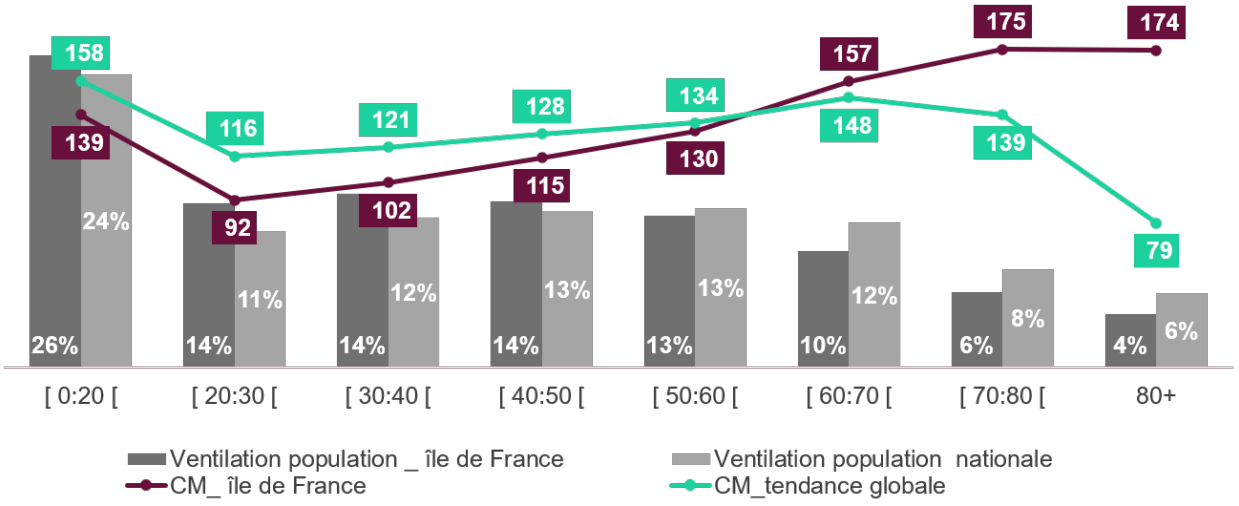


FIGURE 2.24 – Coût moyen par âge : comparaison Ile de France et tendance nationale

— En distinguant les hommes des femmes, le coût moyen en fonction du statut CMU suit des évolutions inverses : les hommes à la CMU ont un coût moyen inférieur de -6% à ceux hors CMU tandis que les femmes enregistrées à la CMU présentent un coût moyen supérieur de 15% à celles hors CMU. La tendance univariée sur la variable CMU est donc principalement influencée par la population féminine.

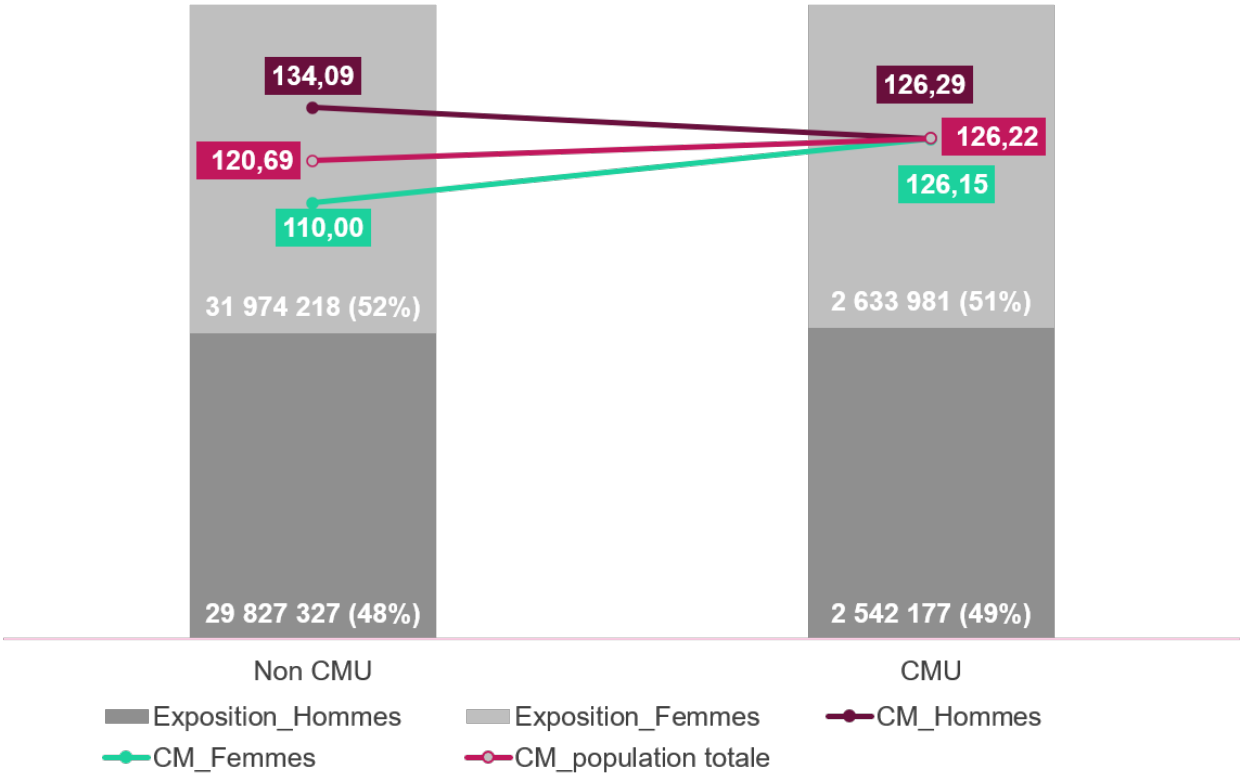


FIGURE 2.25 – Coût moyen par sexe et selon le statut CMU

2.3.5 Analyse bivariée de la fréquence

Les observations réalisées lors des analyses multivariées de la fréquence sont détaillées ci-après :

— La distinction de la population selon l'affiliation à la CMU met en exergue un pic de consommation des actes hospitaliers à 60 ans chez la population CMU au lieu de 80 ans observé sur la population hors CMU. Ce pic représente 5,4 fois la fréquence enregistrée sur la tranche d'âge [60 : 70 [chez la population hors CMU. Ce pic est observé dans toutes les régions, tant chez les hommes que chez les femmes.

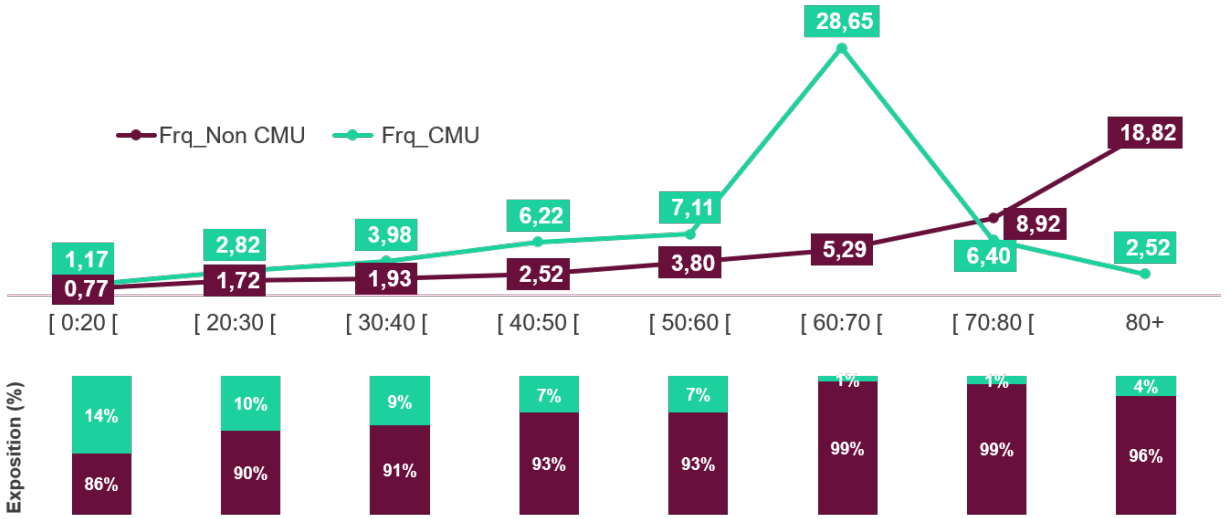


FIGURE 2.26 – Fréquence par tranche d'âge et selon le statut CMU

— Comme le montre la figure suivante, les écarts de fréquence observés entre les populations CMU et non CMU varient plus ou moins significativement selon la région :

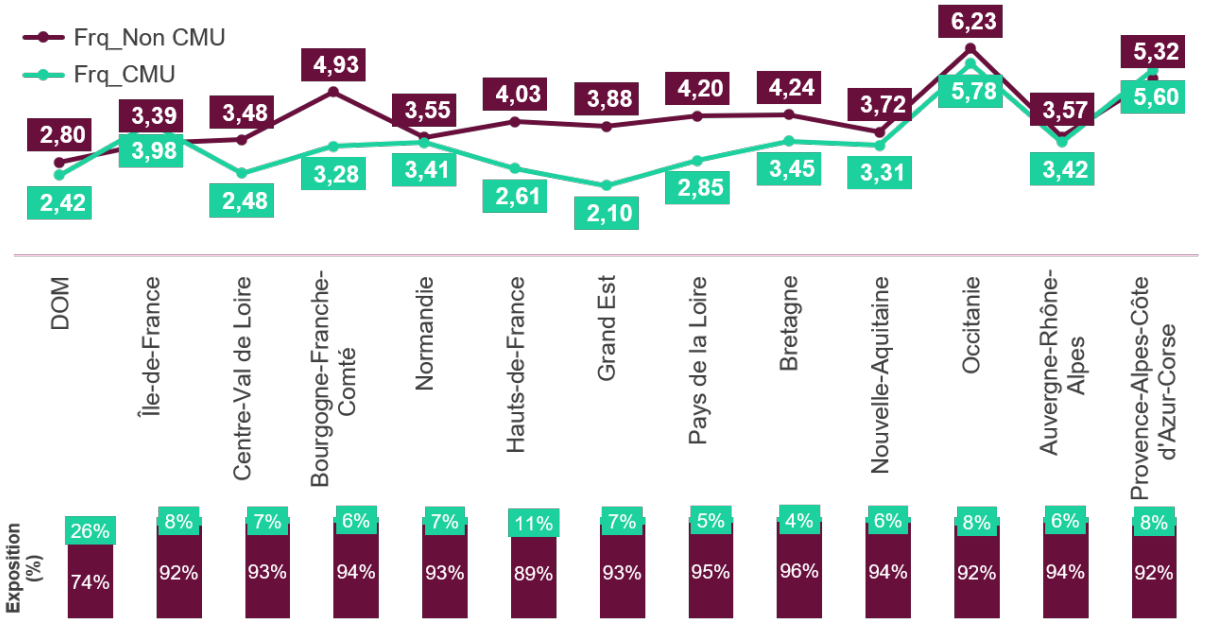


FIGURE 2.27 – Fréquence par région et selon le statut CMU

— Si en matière de fréquence la population hors CMU a une fréquence plus élevée que ceux enregistrés à la CMU, l'analyse par sexe présente un écart plus accentuée chez les femmes (+23%) que chez les hommes (+7%).

	Hommes	Femmes
Hors CMU	3,76	4,40
CMU	3,52	3,59
Ecart (%)	7%	23%

FIGURE 2.28 – Fréquence par sexe et selon le statut CMU

2.3.6 Analyse des dépendances entre variables

Dans les modèles modèles linéaires généralisés (*Generalized Linear model* - GLM), l'existence d'une multicolinéarité rend l'estimation des coefficients instable et introduit du biais dans les tests statistiques réalisés. Ainsi, afin d'éviter d'introduire des variables correlées dans les modélisations à réaliser, cette partie vise à évaluer l'intensité des liens entre les différentes variables explicatives. Toutes les variables étant catégorielles, c'est le V de Cramer qui sera utilisé.

Le test d'indépendance du Khi-deux (χ^2)

Pour tester l'existence d'une liaison entre deux variables qualitatives, il est usuel de recourir à un test d'indépendance du Khi-deux. Le principe consiste à dresser un tableau, dit de contingence, résumant les effectifs pour chaque croisement des modalités des deux variables afin de comparer la distribution qui en ressort à la distribution théorique à obtenir en cas d'indépendance desdites variables.

	Y_1	...	Y_j	...	Y_q	Totaux
X_1	$n_{1,1}$	...	$n_{1,j}$	...	$n_{1,q}$	$n_{1,.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_i	$n_{i,1}$	...	$n_{i,j}$	...	$n_{i,q}$	$n_{i,.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_p	$n_{p,1}$	...	$n_{p,j}$	...	$n_{p,q}$	$n_{p,.}$
Totaux	$n_{.,1}$	...	$n_{.,j}$	...	$n_{.,q}$	$n_{.,.}$

Avec :

— p : nombre de modalités de la variable X ;
— q : nombre de modalités de la variable Y ;
— n_ij : nombre d'individu avec la i-ème modalité de X et la j-ème modalité de Y ;
— $n_{i,.} = \sum_{j=1}^{q} n_{i,j}$;
— $n_{.,j} = \sum_{i=1}^{p} n_{i,j}$.

Ainsi, pour deux variables qualitatives X et Y, les hypothèses du test sont les suivantes :

$\mathcal{H}_0$: la distribution de X est indépendante de celle de Y ;
$\mathcal{H}_1$: les deux variables X et Y sont liées.

La statistique de test utilisée pour répondre à ces hypothèses est construite à partir du tableau de contingence présenté *supra* comme suit :

$$c_{i,j} = \frac{n_{i,.} \times n_{.,j}}{n_{.,.}}$$

Alors,

$$\chi^2(obs) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{i,j} - c_{i,j})^2}{c_{i,j}}$$

et, si $\mathcal{H}_0$ est vérifiée, $\chi^2(obs) \sim \chi^2_{(p-1)(q-1)}$.

Cette statistique indique donc si les fréquences d'observation des modalités de X dépendent des modalités de Y et permet de mesurer la probabilité (p-value) que les distributions des deux variables s'éloignent l'une de l'autre. Si cette probabilité est inférieure à un seuil donné (généralement fixé à 5%), alors il y a peu de chances que les variables soient indépendantes et l'hypothèse $\mathcal{H}_0$ est rejetée.

Le V de Cramer

Si le test d'indépendance du Khi-deux permet de statuer sur la dépendance entre les variables, il ne permet pas d'évaluer l'intensité de la liaison existante. Cette intensité est mesurée par le V de Cramer qui vient normaliser la statistique du test d'indépendance du Khi-deux (χ^2) afin d'obtenir une mesure comprise entre 0 et 1 dont l'interpretation est similaire à la correlation de Pearson : plus la valeur obtenue est proche de 1, plus les variables étudiées sont dépendantes.

Le V de Cramer, de par sa formule, vient également réduire l'instabilité de la statistique χ^2 qui reste dépendante de la taille de l'échantillon et du degré de liberté. En reprennant les notations de la partie précédente, le V de Cramer s'écrit comme suit :

$$V = \sqrt{\frac{\chi^2}{n \times (min(p,q) - 1)}}$$

Application et résultat

L'estimation du V de Cramer entre les variables explicatives disponibles permet de mettre en évidence l'indépendance des variables entre elles. Elles peuvent donc toutes être introduites dans les modèles à construire.

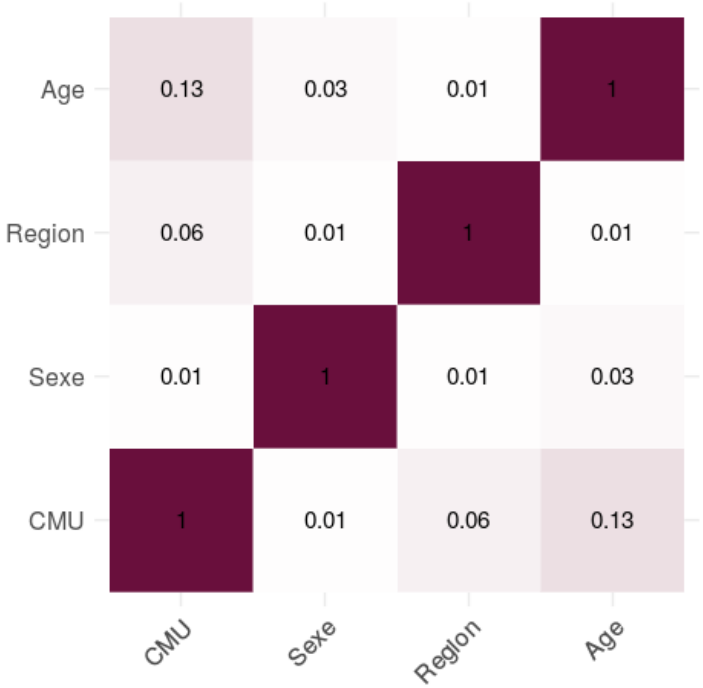


FIGURE 2.29 – V de Cramer

Chapitre 3

Performance de la segmentation par sous-postes de soins usuels dans la tarification du risque

Après l'analyse empirique du risque, l'objet de cette partie est de procéder à sa tarification. Celle-ci est usuellement réalisée à l'aide des modèles linéaires généralisés (GLM) après une segmentation des actes dépendant des garanties proposées. Si l'utilisation des GLM a été retenue comme contrainte de tarification, il est légitime de s'interroger sur l'optimalité statistique des segmentations usuelles. En effet, plus les actes d'un segment donné auront des densités similaires mieux les modèles tarifaires restrancriveront la réalité du risque. Ainsi, cette partie du mémoire vise à évaluer l'efficacité d'une ventilation des actes en sous-postes de soins sur la qualité de la tarification.

Pour ce faire, il sera présenté dans un premier temps les fondements théoriques des modèles linéaires généralisés utilisés pour la tarification. Deux tarifications différentes feront ensuite l'objet des autres sous-parties :

— Tarification 1 : modélisation tous actes confondus supposant l'iso-densité de tous les actes. Elle servira de référentiel de comparaison des performances des différentes segmentations des actes hospitaliers ;

— Tarification 2 : modélisation basée sur la segmentation en sous-postes de soins usuels.

3.1 Théorie des méthodes de tarification utilisées

La tarification du risque hospitalisation est réalisée à partir des modèles linéaires généralisés. L'objectif est de quantifier, via l'estimation de coefficients dits tarifaires, l'impact des variables influençant le risque en matière de fréquence d'une part et de coût de l'autre. Pour ce faire, les étapes suivantes ont été suivies :

- choix des distributions adéquates des variables à expliquer (coût moyen et fréquence de sinistres) ;
- sélection des variables explicatives ;
- estimation des coefficients tarifaires ;
- validation des résultats obtenus.

Pour chacun de ces points, les fondements théoriques seront présentés avant d'en détailler l'application pratique.

3.1.1 Théorie des modèles linéaires généralisés

Soit n individus pour lesquels il est mesuré les variables suivantes :

- $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, la variable à expliquer avec y_i la donnée correspondant au $i^{\text{ième}}$ individu $(i \in \{1, ..., n\})$;
- p variables explicatives tel que $X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n$ contient les observations de la $j^{\text{ième}}$ variable explicative $(j \in \{1, ..., p\})$ pour les n individus.

Par la suite, l'indépendance des n individus et le caractère déterministe des variables X_j sera supposée.

Rappel : la régression linéaire classique

Avant de présenter le modèle linéaire généralisé, il est primordial de comprendre le modèle linéaire gaussien.

La régression linéaire vise à expliquer l'évolution de la variable Y par celle des variables X_j en supposant qu'elles sont liées par la relation affine suivante pour chaque individu :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Qui peut être réécrite sous forme matricielle comme suit :

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

Avec :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_i \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p-1} & x_{1p} \\ 1 & x_{21} & \dots & \dots & \dots & x_{2p-1} & x_{2p} \\ 1 & x_{31} & \dots & \dots & \dots & x_{3p-1} & x_{3p} \\ 1 & x_{41} & \dots & \dots & \dots & x_{4p-1} & x_{4p} \\ \vdots & \vdots & \dots & x_{ij} & \dots & \vdots & \vdots \\ 1 & x_{i1} & \dots & \dots & \dots & x_{ip-1} & x_{ip} \\ \vdots & \vdots & \dots & \dots & \dots & \vdots & \vdots \\ 1 & x_{n-11} & \dots & \dots & \dots & x_{n-1p-1} & x_{n-1p} \\ 1 & x_{n1} & \dots & \dots & \dots & x_{np-1} & x_{np} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{pmatrix} \text{ et } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_{n-1} \\ \epsilon_n \end{pmatrix}$$

où :

— β_0, β_1, …, β_n sont des paramètres constants à estimer. Cette estimation se fait généralement par la méthode des moindres carrés tel que l'estimateur $\widehat{\beta}$ en découlant vérifie :

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

— $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ est le terme d'erreur aléatoire du modèle qui représente l'écart entre la valeur observée et la valeur estimée par combinaison des variables explicatives. Ce terme est supposé vérifier les hypothèses suivantes :

  — Espérance nulle : $\mathbb{E}[\epsilon_i] = 0, \forall\ i \in \{1, \dots, n\}$

  — Homoscédasticité : $Var[\epsilon_i] = \sigma^2, \forall\ i \in \{1, \dots, n\}$

  — Indépendance : $Cov[\epsilon_i \epsilon_j] = 0, \forall\ i \neq j \in \{1, \dots, n\}$

  — Distribution identique : $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \forall\ i \in \{1, \dots, n\}$

Ainsi, ce modèle suppose que la variable à expliquer Y suit une loi normale d'espérance $X\beta$ et de variance $\sigma^2\mathbb{I}_n$ [1]. Cette hypothèse forte rend inappropriée l'utilisation du modèle linéaire dans diverses situations, notamment pour la modélisation de variables ne pouvant prendre de valeurs négatives (nombre/montants de sinistres…).

Le modèle linéaire généralisé permet de supprimer cette contrainte de normalité de la variable à expliquer.

1. la matrice identité

Les modèles linéaires généralisés (GLM)

Dans le cadre du modèle linéaire généralisé, le caractère linéaire de la réponse obtenue par la combinaison linéaire des variables explicatives est maintenu $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Toutefois, le lien avec la variable à expliquer Y mais également la loi suivie par l'erreur de modélisation ϵ sont modifiés. En effet, le modèle linéaire généralisé se distingue du modèle linéaire gaussien par les composantes suivantes :

(i) **La loi de l'erreur ϵ (composante aléatoire)** : cette loi coïncide avec la distribution de probabilité que suit la variable à expliquer Y. Le modèle prévoit que cette loi soit de structure exponentielle.

Une densité f_Y appartient à la famille exponentielle si elle est de la forme :

$$f_Y = f_{\theta,\Phi}(y) = exp(\frac{y\theta - b(\theta)}{a(\Phi)} + c(y,\Phi))$$

Avec :

— $\theta \in \mathbb{R}$, le paramètre naturel de la famille exponentielle

— $\Phi \in \mathbb{R}$, le paramètre de dispersion

— $a(.)$ et $c(.)$, des fonctions dérivables

— $b(.)$, une fonction de classe $\mathbb{C}^3$ et dont la dérivée première est inversible

Une autre caractéristique majeure de cette famille est qu'elle vérifie les résultats suivants :

$$\mathbb{E}[Y] = b'(\theta)$$

$$Var[Y] = b''(\theta)\Phi$$

Dans le cas d'un processus de comptage, une loi de Poisson, de paramètre λ sera le choix privilégié et caractérisé par :

— $\theta = log(\lambda)$, $\Phi = 1$, $b(\theta) = exp(\theta) = \lambda$ et $c(y,\Phi) = -log(y!)$

— $f_Y = f_\lambda(y) = exp(ylog(\lambda) - \lambda - log(y!))$

— $\mathbb{E}[Y] = Var[Y] = \lambda$

(ii) **La fonction de lien** [2] **:**
le modèle linéaire présenté précédemment peut se réécrire à l'aide d'une fonction $f(.)$ tel que $f(\mathbb{E}[Y]) = X\beta$ avec $f(x) = x$, la fonction identité. Dans le cadre des GLM, cette fonction $f(.)$ s'adapte à la nature de la variable Y en imposant une contrainte sur le champ de définition des valeurs estimées par le modèle. Par exemple :

— la fonction logarithme népérien : $f(x) = log(x)$ permet par transformation inverse de respecter la nature d'une variable Y qui doit être positive ;

— la fonction logit : $f(x) = log(\frac{x}{1-x})$ permet, par transformation inverse, de respecter la nature d'une variable Y qui serait une probabilité.

2. Lien entre les variables explicatives et la variable à expliquer

Ainsi, dans le cas des GLM, l'équation à calibrer est toujours sous la forme $f(\mathbb{E}[Y]) = X\beta$ avec $\widehat{\beta}$ à estimer, mais cette fois par la méthode du maximum de vraisemblance.
Pour Y issue d'une famille exponentielle, l'expression de la vraisemblance s'écrit :

$$L(y_1, \ldots, y_n; \theta, \Phi) = \exp(\sum_{k=1}^{n} \frac{y_k\theta_k - b(\theta_k)}{a_k(\Phi)} + c(y_k, \Phi))$$

En découle ainsi l'expression de la log-vraisemblance :

$$log(L) = \sum_{k=1}^{n} \frac{y_k\theta_k - b(\theta_k)}{a_k(\Phi)} + c(y_k, \Phi)$$

Enfin, les paramètres $\widehat{\beta_i}$ sont obtenus suite à la résolution du système d'équation suivant :

$$\frac{\partial}{\partial\beta_i} log(L) = \sum_{k=1}^{n} \frac{\partial}{\partial\beta_i} (\frac{y_k\theta_k - b(\theta_k)}{a_k(\Phi)} + c(y_k, \Phi)) = 0$$

Ces estimations sont réalisées par le biais de l'algorithme d'optimisation de Newton-Raphson.

3.1.2 Théorie du modèle coût -fréquence pour la tarification

L'approche « coût-fréquence » est une méthode traditionnellement utilisée en tarification sur la branche non-vie, notamment en tarification santé. Elle consiste à estimer la sinistralité moyenne en modélisant distinctement la fréquence d'occurence des sinistres et leur coût moyen en supposant que ceux-ci sont indépendants. Ainsi, cette décomposition de la sinistralité permet à l'actuaire de prendre en compte les facteurs de risques les plus pertinents selon la grandeur modélisée.

Fondement du modèle

Soient les notations suivantes pour un groupe d'assurés donné :

— $N \in \mathbb{N}$: nombre total de sinistres enregistrés au cours d'une année donnée (variable aléatoire) ;
— $X_i \in \mathbb{R}+$: montant du i^{ieme} sinistre (ces montants sont supposés indépendants les uns des autres et identiquement distribués) ;
— $S = \sum_{i=1}^{N} X_i$ la charge sinistre totale.

L'objectif est d'obtenir une estimation moyenne du montant de sinistres attendu par l'assureur à horizon 1 an i.e. la prime pure annuelle. Mathématiquement, cette prime pure sur l'ensemble du portefeuille est égale à l'espérance des pertes $\mathbb{E}[S]$.

Pour l'estimer, la formule des espérances conditionnelles totales permet d'écrire l'égalité suivante :

$$\begin{aligned}
\mathbb{E}[S] &= \sum_{k=0}^{+\infty} \mathbb{E}[S|N=k] \times \mathbb{P}(N=k) \\
&= \sum_{k=0}^{+\infty} \mathbb{E}[\sum_{i=1}^{N} X_i|N=k] \times \mathbb{P}(N=k) \\
&= \sum_{k=0}^{+\infty} \mathbb{E}[\sum_{i=1}^{k} X_i|N=k] \times \mathbb{P}(N=k) \\
&= \sum_{k=0}^{+\infty} \mathbb{E}[\sum_{i=1}^{k} X_i] \times \mathbb{P}(N=k) \text{ car } X_i \perp N \\
&= \sum_{k=0}^{+\infty} \sum_{i=1}^{k} \mathbb{E}[X_1] \times \mathbb{P}(N=k) \text{ car } X_i \text{ i.i.d.} \\
&= \sum_{k=0}^{+\infty} k\mathbb{E}[X_1] \times \mathbb{P}(N=k) \\
&= \mathbb{E}[X_1] \sum_{k=0}^{+\infty} k\mathbb{P}(N=k)
\end{aligned}$$

$$\mathbb{E}[S] = \mathbb{E}[X_1] \times \mathbb{E}[N] = \text{coût moyen} \times \text{nombre moyen de sinistres}$$

Cette espérance rapportée à l'exposition globale [3] des assurés en portefeuille permet d'obtenir la prime pure selon le modèle « coût-fréquence » :

$$\text{Prime Pure} = \frac{\mathbb{E}[S]}{\text{Exposition}} = \mathbb{E}[X_1] \times \frac{\mathbb{E}[N]}{\text{Exposition}} = \text{coût moyen} \times \text{fréquence moyenne}$$

La modélisation du coût moyen et de la fréquence moyenne de sinistres nécessitera l'utilisation des modèles linéaires généralisés.

Modélisation du nombre de sinistres

Comme mentionné précédemment, afin d'établir la prime pure, il est nécessaire de déterminer le nombre probable de sinistres qui devront être indemnisés durant un temps de couverture donné. Cette variable correspond à un processus de comptage classiquement modélisé par une loi de Poisson supposant une équi-dispersion des données. Toutefois, lorsque cette propriété n'est pas observée, des lois alternatives sont utilisées.

La loi classique : loi de Poisson

Une variable X suit une loi de Poisson de paramètre $\lambda > 0$ ($X \sim P(\lambda)$) si pour tout $k \in \mathbb{N}$:

$$\mathbb{P}(X = k; \lambda) = \frac{\lambda^k \times exp(-\lambda)}{k!}$$

3. correspond à la somme des expositions individuelles des assurés. Un assuré donné bénéficiant d'un contrat prenant effet le 01/01/2020 avec comme date de fin de couverture le 30/06/2020, comptera pour 0,5 personne pour l'année 2020.

Cette loi garantie des nombres de sinistres entiers et positifs et vérifie :

$$\mathbb{E}[X] = Var[X] = \lambda$$

Ainsi, l'estimation du nombre de sinistres d'un individu i en fonction de diverses variables explicatives sera réalisé à travers un modèle linéaire généralisé avec la fonction logarithmique $g(x) = ln(x)$ pour fonction de lien canonique.

Pour y_i, la réalisation d'une loi de poisson de paramètre λ correspondant au nombre de sinistres de l'individu i, et $X_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}$, les variables explicatives influençant la valeur de y_i, l'équation de la régression s'écrira alors :

$$g(\mathbb{E}[y_i]) = ln(\lambda(x_{1i}, x_{2i} \dots x_{ni})) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Cela revient à supposer que $y_i \sim P(exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))$ et à estimer les coefficients par la méthode du maximum de vraisemblance présentée *supra*.

La question de la dispersion des données

Bien que la loi de Poisson soit théoriquement une loi adéquate pour la modélisation d'une variable discrète et positive, elle est régie par une hypothèse forte qui est l'équi-dispersion des données via l'égalité de l'espérance et de la variance. Toutefois, en pratique, il peut être observé des phénomènes de sur-dispersion ou de sous-dispersion (plus rare) des données :

$$Var(X) = \Phi\mathbb{E}[X] \text{ avec } \begin{cases} \Phi > 1 \text{ sur-dispersion} \\ \Phi < 1 \text{ sous-dispersion} \end{cases}$$

Plusieurs causes peuvent être à l'origine d'une sur-dispersion : par exemple dans le cas de la santé, elle peut découler d'une consommation importante de certains actes contrairement à d'autres qui conduit à avoir, en volume, des niveaux de sinistralité très faibles sur certains segments pour la variable réponse dans une base sinistres par actes. Par conséquent, l'unique paramètre de la loi de Poisson ne serait pas suffisant afin de rendre compte fidèlement du phénomène modélisé. Il sera donc nécessaire d'utiliser des lois alternatives car la présence de sur/sous-dispersion peut affecter les estimations des statistiques qui interviennent dans la sélection des variables explicatives et l'évaluation des modèles.

Les lois alternatives

Nous retrouvons dans la littérature plusieurs alternatives permettant de prendre en compte le phénomène de sur-dispersion, dont les principales sont :

(i) **Le modèle quasi-Poisson :** permet de corriger les écarts-types des paramètres β par l'intermédiaire de l'estimation d'un coefficient de sur-dispersion Φ. Ainsi, pour chaque individu i on a :

$$\mathbb{E}[y_i] = \lambda_i \text{ et } Var(y_i) = \Phi\lambda_i, \Phi > 1$$

Ce coefficient est estimé par le coefficient de Pearson généralisé :

$$\widehat{\Phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{Var(\lambda_i)}}{n - p}$$

avec :

— n : le nombre d'observations considérées (nombre d'individu dans la base)

— p : le nombre de variables explicatives

Ce modèle permet d'obtenir des estimations des termes λ_i et β_j identiques aux estimations du modèle de Poisson tout en annulant l'hypothèse d'équi-dispersion des données.

(ii) **Le modèle binomial négatif :** c'est la distribution classiquement utilisée pour inclure une sur-dispersion. Celle-ci est introduite via un paramètre τ_i permettant d'inclure une hétérogénéité dans l'espérance conditionnelle de la loi de Poisson telle que pour chaque individu i, les expressions de l'espérance conditionnelle et de la densité conditionnelle en fonction des variables explicatives sont les suivantes :

$$\mathbb{E}[y_i|X_i, \tau_i] = \exp(X_i^t \beta + \epsilon_i) = \mu_i \tau_i$$

$$f(y_i|X_i, \tau_i) = \exp(-\mu_i \tau_i) \frac{(\mu_i \tau_i)^{y_i}}{y_i!}$$

Le terme τ_i suit une loi Gamma d'espérance 1 et de variance $\frac{1}{\theta}$ qui permet de calculer la densité de la variable y_i conditionnellement à X_i :

$$f(y_i|X_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(y_i + 1)\Gamma(\theta)} (\frac{\theta}{\theta + \mu_i})^{\theta} \frac{\mu_i}{\mu_i + \theta}^{y_i}$$

Ce qui permet d'obtenir :

$$\mathbb{E}[y_i|X_i] = \mu_i \text{ et } Var[y_i|X_i] = (1 + \frac{\mu_i}{\theta})\mu_i = \Phi \mathbb{E}[y_i|X_i]$$

où $\Phi = (1 + \frac{\mu_i}{\theta}) > 1$ correspond au coefficient de dispersion à estimer.

Ainsi, la loi binomiale négative prend bien en compte l'effet de sur-dispersion.

Par ailleurs, ce modèle peut également être assimilé à un modèle quasi-Poisson où la variance d'une distribution de Poisson est multipliée par un paramètre estimé.

Modélisation du coût moyen des sinistres

Dans le cas du coût moyen des sinistres, il s'agit de modéliser une variable continue et positive. La loi normale classique n'est donc pas appropriée. Les distributions les plus utilisées dans ce cas sont les lois Gamma et log-normale. Le choix final portera sur la loi s'ajustant au mieux aux données étudiées.

Loi Gamma

Cette loi correspond aux caractéristiques du coût moyen des sinistres car elle admet des valeurs strictement positives et est asymétrique : sa queue de distribution de droite est nettement plus étendue que celle de gauche.

Mathématiquement, une variable $Y \geq 0$ suit une loi Gamma de paramètres $\alpha > 0$ et $\beta > 0$ si elle vérifie :

— Fonction de densité : $f(Y) = \frac{\beta^\alpha}{\Gamma(\alpha)} Y^{\alpha-1} exp(-\beta Y)$ avec $\Gamma(.)$ la fonction Gamma [4].

— Espérance : $\mathbb{E}[Y] = \frac{\alpha}{\beta}$

— Variance : $Var[Y] = \frac{\alpha}{\beta^2} = \Phi\mathbb{E}[Y]$ avec $\Phi = \frac{1}{\beta}$

Il est important de souligner que cette distribution peut prendre de nombreuses formes : par exemple, $\alpha = 1$ conduit à la loi exponentielle. De plus, étant donné le lien entre la moyenne et la variance, elle tient compte d'une éventuelle hétéroscédasticité dans les données. En matière de régression, bien que le lien canonique de la loi Gamma soit la fonction inverse, il est fréquent d'utiliser un lien logarithmique pour privilégier une forme multiplicative permettant des interprétations plus simples.

Loi log-normale

Contrairement à loi Normale, la loi log-normale permet de restreindre le champ de définition de la variable réponse à $\mathbb{R}+$. En effet, une variable aléatoire $Y \geq 0$ suit une loi log-normale de paramètres μ et σ (μ, $\sigma \in \mathbb{R}$) si elle vérifie :

— Fonction de densité : $f(Y) = \frac{1}{Y} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{ln(Y)-\mu}{\sigma})^2)$

— Espérance : $\mathbb{E}[Y] = \exp(\mu + \frac{\sigma^2}{2})$

— Variance : $Var[Y] = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$

L'utilisation de cette loi dans un modèle GLM se fait en ajustant le modèle pour $ln(Y)$ et non plus Y.

Passage à la prime pure

L'objectif final des modélisations susmentionnées est de déterminer la prime pure que devra facturer un assureur pour couvrir un risque donné. Ainsi dans le contexte de cette étude :

— L'espérance du nombre de sinistres d'un individu i donné s'écrit :

$$\mathbb{E}(y_i) = \exp(\widehat{\beta}_0^y + \widehat{\beta}_1^y x_{i1}^y + \ldots + \widehat{\beta}_p^y x_{ip}^y) \text{ , } \begin{cases} \widehat{\beta}_k^y \text{ , } k^{\text{ième}} \text{ coefficient tarifaire du modèle sur y} \\ x_{ik}^y k^{ime} \text{ variable explicative du modèle sur y} \end{cases}$$

— L'espérance du coût moyen :

$$\mathbb{E}(z_i) = \exp(\widehat{\beta}_0^z + \widehat{\beta}_1^z x_{i1}^z + \ldots + \widehat{\beta}_p^z x_{ip}^z)$$

— La prime pure (PP) :

$$PP = \mathbb{E}(y_i) \times \mathbb{E}(z_i)$$

Lorsque cette modélisation est effectuée sur plusieurs sous-segments d'une garantie donnée, la prime pure de la garantie correspond à la somme des primes pures obtenues sur chaque sous-segment.

4. $\Gamma(1) = 1$, $\forall$ x $\in \mathbb{R}+ : \Gamma(x+1) = x\Gamma(x)$ et $\forall$ n $\in \mathbb{N} : \Gamma(n) = (n-1)!$

3.1.3 Critères de choix de modèle

Dans le processus de modélisation, l'objectif est d'aboutir à un modèle retranscrivant au mieux la variable à expliquer. Il est donc important de sélectionner minutieusement les variables explicatives. Pour ce faire, plusieurs méthodes et critères de sélection existent dont les plus utilisés sont présentés ci-après.

La sélection de variables

La sélection de variables vise à identifier la meilleure combinaison de variables explicatives à retenir. Le principe est d'inclure dans le modèle uniquement celles ayant une influence réelle sur la variable à expliquer. L'objectif est d'éviter d'avoir un modèle complexe qui intègrerait, sans tri, toutes les variables explicatives disponibles pour la modélisation. Pour ce faire, il existe plusieurs méthodes dont les plus couramment utilisées sont :

— **La sélection *forward*** : le point de départ est le modèle vierge (la variable réponse dépend uniquement d'une constante) auquel on rajoute une à une les potentielles variables explicatives. Si l'ajout d'une variable explicative donnée améliore la qualité du modèle selon un critère donné, elle est retenue. Et ainsi de suite, jusqu'à ce que toutes les variables explicatives aient été testées. Sachant que la significativité d'une variable donnée peut être impactée lorsqu'elle est combinée à une autre, l'inconvénient de cette méthode est que lorsqu'une variable est retenue, elle ne peut être écartée ensuite au fil de l'ajout des autres variables.

— **La sélection *backward*** : c'est le procédé inverse de la méthode *forward*. En effet, le modèle complet, i.e. contenant toutes les variables explicatives, est calibré puis il lui est retiré une variable à la fois. Si le modèle sans une variable donnée s'avère meilleur que le modèle l'incluant, la variable n'est pas retenue pour la modélisation.

— **La sélection *stepwise*** : c'est une combinaison des méthodes *forward* et *backward* qui, contrairement à la méthode *forward*, permet à chaque nouveau modèle calibré de retirer une variable précédemment testée s'il s'avère qu'elle n'est finalement plus significative.

— **La sélection exhaustive** : elle consiste à tester toutes les combinaisons possibles de variables afin de sélectionner le meilleur modèle selon un critère donné. L'inconvénient est le temps de calcul qui peut être considérable, particulièrement si le nombre de variables explicatives est important.

Les critères de comparaison de modèle

La mise en œuvre des méthodes de sélection de variables, vues *supra*, nécessite des critères d'évaluation de qualité des différents modèles qui seront testés pour en retenir le meilleur. Plusieurs critères peuvent être considérés dont :

— **La déviance** : critère basé sur la vraisemblance du modèle. Plus la vraisemblance est proche de 1, mieux le modèle s'ajuste aux données. La déviance compare donc la vraisemblance du modèle calibré à celle du modèle parfait en matière d'adéquation aux données, dit modèle saturé. L'expression de la déviance est la suivante :

$$D = 2(\underbrace{\log L(y, y, \Phi)}_{\text{modèle saturé}} - \underbrace{\log L(y, \mu, \Phi)}_{\text{modèle calibré}})$$

Ainsi, le meilleur modèle sera celui qui aura la déviance la plus faible. Toutefois, il est important de préciser que ce critère est sensible, à l'instar de la somme des carrés des résidus dans la régression linéaire, au nombre de variables considérées dans les différents modèles à comparer : moins il y a de variables, moins il y a d'information dans le modèle, ce qui conduira à une augmentation de la déviance. Ainsi, la déviance est à privilégier dans la comparaison de modèles ayant le même nombre de variables explicatives, et non pour juger si l'ajout d'une variable ou non dans un modèle donné apporte de la précision.

— **Les critères AIC et BIC** : critères apportant une information sur le niveau des maximums de vraisemblance des modèles potentiels tout en pénalisant ceux comportant un nombre important de variables pouvant conduire à du sur-apprentissage [5]. Ils permettent ainsi un arbitrage entre l'information apportée par des variables supplémentaires et la complexité du modèle (nombre de paramètres à estimer).

L'AIC et le BIC sont donnés par les formules suivantes :

$$AIC = -2log(L) + 2k$$

$$BIC = -2log(L) + klog(n)$$

avec :

- $log(L)$ la log vraisemblance
- k le nombre de paramètres estimés
- n le nombre d'observations

Ainsi, le modèle à retenir est celui minimisant ces critères. En effet, l'ajout d'une variable pertinente fera décroître ces indicateurs tandis que l'ajout de variables peu significatives, ou induisant une redondance d'information, conduira à leur augmentation.

— **L'erreur quadratique moyenne (RMSE)** : mesure la qualité de prédiction d'un modèle en s'intéressant aux résidus induits par cette prédiction. Elle est très souvent utilisée dans un processus de validation croisée consistant à scinder l'échantillon de données disponibles en deux parties pour calibrer le modèle sur une première partie, dite base d'apprentissage, puis réaliser la prédiction sur la seconde partie appelée base de test. Ainsi, la racine carrée de la moyenne du carré des écarts entre chaque valeur prédite et sa valeur réelle constitue la RMSE :

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

avec :

- y_i la valeur observée de la variable à expliquer pour l'individu i
- $\widehat{y_i}$ la valeur prédite de la variable à expliquer pour l'individu i
- n le nombre d'observations dans la base de test

Le modèle à privilégier sera alors celui qui présente la RMSE la plus faible.

5. reproduction parfaite des données existantes et difficultés à prédire sur un échantillon tiers.

3.2 Constitution d'une base d'apprentissage et d'une base de test

Pour réaliser cette étude, la base de données sera segmentée en base d'apprentissage et base de test. La base d'apprentissage servira de base de calibrage des modèles et la base de test permettra de comparer les différentes méthodes de tarification sur la qualité de leur pouvoir prédictif.

Toutefois, la base de données étant une base en *model point*, le mode de constitution classique d'une base d'apprentissage et d'une base de test ne peut s'appliquer, au risque d'écarter de l'une ou l'autre des bases certains profils d'individus. Un autre procédé a donc été retenu : une sélection aléatoire des données de 10 mois de soins parmi les 12 observés pour constituer la base d'apprentissage et les 2 mois restants représenteront la base de test.

Une telle segmentation pose évidemment la question de la stabilité mensuelle de la fréquence de sinistres et des coûts des soins. Ainsi, pour éviter le biais que pourrait induire cette segmentation dans la modélisation, les coefficients tarifaires des différents modèles seront évalués sur toutes les combinaisons de base d'apprentissage/base de test possibles. Si des variations importantes sont observées, il sera retenu un niveau moyen pour chaque coefficient. Autrement, les résultats d'une seule combinaison de base d'apprentissage/base de test seront retenus.

3.3 Approche globale : tarification à la maille acte

Cette tarification est la première d'une serie de trois tarifications comme précisé *infra*. Elle a été réalisée en considérant que tous les actes présentaient les mêmes caractéristiques en matière de fréquence de consommation et de dépense engagée. C'est le modèle qui servira de base d'évaluation de l'apport des différents modes de segmentation à la tarification du risque. Cette partie en présente les résultats.

3.3.1 Modèle de coût

Choix de la distribution adéquate

Deux distributions candidates ont été testées : la loi Gamma et la loi Log-normale. Pour ce faire, leurs paramètres ont été estimés sur les données par la méthode du maximum de vraisemblance pour retenir les paramètres s'ajustant au mieux aux données. Les résultats sont présentés ci-après :

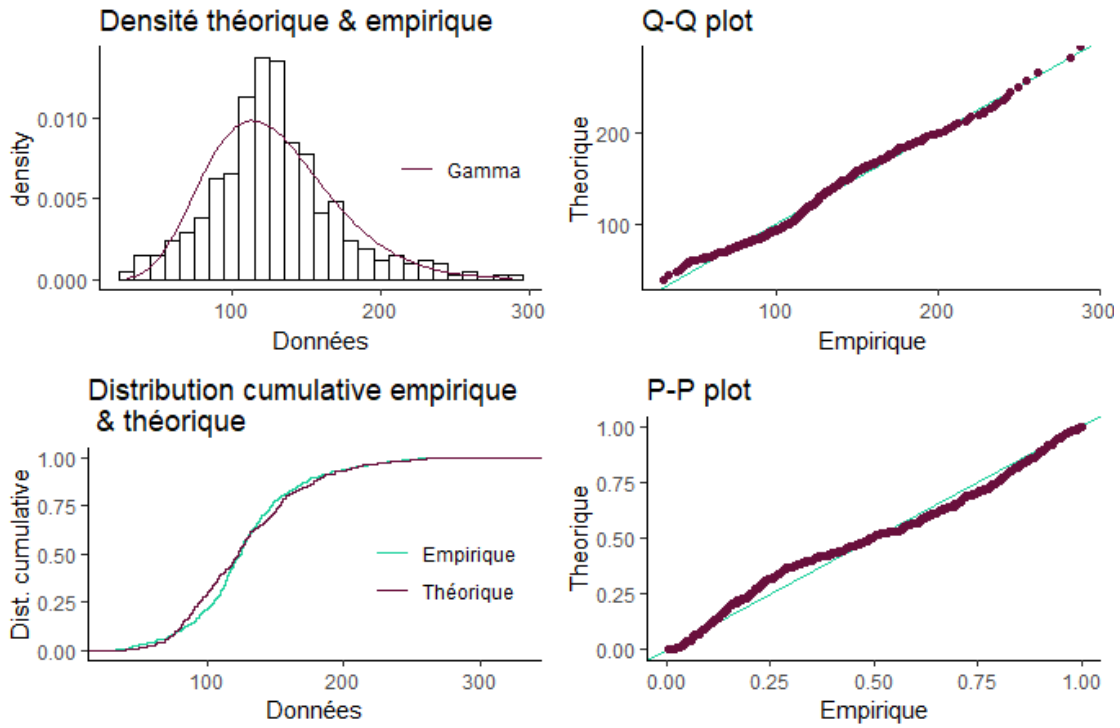


FIGURE 3.1 – Adéquation loi Gamma - modèle de coût - tarification globale

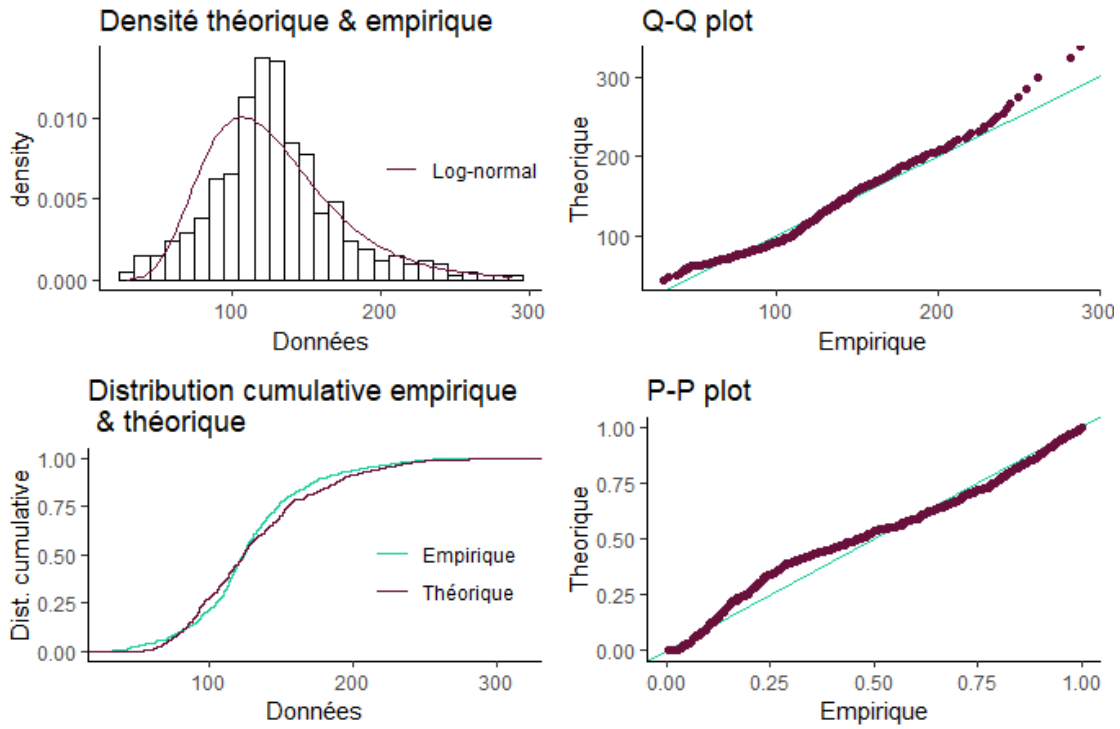


FIGURE 3.2 – Adéquation loi log-normale - modèle de coût - tarification globale

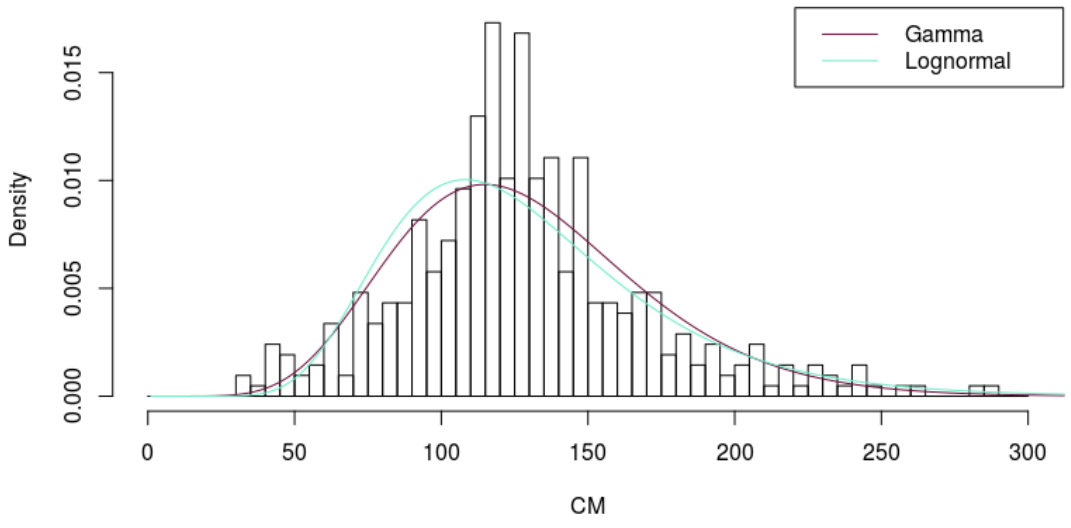


FIGURE 3.3 – Distribution du coût moyen tous actes confondus

L'analyse des graphiques précédents permet de constater que les loi Gamma et log-normal ne sont pas les plus adéquates pour la distribution des coûts moyens observés. Toutefois, les Q-Q plot montrent que la loi Gamma est la plus adéquate des deux avec des quantiles qui coïncident mieux avec les quantiles empiriques (points de la courbe suivant la droite x=y). C'est cette loi qui sera retenue pour la modélisation.

Sélection de variables

La sélection de variables a été réalisée selon la méthode exhaustive par l'utilisation de la fonction R « glmulti », qui permet de tester tous les modèles possibles en y intégrant les interactions d'ordre 2 pour en sortir le meilleur modèle selon un critère donné. Ici, le critère retenu a été l'AIC car, étant face à un nombre restreint de variables explicatives, le BIC, très pénalisant, réduirait encore plus le nombre de variables à retenir dans le modèle. Ainsi, 100 modèles ont été simulés et le meilleur modèle en découlant est le suivant :

$$\beta_0 + \beta_1 Age + \beta_2 Region + \beta_3 Sexe + \beta_4 CMU + \beta_5 Region : Age + \beta_6 Sexe : Age + \beta_7 Sexe : Region + \beta_8 CMU : Age + \beta_9 CMU : Region + \beta_{10} CMU : Sexe$$

Ce modèle est cohérent avec l'analyse multivariée des données réalisées précédemment car il a retenu les impacts bivariés sur lesquels des tendances particulières ont été observées. Toutefois, s'il présente l'AIC le plus faible, certaines variables retenues présentent finalement des coefficients peu significatifs : elles ont donc été retirées ou réduites à certaines modalités pour réduire le bruit que celà pourrait introduire dans le modèle. En effet, une analyse des coefficients du modèle et des p-values du test de Student [6] permet de montrer que les croisements Sexe : Age, Sexe : Region, CMU : Age et CMU : Region peuvent être omis du modèle. De plus, l'analyse de l'impact bivarié de l'âge et la région montre qu'une correction de la somme des effets marginaux des deux variables n'est nécessaire que sur la tranche d'âge 80 et plus.

Le modèle retenu *in fine* est alors :

$$\beta_0 + \beta_1 Age20 + \beta_2 Age30 + \beta_3 Age40 + \beta_4 Age50 + \beta_5 Age60 + \beta_6 Age70 + \beta_7 Age80 + \beta_8 Age80 : Region + \beta_9 CMU + \beta_{10} Sexe + \beta_{11} Region + \beta_{12} CMU : Sexe$$

Le détail des coefficients obtenus, les p-value associées ainsi que leur niveau de significativité [7] sont présenté ci-après :

Variable explicative	Estimation	p-value	Significativité
(Intercept)	5,66	8,70E-295	***
Age20	-0,3	1,20E-15	***
Age30	-0,25	9,81E-13	***
Age40	-0,19	2,57E-09	***
Age50	-0,15	1,88E-07	***
Age60	-0,06	0,024231148	*
Age70	-0,12	7,45E-06	***
Age80	-0,4	0,000114467	***
Region11	-0,54	8,77E-28	***
Region24	-0,41	1,66E-12	***
Region27	-0,73	1,22E-34	***
Region28	-0,39	1,21E-11	***
Region32	-0,5	1,80E-22	***
Region44	-0,59	1,76E-28	***
Region52	-0,64	4,14E-30	***
Region53	-0,8	4,84E-42	***
Region75	-0,44	7,49E-18	***
Region76	-0,56	6,55E-28	***
Region84	-0,5	2,49E-23	***
Region93	-0,47	8,97E-21	***

Variable explicative	Estimation	p-value	Significativité
CMU1	-0,21	8,74E-10	***
Sexe2	-0,16	5,39E-33	***
Age80:Region11	0,54	1,43E-06	***
Age80:Region24	-0,08	0,519524423	
Age80:Region27	-0,57	9,33E-07	***
Age80:Region28	-0,39	0,001749985	**
Age80:Region32	-0,66	3,27E-09	***
Age80:Region44	-0,56	7,57E-07	***
Age80:Region52	-0,57	1,62E-06	***
Age80:Region53	-0,55	1,73E-06	***
Age80:Region75	-0,39	0,000368665	***
Age80:Region76	-0,36	0,000676157	***
Age80:Region84	-0,28	0,009919453	*
Age80:Region93	0,06	0,56913341	
CMU1:Sexe2	0,18	7,52607E-05	***

FIGURE 3.4 – Coefficients du modèle final de coût moyen - tarification globale

6. H0 : $\beta_i = 0$ avec 5% comme seuil critique de significativité.Si la p-value dépasse ce seuil alors la modalité associée à ce coefficient n'est pas significative et peut être omise du modèle ou regroupée avec une autre modalité.

7. * influence significative, ** influence très significative, *** influence hautement significative.

Analyse des coefficients

La première étape de cette analyse a été de vérifier la stabilité des coefficients obtenus en fonction des mois composant la base d'apprentissage. 66 simulations du modèle précédent ont donc été réalisées pour comparer la variation des coefficients obtenus. Il en ressort que peu importe la segmentation Base d'apprentissage/Base de test, les résultats sont stables : les courbes des coefficients obtenus se confondent comme le montre le graphique ci-dessous.

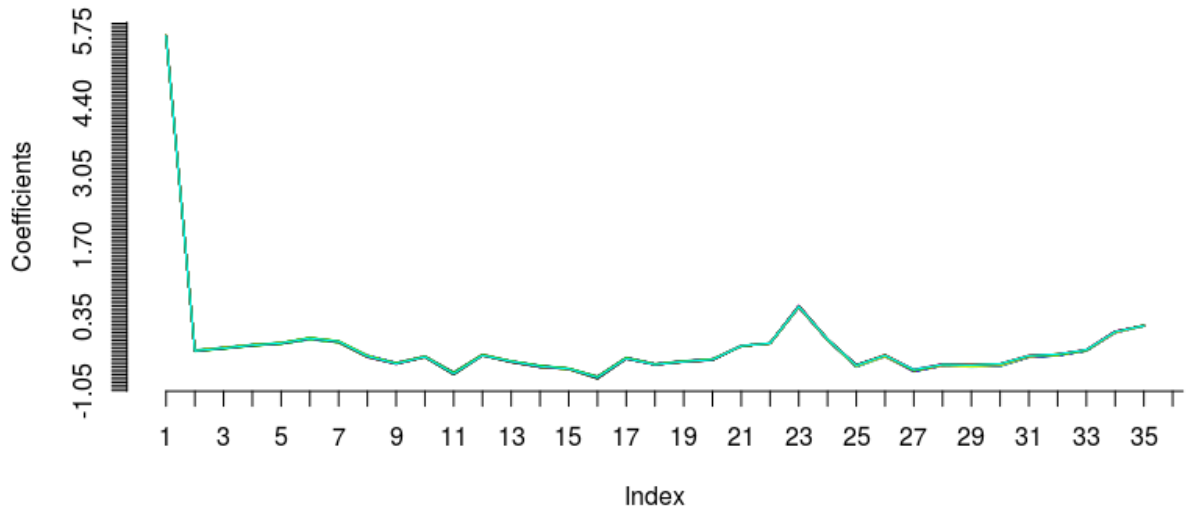


FIGURE 3.5 – Stabilité temporelle des coefficients tarifaires

Les coefficients du modèle présentés ci-dessus sont des coefficients bruts traduisant les effets de différenciation des différents profils d'assurés par rapport à un profil de référence dont les caractéristiques sont les suivantes :

- tranche d'âge : 0-19 ans (code : 0) ;
- sexe : masculin (code : 1) ;
- affiliation à la CMU : non CMU (code : 0) ;
- région de résidence : Outre-mer (code : 5).

Pour avoir les coefficients tarifaires finaux liés à chaque profil, les coefficients bruts sont à additionner à l'intercept qui est le coefficient tarifaire du profil de référence. Ces coefficients tarifaires ainsi obtenus sont comparés ci-après aux tendances observées lors de l'analyse descriptive des données pour évaluer leur cohérence.

— **Variable « âge »**

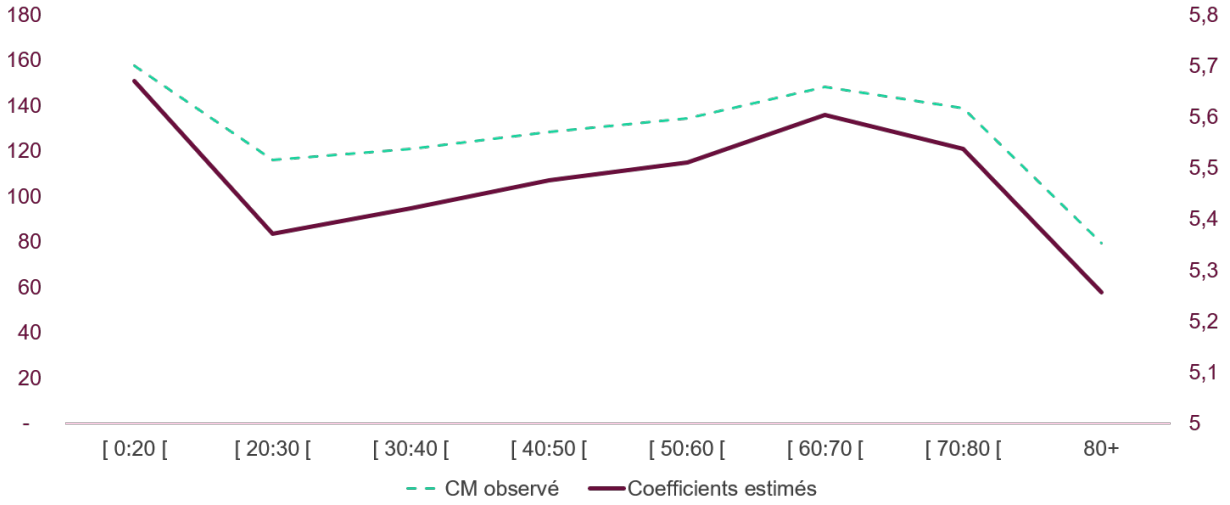


FIGURE 3.6 – Coefficients associés à l'âge

— **Variable « Sexe »**

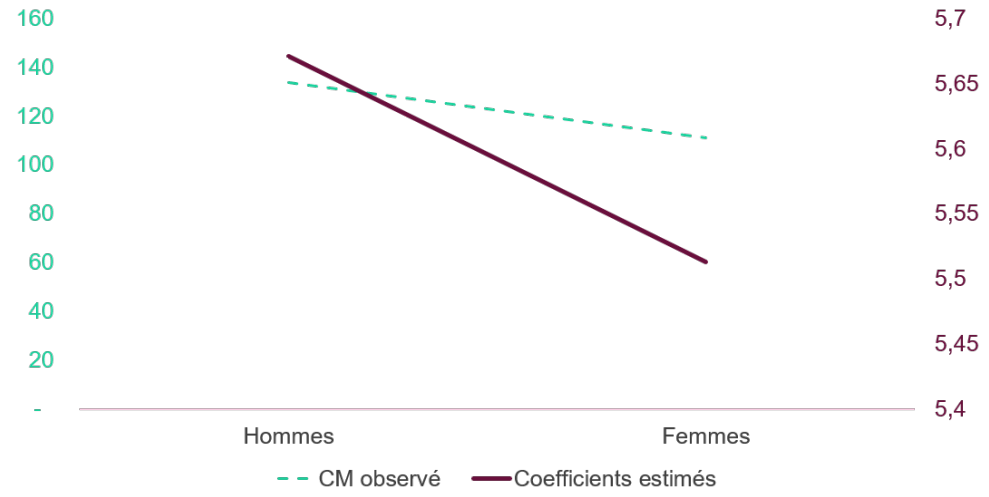


FIGURE 3.7 – Coefficients associés à la variable sexe

— **Variable « région »**

(i) **Sur les âges inférieurs à 80 ans**

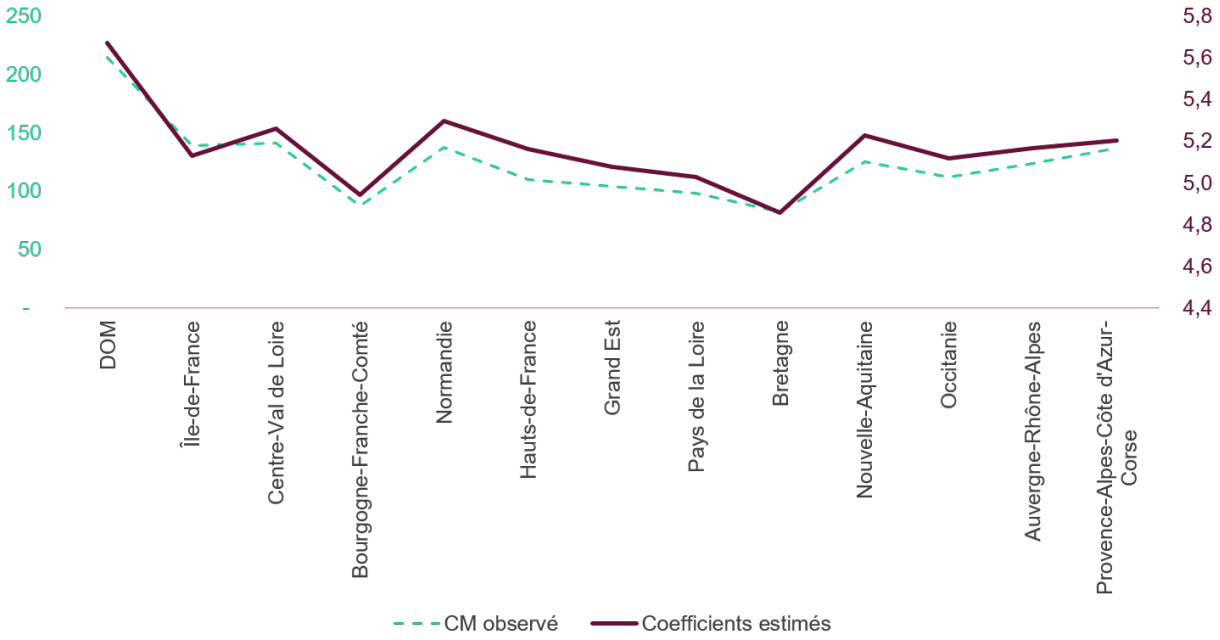


FIGURE 3.8 – Coefficients associés à la région chez les moins de 80 ans

(ii) **Sur les âges supérieurs ou égaux à 80 ans**

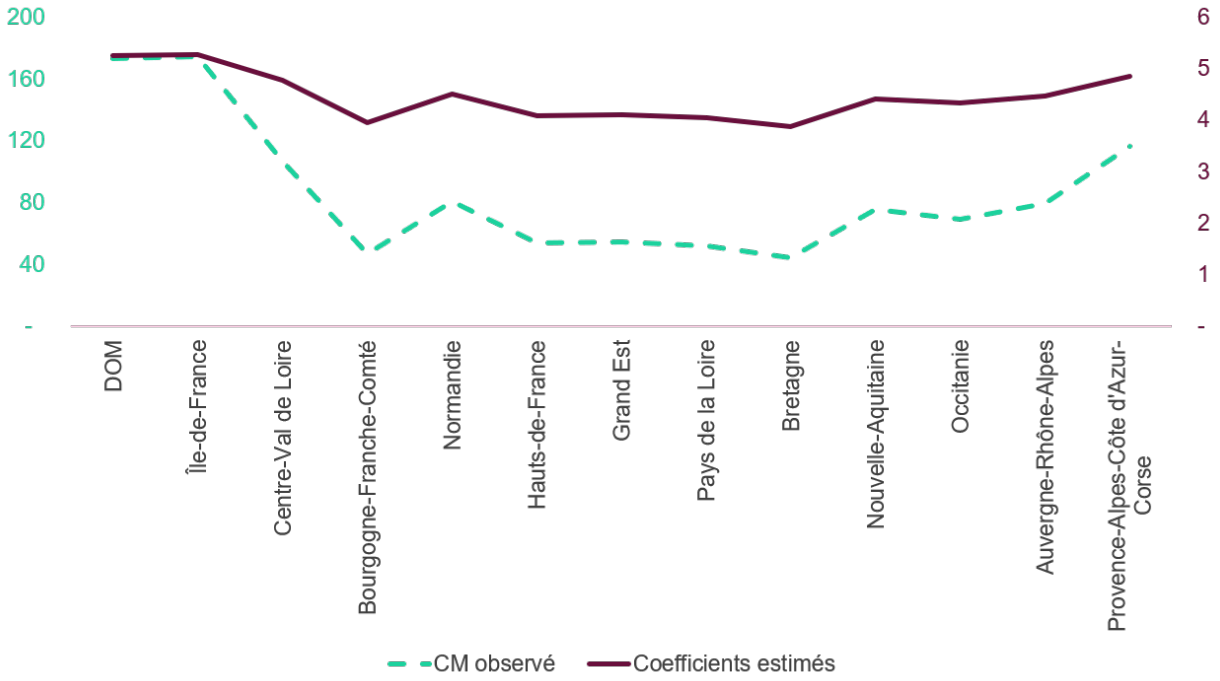


FIGURE 3.9 – Coefficients associés à la région chez les 80 ans et plus

— **Variable « CMU »**

(i) **Sur la population masculine**

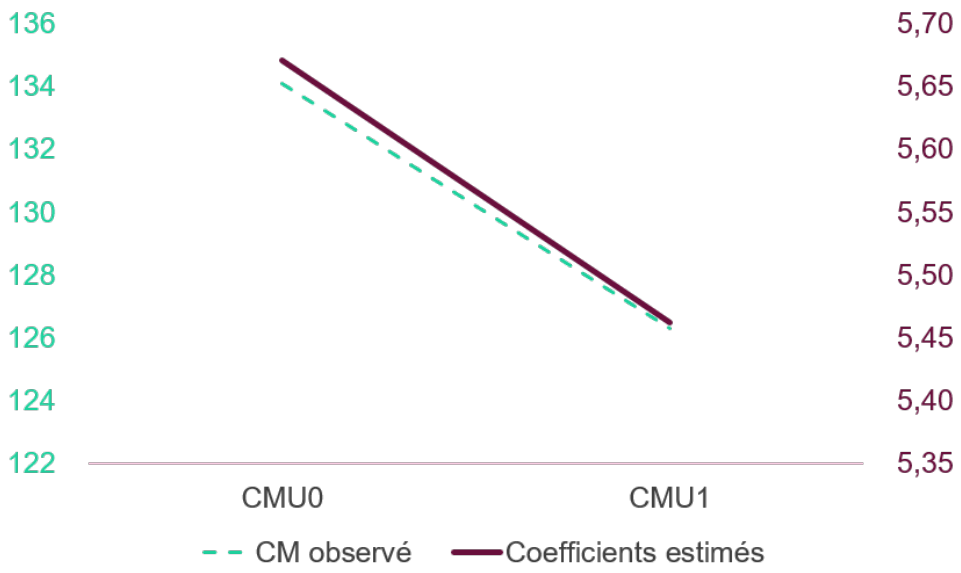


FIGURE 3.10 – Coefficients associés à la variable CMU chez les hommes

(ii) **Sur la population féminine**

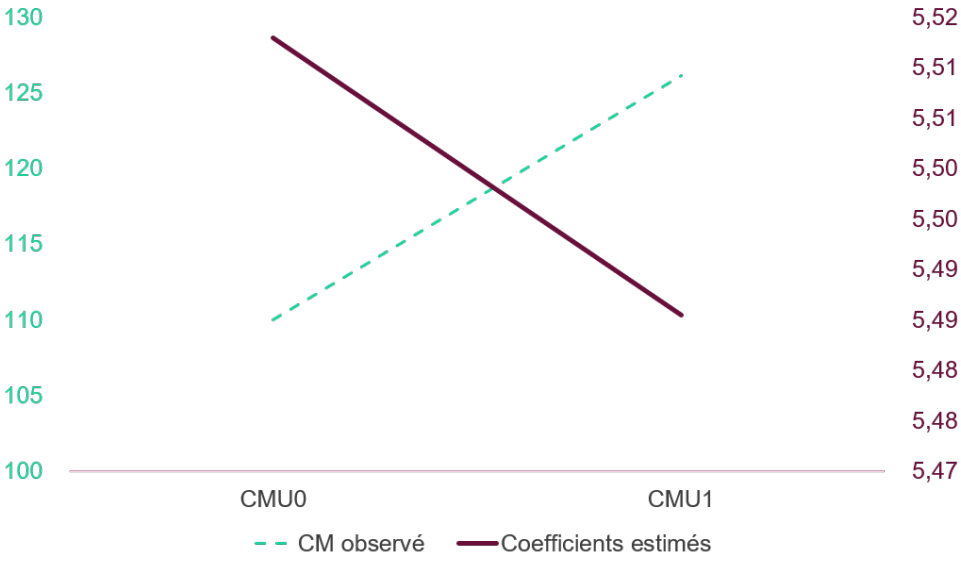


FIGURE 3.11 – Coefficients associés à la variable CMU chez les femmes

Si le modèle retranscrit relativement bien les tendances observées, l'effet conjoint du sexe et de la CMU sur le coût moyen observé n'est quant à lui pas correctement reproduit. L'interaction Sexe*CMU vient donc complexifier le modèle sans pour autant en améliorer la pertinence. Celle-ci a par conséquent été retirée du modèle final.

Enfin, il est à noter que le retrait de cette interaction ne modifie pas l'évolution des coefficients en fonction des modalités des variables tarifaires maintenues.

Variable explicative	Estimation	p-value	Significativité
(Intercept)	5,662037868	2,31E-292	***
Age20	-0,296209468	6,81E-15	***
Age30	-0,24642225	4,52E-12	***
Age40	-0,192189496	7,06E-09	***
Age50	-0,157564251	4,16E-07	***
Age60	-0,064062742	0,032882474	*
Age70	-0,130525496	1,54E-05	***
Age80	-0,411515117	0,000166168	***
Region11	-0,540188863	4,55E-27	***
Region24	-0,411677167	3,69E-12	***
Region27	-0,728731438	9,81E-34	***
Region28	-0,374663331	2,32E-11	***
Region32	-0,508266206	7,74E-22	***
Region44	-0,595875502	1,08E-27	***
Region52	-0,644127783	2,60E-29	***
Region53	-0,814470361	5,93E-41	***
Region75	-0,444263407	2,33E-17	***
Region76	-0,553546571	3,88E-27	***
Region84	-0,50402207	1,11E-22	***
Region93	-0,468468601	3,04E-20	***
CMU1	-0,111636222	4,65E-06	***
Sexe2	-0,145333291	8,00E-30	***
Age80:Region11	0,543197457	2,27E-06	***
Age80:Region24	-0,08243698	0,52103387	
Age80:Region27	-0,579700327	1,34E-06	***
Age80:Region28	-0,380396201	0,002056604	**
Age80:Region32	-0,675420678	5,35E-09	***
Age80:Region44	-0,565694859	1,08E-06	***
Age80:Region52	-0,560163213	2,31E-06	***
Age80:Region53	-0,563958142	2,41E-06	***
Age80:Region75	-0,402616475	0,000446833	***
Age80:Region76	-0,372946418	0,000802857	***
Age80:Region84	-0,28846212	0,010861761	*
Age80:Region93	0,062312919	0,580942802	

FIGURE 3.12 – Modèle final de coût moyen - tarification globale

Validation du modèle

Bien que le modèle retranscrive bien les tendances observées sur les données, il est important qu'il vérifie les hypothèses sur lesquelles il est fondé à savoir :

— **Hypothèse 1 : résidus centrés**
Bien que quelques points s'en écartent, le nuage de point des résidus est bien centré autour de l'axe des abscisses. Par conséquent, cette hypothèse est vérifiée.

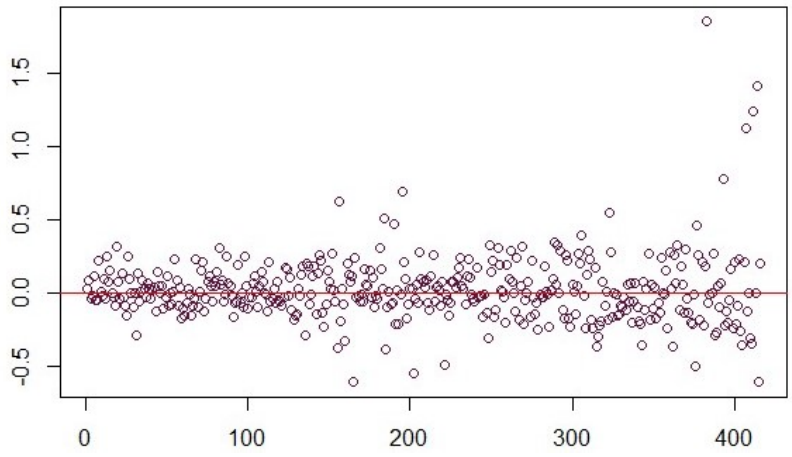


FIGURE 3.13 – Nuage de points des résidus de Pearson - modèle de coût - tarification globale

Il aurait été possible d'écarter les individus présentant des résidus erronés vis-à-vis de l'attendu pour s'assurer du respect des hypothèses. Toutefois, il en résulterait une perte importante d'information, la base de calibration étant en *model point*.

— **Hypothèse 2 : résidus homoscédastiques**
La dispersion des résidus en fonction des valeurs prédites ne présente pas de déformation particulière ce qui permet d'accepter l'hypothèse d'homoscédasticité.

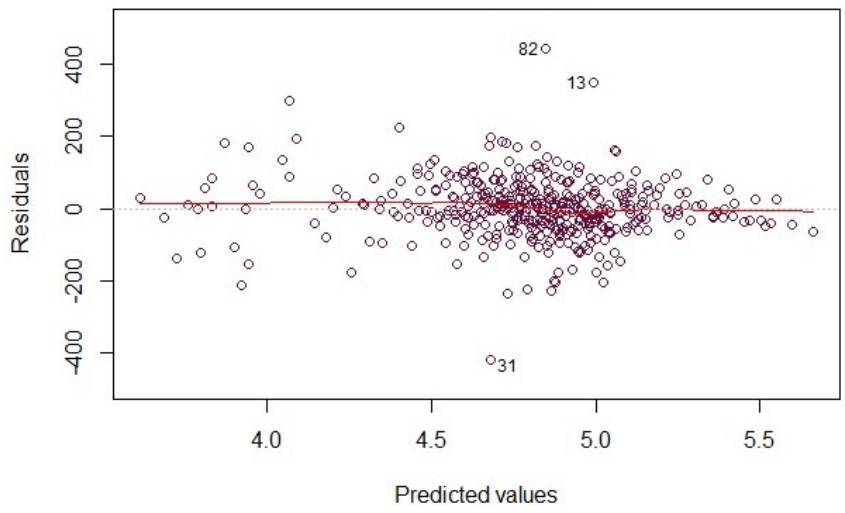


FIGURE 3.14 – Tracé des résidus en fonction des valeurs prédites - modèle de coût - tarification globale

— **Hypothèse 3 : résidus indépendants**
Avec le logiciel R, le calcul de la statistique de Durbin-Watson ne prend pas en compte les modèles avec des poids comme c'est le cas avec le modèle Gamma réalisé. Par conséquent, l'indépendance des résidus a été évalué par l'équivalent graphique du test de Durbin-Watson à savoir une ACF (fonction d'auto-correlation) de lag 1. Ainsi, le graphique suivant présente une autocorrélation proche de 0, permettant de conclure que les résidus du modèle sont indépendants.

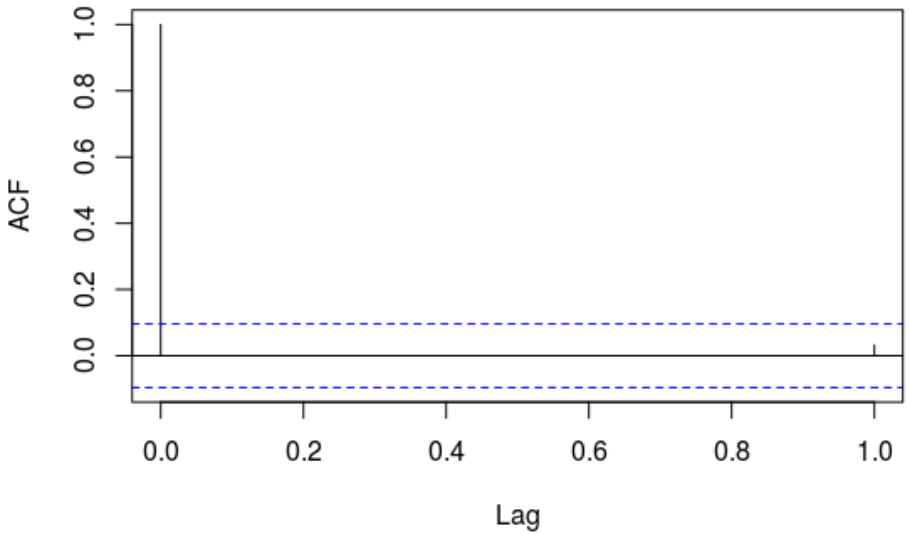


FIGURE 3.15 – ACF lag 1 - modèle de coût - tarification globale

Evaluation de la qualité prédictive du modèle

Afin de juger de la qualité du modèle, un exercice de prédiction a été réalisé sur **la base de test**. La qualité de cette prédiction a été évaluée selon plusieurs axes :

— **Calcul de la RMSE** : elle s'élève à 26,81. Mis en perspective avec le coût moyen, tous profils confondus (119,12 euros), cette RMSE traduit un biais de prédiction (RMSE /coût moyen) de 22,51%.

— **Evaluation de la dépense totale prédite toutes choses égales par ailleurs** : pour chaque profil, le coût moyen prédit a été multiplié par le nombre de sinistres réels observés, afin de déterminer la dépense totale prédite qui pourrait en résulter. Il en ressort un écart de seulement 1,85% à la dépense totale observée.

— **Représentation graphique des coûts moyens observés et des coûts moyens prédits** : les prédictions ont des niveaux relativement similaires à l'observé comme le montre le graphique ci-après.

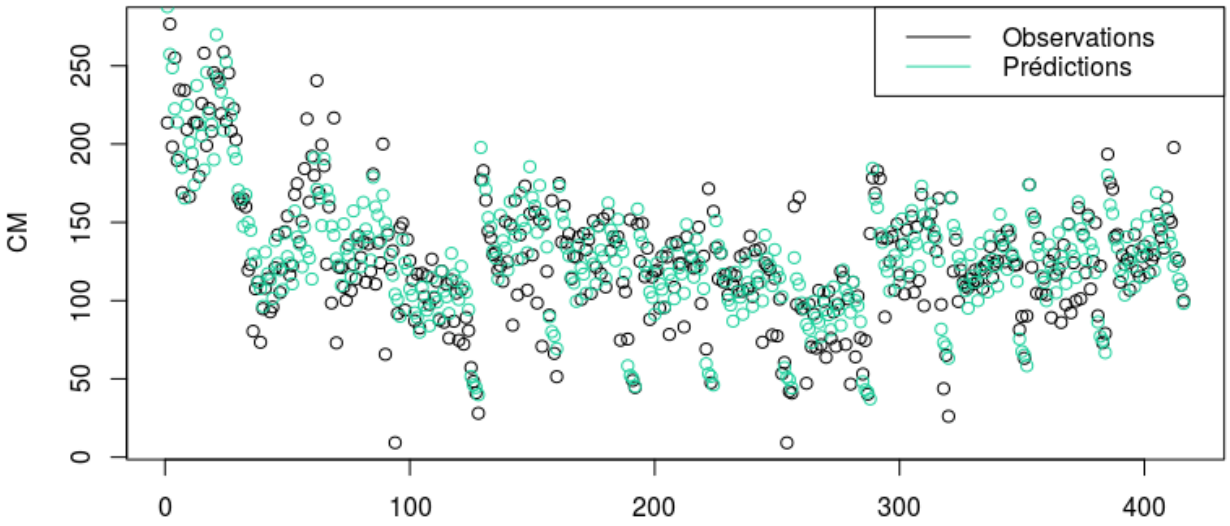


FIGURE 3.16 – Représentation coûts moyens prédits et observés - tarification globale

3.3.2 Modèle de fréquence

Distribution adéquate

De manière similaire aux travaux réalisés sur le modèle de coût, la loi de Poisson, la loi binomiale négative et la loi quasi-Poisson ont été testées comme lois potentielles de modélisation. La forte dispersion observée dans les données a conduit à écarter la loi de Poisson. Quant aux deux autres lois, elles ne s'ajustent pas non plus correctement à l'ensemble de la distribution, bien qu'ayant des formes différentes (cf. 3.17). La modélisation se fera donc simultanément selon chacune de ces lois puis il sera retenu celle avec le meilleur pouvoir prédictif.

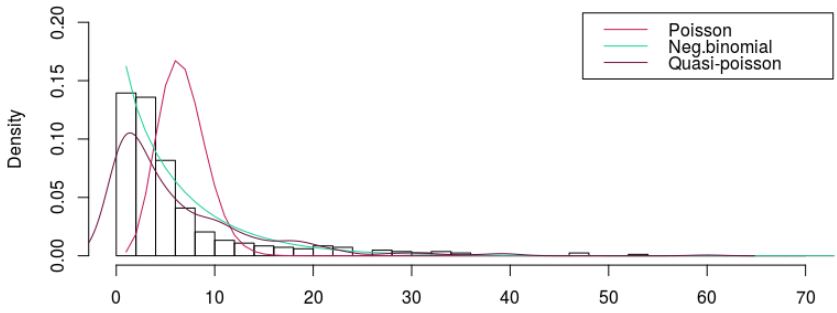


FIGURE 3.17 – Distribution fréquence - tarification globale

Sélection de variables

Dans le cas du modèle de fréquence, il apparait que le meilleur modèle est :

$$Age + Region + Sexe + CMU + Region : Age + Sexe : Age + CMU : Age + CMU : Region + CMU : Sexe$$

Sa calibration, selon la loi binomiale négative et après retraitement afin d'obtenir un modèle composé uniquement de coefficients significatifs, aboutit à retenir le modèle suivant :

Variable explicative	Estimation	p-value	Significativité
(Intercept)	5,662037868	2,31E-292	***
Age20	-0,296209468	6,81E-15	***
Age30	-0,24642225	4,52E-12	***
Age40	-0,192189496	7,06E-09	***
Age50	-0,157564251	4,16E-07	***
Age60	-0,064062742	0,032882474	*
Age70	-0,130525496	1,54E-05	***
Age80	-0,411515117	0,000166168	***
Region11	-0,540188863	4,55E-27	***
Region24	-0,411677167	3,69E-12	***
Region27	-0,728731438	9,81E-34	***
Region28	-0,374663331	2,32E-11	***
Region32	-0,508266206	7,74E-22	***
Region44	-0,595875502	1,08E-27	***
Region52	-0,644127783	2,60E-29	***
Region53	-0,814470361	5,93E-41	***
Region75	-0,444263407	2,33E-17	***
Region76	-0,553546571	3,88E-27	***
Region84	-0,50402207	1,11E-22	***
Region93	-0,468468601	3,04E-20	***

Variable explicative	Estimation	p-value	Significativité
CMU1	-0,111636222	4,65E-06	***
Sexe2	-0,145333291	8,00E-30	***
Age80:Region11	0,543197457	2,27E-06	***
Age80:Region24	-0,08243698	0,52103387	
Age80:Region27	-0,579700327	1,34E-06	***
Age80:Region28	-0,380396201	0,002056604	**
Age80:Region32	-0,675420678	5,35E-09	***
Age80:Region44	-0,565694859	1,08E-06	***
Age80:Region52	-0,560163213	2,31E-06	***
Age80:Region53	-0,563958142	2,41E-06	***
Age80:Region75	-0,402616475	0,000446833	***
Age80:Region76	-0,372946418	0,000802857	***
Age80:Region84	-0,28846212	0,010861761	*
Age80:Region93	0,062312919	0,580942802	

FIGURE 3.18 – Coefficients modèle binomial négatif - tarification globale

Quant à la loi quasi-Poisson, le modèle retenu est présenté ci-après :

Variable explicative	Estimation	p-value	Significativité
(Intercept)	-0,692285	5,38E-32	***
Age20	0,7398831	6,86E-31	***
Age30	0,9431701	5,04E-50	***
Age40	1,2540324	4,65E-82	***
Age50	1,7806577	6,03E-142	***
Age60	2,1938006	8,38E-166	***
Age70	2,7815156	5,07E-201	***
Age80	3,5366103	5,61E-231	***
Sexe2	-0,071819	5,13E-06	***
CMU1	0,4264802	6,04E-11	***
Region11	0,8785453	6,76E-31	***
Region24	-0,043731	3,79E-01	
Region27	0,2913494	7,69E-10	***
Region28	0,6405875	7,54E-08	***
Region32	0,2280188	1,89E-07	***
Region44	0,1232742	4,86E-03	**
Region52	0,2113544	3,90E-06	***
Region53	0,1849039	6,60E-05	***
Region75	0,0023326	9,57E-01	
Region76	0,7064194	8,98E-40	***
Region84	0,511502	1,44E-08	***
Region93	0,6859633	4,81E-37	***
Age20:Sexe2	0,2592896	1,15E-06	***
Age30:Sexe2	0,1626155	5,73E-04	***
Age40:Sexe2	0,1508399	2,53E-04	***
Age80:Sexe2	0,1099592	6,87E-05	***
Age20:CMU1	0,0838684	3,73E-01	
Age30:CMU1	0,3022626	5,40E-04	***
Age40:CMU1	0,4745732	1,93E-08	***
Age50:CMU1	0,1962839	1,43E-02	*

Variable explicative	Estimation	p-value	Significativité
Age60:CMU1	1,261747	2,06E-36	***
Age70:CMU1	-0,756555	4,30E-07	***
Age80:CMU1	-2,419048	9,62E-38	***
Age20:Region11	-0,23439	4,36E-03	**
Age30:Region11	-0,297346	1,55E-04	***
Age40:Region11	-0,427165	2,11E-08	***
Age50:Region11	-0,562008	3,43E-14	***
Age60:Region11	-0,698456	6,57E-20	***
Age70:Region11	-0,887782	5,93E-29	***
Age80:Region11	-1,288417	2,48E-52	***
Age20:Region28	-0,409139	1,72E-02	*
Age30:Region28	-0,453746	4,73E-03	**
Age40:Region28	-0,501995	7,52E-04	***
Age50:Region28	-0,52029	1,43E-04	***
Age60:Region28	-0,612231	3,81E-06	***
Age70:Region28	-0,671783	3,28E-07	***
Age80:Region28	-0,844923	3,41E-11	***
Age20:Region84	-0,14526	2,02E-01	
Age30:Region84	-0,245098	2,42E-02	*
Age40:Region84	-0,243701	1,61E-02	*
Age50:Region84	-0,33285	5,21E-04	***
Age60:Region84	-0,420785	9,27E-06	***
Age70:Region84	-0,463057	7,04E-07	***
Age80:Region84	-0,651136	1,29E-12	***
Age60:Region93	-0,194208	5,22E-04	***
Age70:Region93	-0,422156	6,66E-14	***
Age80:Region93	-0,604213	1,13E-29	***
Age60:Region76	-0,178782	7,58E-04	***
Age70:Region76	-0,272321	1,26E-07	***
Age80:Region76	-0,231339	1,23E-07	***

FIGURE 3.19 – Coefficients modèle quasi-Poisson - tarification globale

Afin de choisir la loi qui convient le mieux, la RMSE et la dépense totale prédite de chaque modèle a été calculée.

	Quasi-Poisson	Binomiale négative
RMSE	16 957	70 267
Nombre moyen de sinistres (pondéré par l'exposition)	196 159	196 159
RMSE / nombre moyen de sinistres	8,64%	35,82%
Nombre total de sinistres observés	43 747 762	43 747 762
Nombre total de sinistres prédits	45 537 080	48 780 995
Ecart	4%	12%

FIGURE 3.20 – Modèle de fréquence quasi-Poisson VS binomiale négative - tarification globale

Il en ressort que la loi quasi-Poisson est la plus performante. C'est donc le modèle découlant de cette loi qui sera analysé dans la suite de cette partie.

Analyse des coefficients

Tout comme le modèle de coût moyen, la segmentation base d'apprentissage/ base de test a un impact négligeable sur les coefficients du modèle de fréquence comme le montre le graphique ci-dessous :

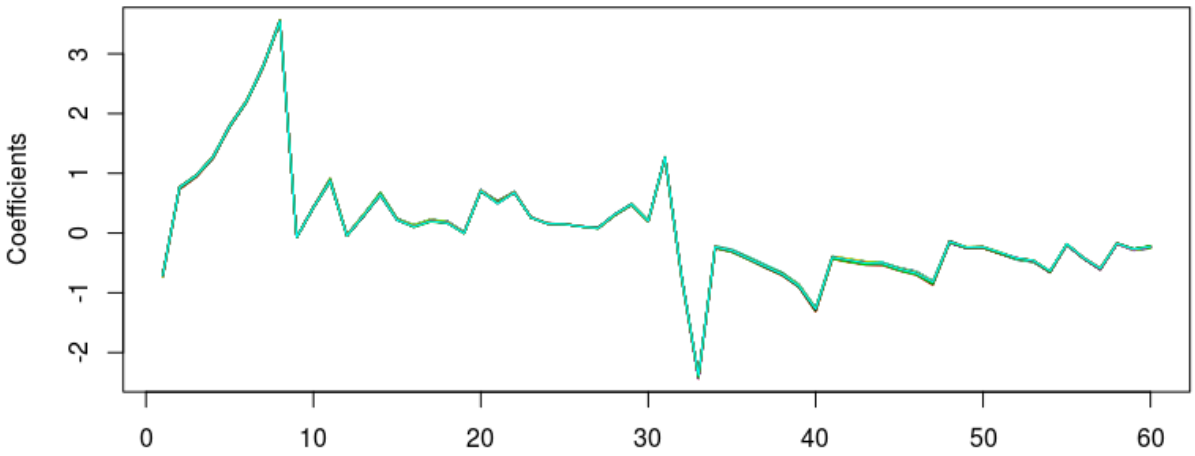


FIGURE 3.21 – Stabilité des coefficients - modèle de fréquence - tarification globale

Mis en perspective avec les tendances empiriques, ces coefficients retranscrivent globalement bien le risque comme il est détaillé ci-après :

— **Variable « Sexe »**
Avec ce modèle, les hommes ont bien une fréquence de recours aux soins hospitaliers supérieure à celle des femmes.

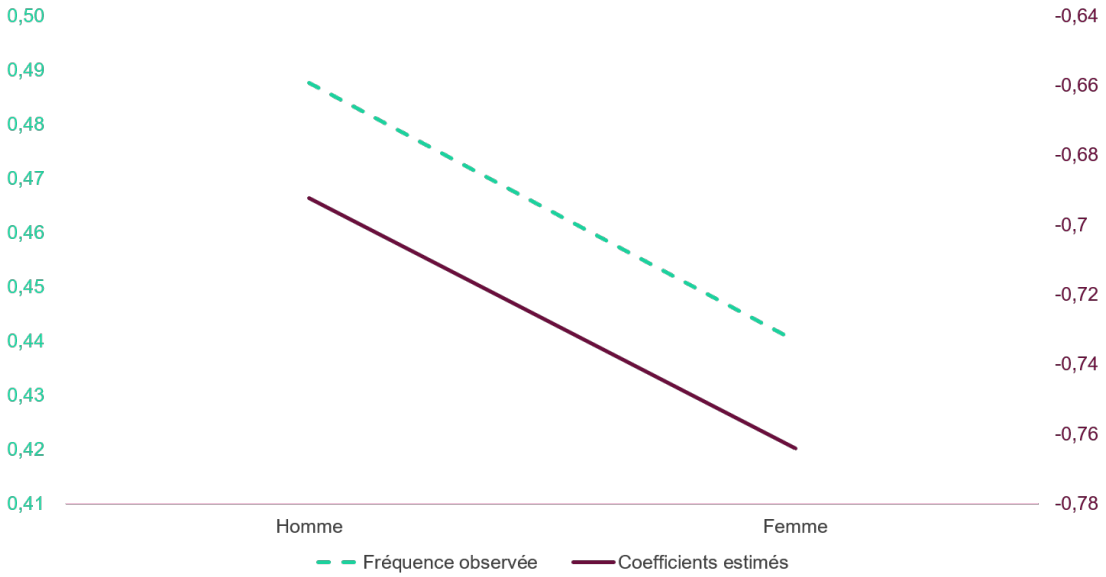


FIGURE 3.22 – Coefficients associés au sexe - modèle de fréquence - tarification globale

— **Variable « âge » et « CMU »**
Sur la fréquence de recours aux soins hospitaliers, les variables « âge » et « CMU » sont liées car des écarts importants de fréquence sont observées par âge selon la population considérée. L'analyse des coefficients inhérents à ces variables montre que le modèle retranscrit bien les tendances d'évolution observées, notamment le pic de fréquence observé sur la tranche d'âge 60-69 chez la population à la CMU.

(i) **Population hors CMU**

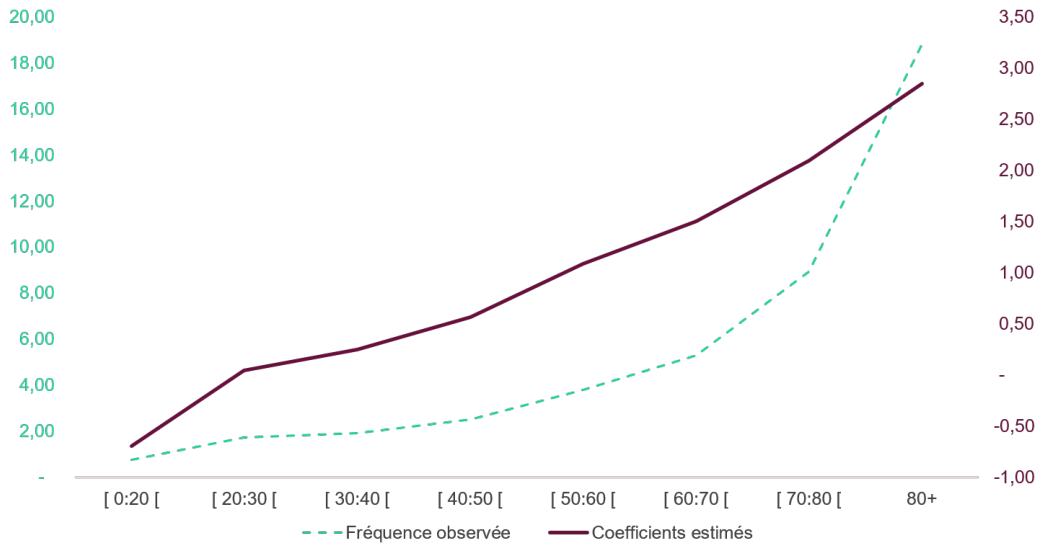


FIGURE 3.23 – Coefficients associés à l'âge chez la population hors CMU - modèle de fréquence - tarification globale

(ii) **Population CMU**

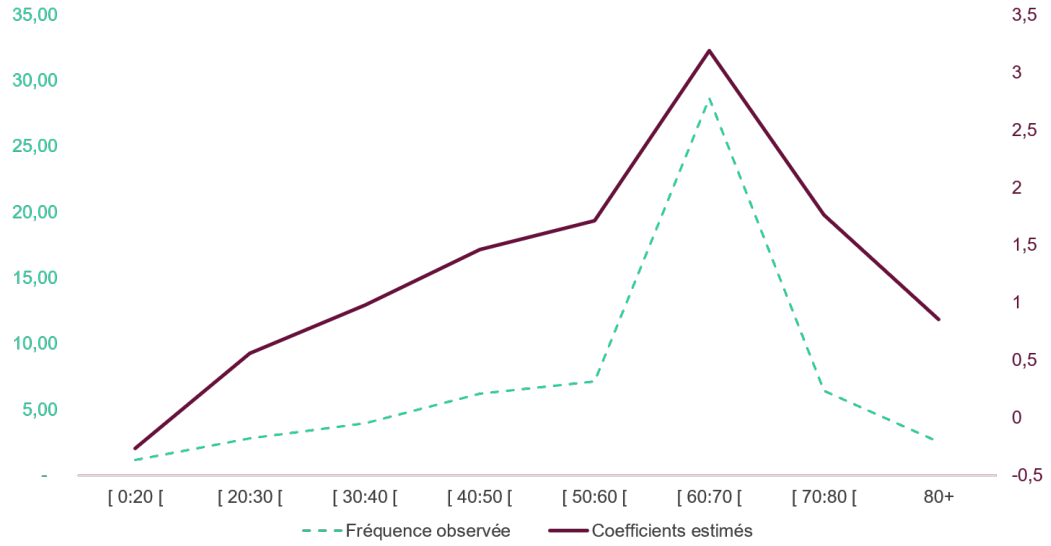


FIGURE 3.24 – Coefficients associés à l'âge chez la population CMU - modèle de fréquence - tarification globale

— **Variable « région »**
A l'exception de certaines régions, le modèle respecte bien les écarts de niveaux de fréquence entre les régions comme le montre le graphique ci-après.

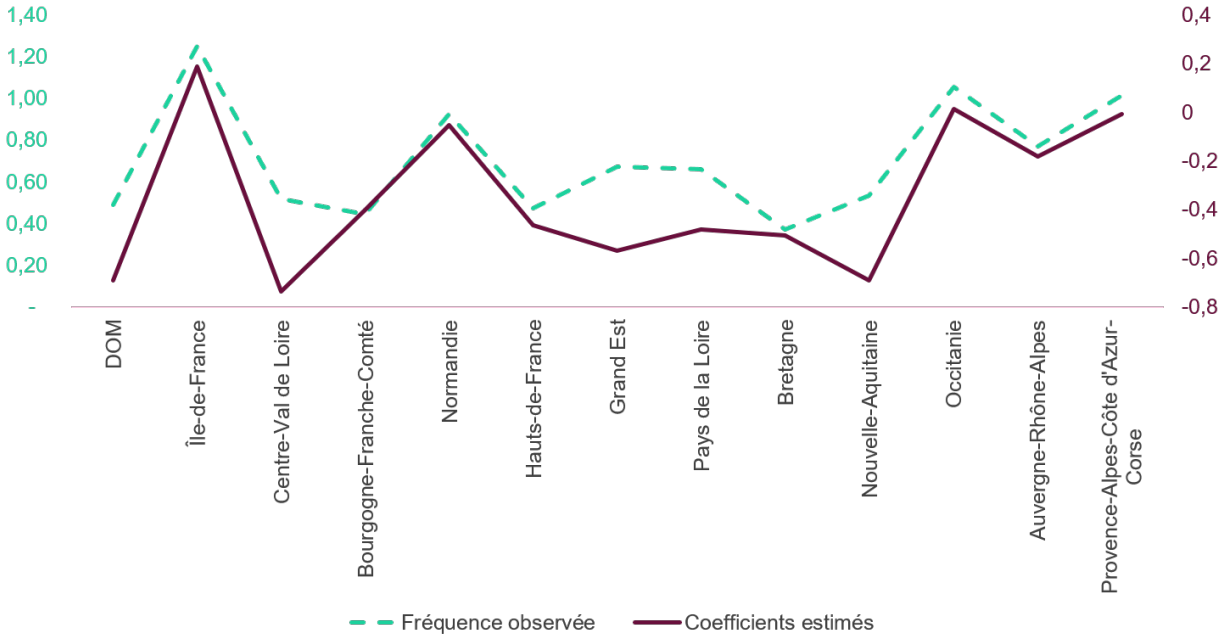


FIGURE 3.25 – Coefficients associés à la région - modèle de fréquence - tarification globale

Validation du modèle

Comme le montrent les graphes ci-dessous, les résidus de ce modèle sont centrés (autour de la droite y=0) et homoscédastiques (pas de déformation de la distribution des résidus). Toutefois, les résidus ne sont pas indépendants car la statistique de test de Durbin Watson (DW) calculée est proche de 4 : DW = 3.1628.

— **Résidus centrés** :

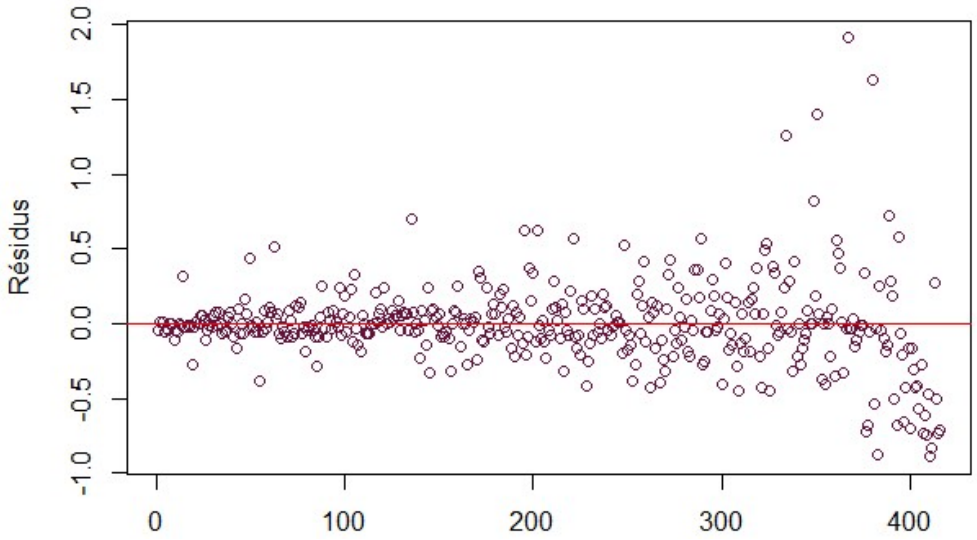


FIGURE 3.26 – Nuage de points des résidus de Pearson - modèle de coût - tarification globale

— **Résidus homoscédastiques** :

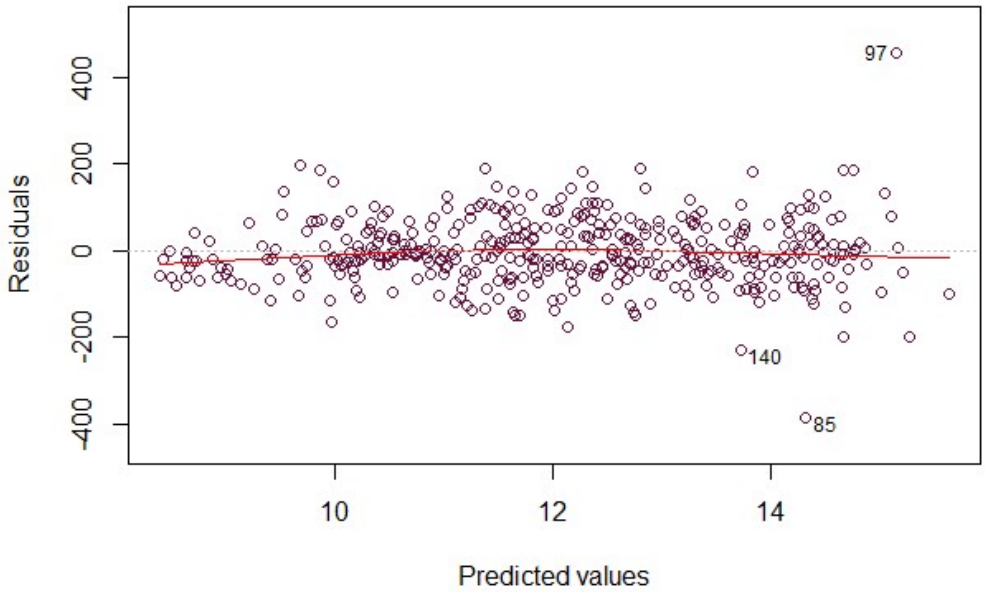


FIGURE 3.27 – Tracé des résidus en fonction des valeurs prédites - modèle de coût - tarification globale

3.3.3 Prime pure et performance globale du modèle

Suite au calibrage des modèles de coût et de fréquence, la dépense totale ainsi que la prime pure moyenne ont été évaluées sur la base de test. Les résultats sont résumés dans le tableau ci-après :

Dépense totale observée	5 211 083 140
Dépense totale prédite	5 538 699 125
Ecart	6,287%

Prime pure moyenne observée	466,82
Prime pure moyenne prédite	496,17
Ecart	6,287%

FIGURE 3.28 – Résultat global - modèle de coût-fréquence - tarification globale

3.4 Tarification par sous-postes de soins usuels

Cette partie présente la deuxième tarification, réalisée en ligne avec les pratiques actuelles de classification des actes sur la garantie hospitalisation. Il s'agit d'un regroupement des actes hospitaliers par nature de prestations médicales afin de commercialiser des packages de garantie selon une cohérence médicale.

3.4.1 Classification des actes en sous-postes de soins usuels

Usuellement, trois sous-postes de soins sont considérés en hospitalisation :

— Les frais de séjour (FSEJ) : actes de chirurgie, d'anesthésie, forfait journalier, honoraires du personnel médical... ;

— Les frais de chambre (CHBR) : supplément chambres particulières, diverses demandes de conforts... ;

— Les frais divers (AUTR) : consultations spécifiques et majorations diverses.

Les actes de la base d'étude ont donc été regroupés selon ces sous-postes de soins en effectuant un croisement avec la base sinistres d'un assureur de place. Les actes absents de la base assureur ont quant à eux été classés à dire d'expert. Le bilan de ces traitements est résumé dans le tableau ci-après :

	Nombre d'actes	%	Nombre de sinistres	%	Coût sinistres	%
Actes mappés	120	*43,80%*	219 103 752	*80,72%*	30 456 914 970	*92,68%*
Actes classés à dire d'experts	154	*56,20%*	52 329 409	*19,28%*	2 403 919 682	*7,32%*
Total	**274**		**271 433 161**		**32 860 834 652**	

FIGURE 3.29 – Bilan ventilation des actes en sous-postes de soins

Ainsi, la majorité des actes hospitaliers sollicités concerne le sous-poste de soins FSEJ qui représente alors 98,78% du coût des sinistres et 86,4% du volume de sinistres.

Sous-postes	Nombre d'actes	%	Nombre de sinistres	%	Coût sinistres	%
FSEJ	239	*87,23%*	234 505 031	*86,40%*	32 459 483 209	*98,78%*
CHBR	3	*1,09%*	1 558 493	*0,57%*	55 640 598	*0,17%*
AUTR	32	*11,68%*	35 369 637	*13,03%*	345 710 845	*1,05%*
Total	**274**		**271 433 161**		**32 860 834 652**	

FIGURE 3.30 – Bilan de la consommation des actes par sous-postes de soins

3.4.2 Analyse descriptive des sous-postes de soins

L'analyse des coûts moyens et de la fréquence de consommations des actes par sous-poste de soins montre de réels écarts de comportements entre ces différents groupes d'actes.

De manière globale, les frais de séjour ont un coût moyen 14 fois plus élevé que les soins classés en "autres soins" et 4 fois plus élevés que les frais de chambre. Au niveau de la fréquence, ces écarts sont respectivement de 7 fois et de 150 fois.

Sous-poste	Coût sinistres	Nombre de sinistres	Coût moyen	Exposition	Fréquence
FSEJ	32 459 483 209	234 505 031	138	66 977 703	3,501
AUTR	345 710 845	35 369 637	10	66 977 703	0,528
CHBR	55 640 598	1 558 493	36	66 977 703	0,023
Total	32 860 834 652	271 433 161			

FIGURE 3.31 – Coût moyen et fréquence de consommation par sous-postes de soins

Vision selon le sexe : tous les segments présentent un coût moyen plus élevé chez la population masculine que chez les femmes. Toutefois l'écart est réduit sur les frais de chambre par rapport aux deux autres segments. Quant à la fréquence, elle est plus élevée chez les femmes peu importe le segment considéré avec un écart plus prononcé sur le segment "autres".

		CHBR	FSEJ	AUTR
Coût moyen	Hommes	36	151	10
	Femmes	35	128	9
Fréquence	Hommes	0,02	3,27	0,46
	Femmes	0,03	3,72	0,59

FIGURE 3.32 – Coût moyen et fréquence de consommation par sexe et sous-postes de soins

Vision selon l'âge : les évolutions du coût moyen par âge sont très différentes selon les postes de soins. Toutefois sur la fréquence, chaque sous-poste présente une évolution croissante avec l'âge.

FIGURE 3.33 – Coût moyen et fréquence de consommation par âge et sous-poste de soins

Lecture : sur le graphique de coût moyen, les coût moyen des FSEJ sont à lire sur l'axe principal et les autres sur l'axe secondaire. Sur celui de la fréquence, seul CHBR est à lire sur l'axe secondaire

Vision selon la région : la ventilation du coût moyen et de la fréquence par région montrent des tendances quasi similaires entre les frais de séjour et les frais de chambre. Toutefois, sur le poste "autres", les comportements observés sont particulièrement différents.

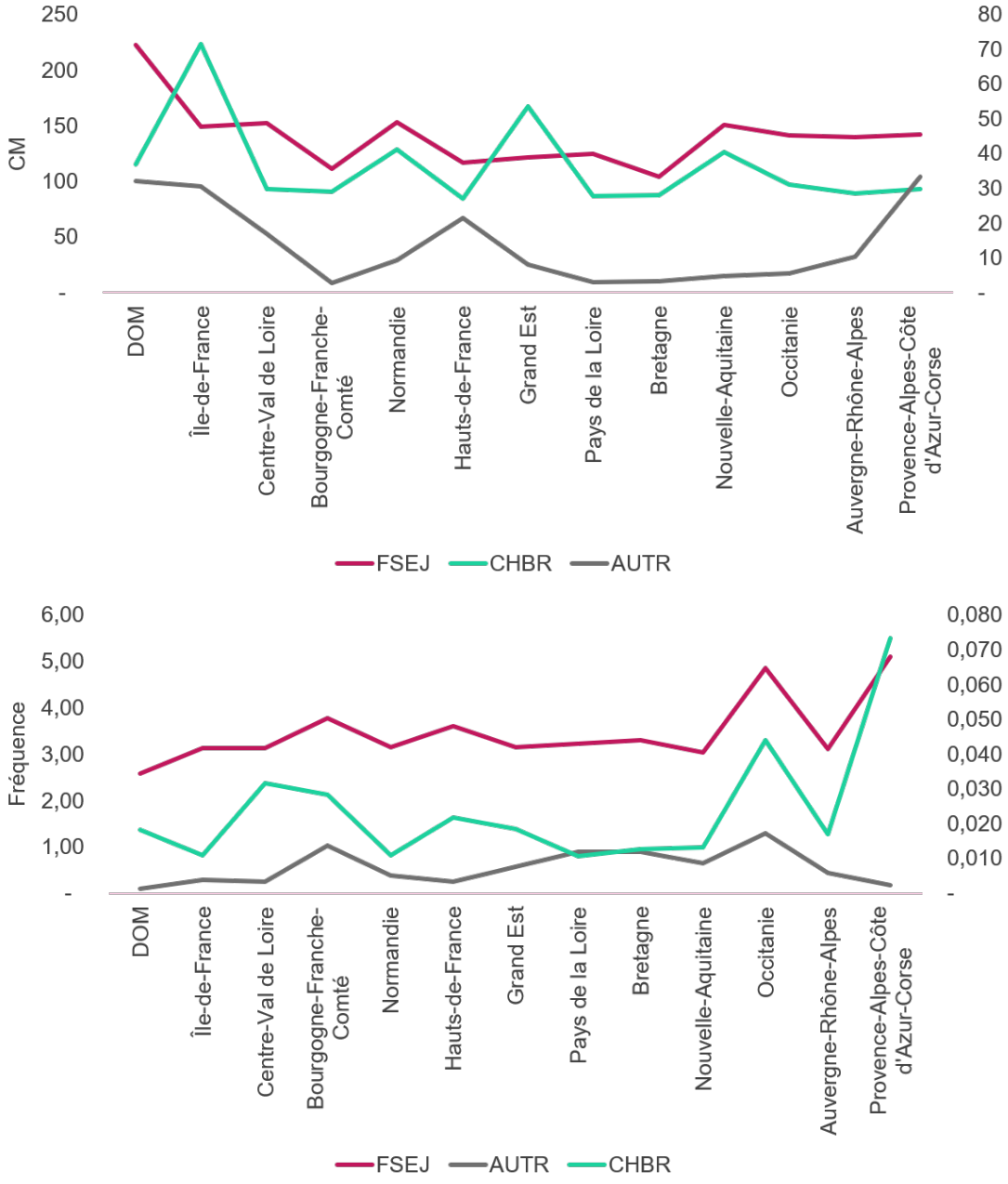


FIGURE 3.34 – Coût moyen et fréquence de consommation par région et sous-poste de soins

Lecture : sur le graphique de coût moyen, les coûts moyens des FSEJ sont à lire sur l'axe principal et les autres sur l'axe secondaire. Sur celui de la fréquence, seul CHBR est à lire sur l'axe secondaire

Vision selon l'affiliation à la CMU : les frais de séjour présentent des coûts moyens similaires sur les deux populations d'assurés. Toutefois, sur les deux autres segments la population non affiliée à la CMU a des coûts moyens supérieurs à la population CMU. En termes de fréquence, les niveaux sont identiques sur les frais de séjour, plus élevés pour la population CMU sur CHBR et plus élevés pour la population non CMU sur les actes AUTR.

		FSEJ	CHBR	AUTR
Coût moyen	Non CMU	138	37	9
	CMU	137	28	25
Fréquence	Non CMU	3,5	0,02	0,55
	CMU	3,2	0,04	0,32

FIGURE 3.35 – Coût moyen et fréquence de consommation par sous-poste de soins selon l'affiliation à la CMU

3.4.3 Modèle de coût

Sélection des distributions adéquates

Comme pour la tarification globale, une recherche des lois de modélisation adéquates a été réalisée pour chaque sous-poste. Les distributions empiriques ainsi que celles des lois théoriques candidates ajustées aux données sont présentées ci-après :

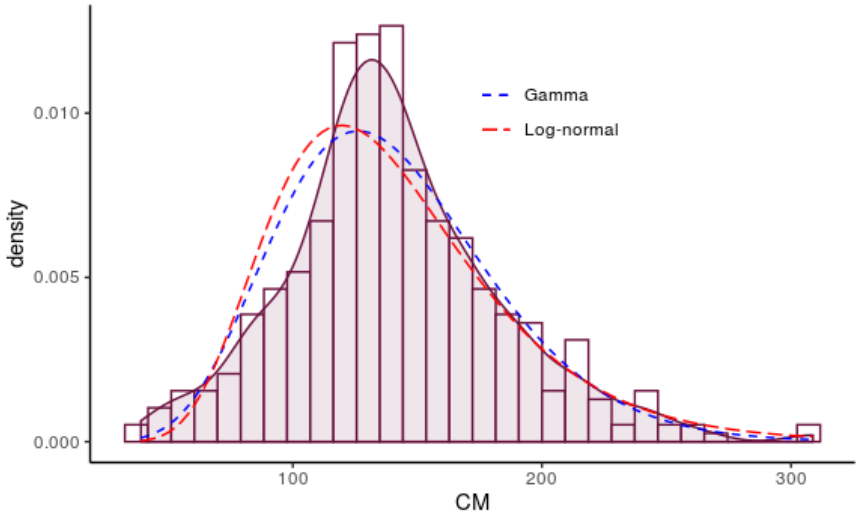


FIGURE 3.36 – Distribution CM - FSEJ

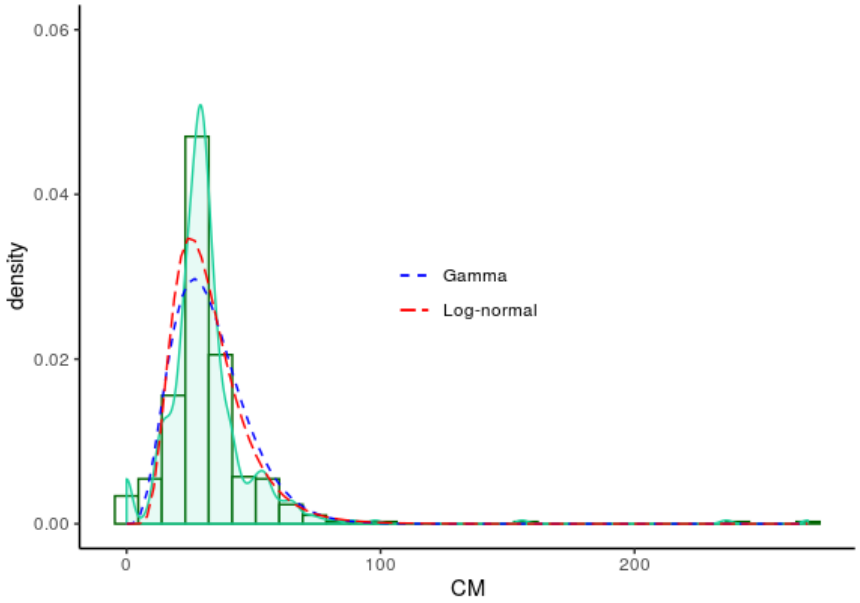


FIGURE 3.37 – Distribution CM - CHBR

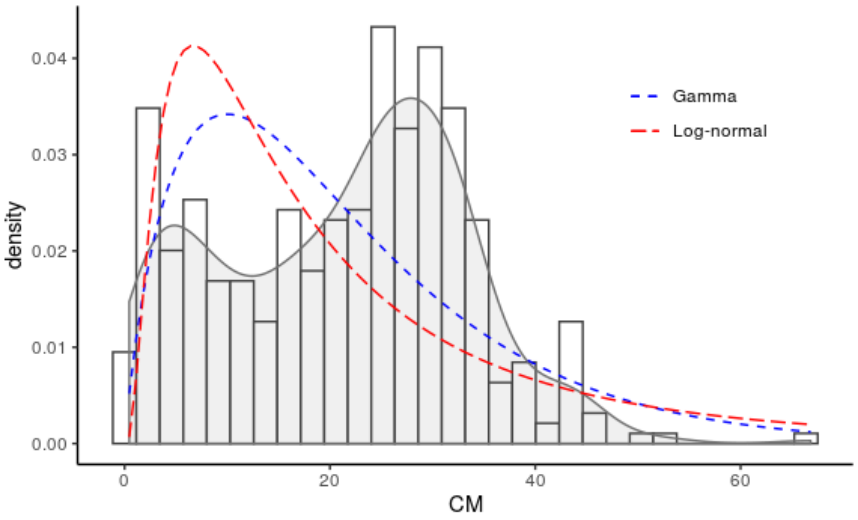


FIGURE 3.38 – Distribution CM - AUTR

Il ressort de ces graphiques que les lois usuelles des GLM ne sont pas les plus adaptées à cette modélisation. Toutefois, ces modèles ont été retenus afin d'être en ligne avec les pratiques de marché. Ainsi, l'analyse des QQ-plot et des PP-plot associés aux différentes lois a conduit aux choix présentés ci-après :

Sous-poste de soins	Distribution retenue
FSEJ	Gamma
CHBR	Log-normale
AUTR	Gamma

FIGURE 3.39 – Choix de distribution par sous-poste de soins - modèle de coût

Le lecteur intéressé, peut retrouver l'ensemble des tests d'adéquation graphiques réalisés en annexe B.

Sélection de variables

Un procédé de sélection de variables similaire (utilisation de la fonction *glmulti* et sélection des coefficients significatifs) à celui utilisé dans la tarification globale a été appliqué à chaque sous-poste dans le but d'obtenir des modèles cohérents avec les tendances empiriques observées. Les résultats sont les suivants :

Sous-poste de soins	Variables retenues
FSEJ	Age, Region, Sexe, CMU , Age80 :Region
CHBR	Age, Region, Sexe, CMU , (Region11, Region44,Region76,Region84,Region93):Age
AUTR	Age, Region, Sexe, CMU , (Region27, Region28,Region 28, Region44, Region52, Region53, Region75, Region76, Region84):(Age40,Age50,Age60)

FIGURE 3.40 – Choix des variables par sous-poste de soins - modèle de coût

Analyse des coefficients

Globalement, les modèles de coût réalisés sur chaque sous-poste de soins retranscrivent bien les tendances observées. Ci-après une vision univariée de l'adéquation des coefficients obtenus aux évolutions observées.

— **Sous-poste frais de séjour (FSEJ) :**

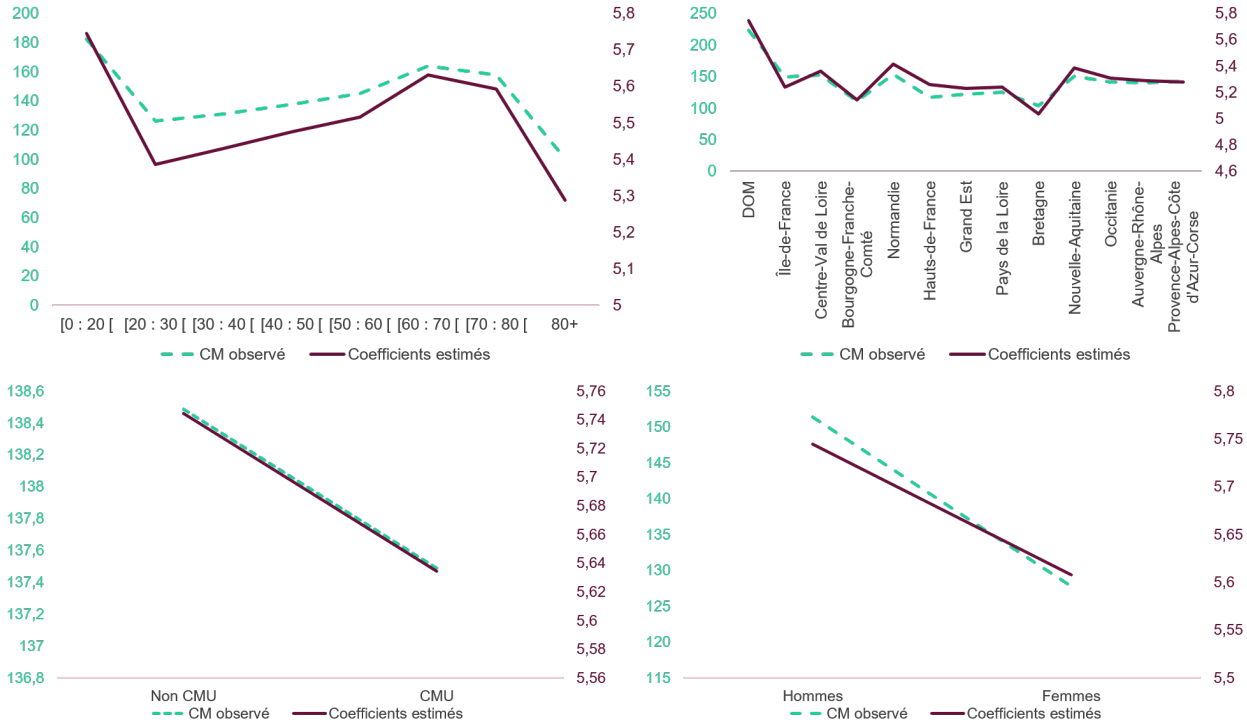


FIGURE 3.41 – Coefficients tarifaires FSEJ - modèle de coût

— **Sous-poste chambres (CHBR) :**

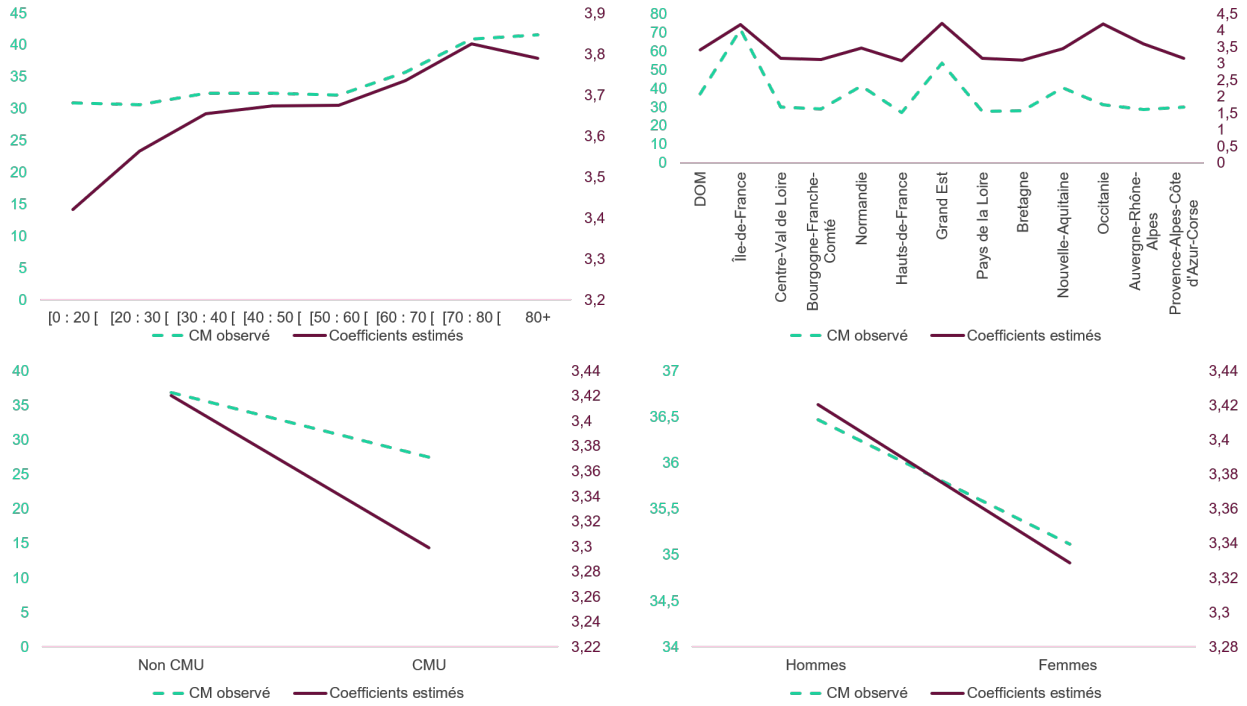


FIGURE 3.42 – Coefficients tarifaires CHBR - modèle de coût

— **Sous-poste autres (AUTR) :**

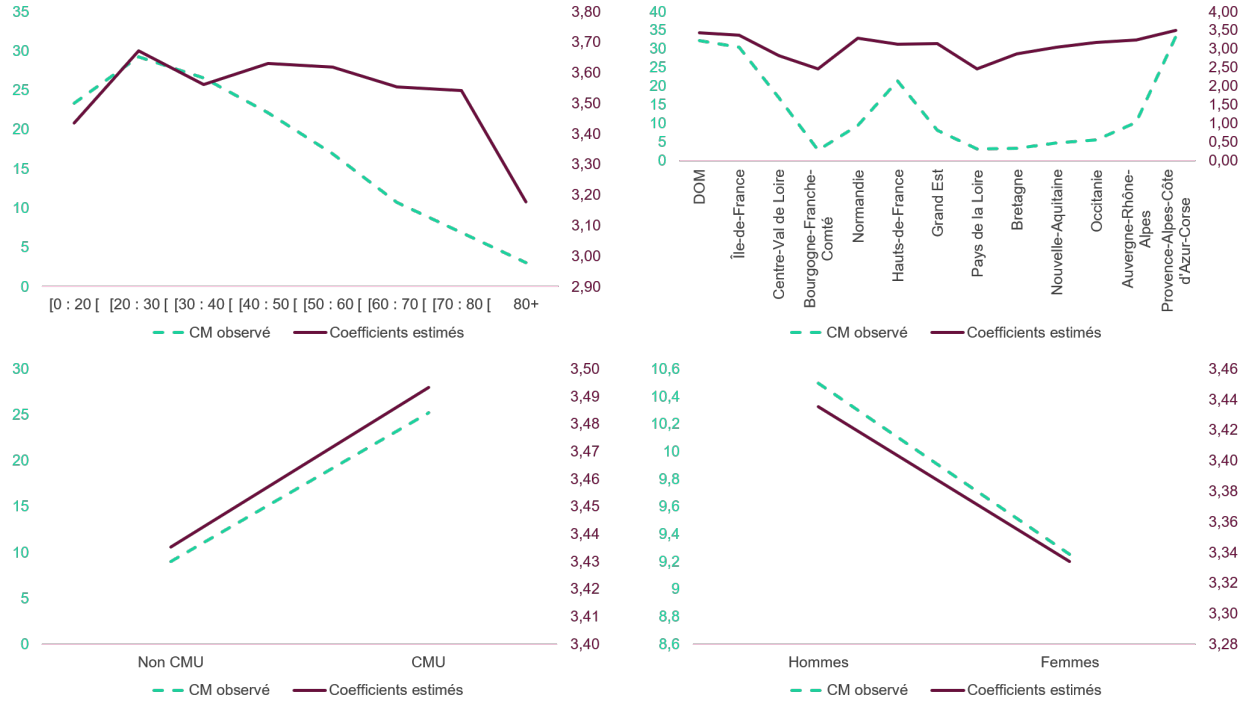


FIGURE 3.43 – Coefficients tarifaires AUTR - modèle de coût

Validation du modèle

L'analyse des résidus de chaque modèle conduit aux résultats suivants :

Sous-poste de soins	Moyenne nulle	Homoscédasticité	Indépendance
FSEJ	OK	OK	OK
CHBR	OK	OK	OK
AUTR	OK	OK	Non OK

FIGURE 3.44 – Bilan analyse des résidus - modèles de coût

Contrairement à la tarification globale, la segmentation des actes en sous-postes de soins permet d'obtenir l'indépendance des résidus sur deux des trois modèles.

Qualité prédictive des modèles et comparaison au modèle de coût global

Si les frais de séjour (FSEJ) présentent un biais de prédiction (RMSE/coût moyen) du même ordre que celui de la tarification 1, sur les segments chambres (CHBR) et autres (AUTR), les prédictions présentent des écarts importants à l'observé.

Sous-poste	Poids dans la dépense engagée	Coût moyen	RMSE	RMSE / Coût moyen
FSEJ	98,74%	136,77	28,93	21%
CHBR	0,18%	35,70	18,86	53%
AUTR	1,08%	9,59	9,53	99%

FIGURE 3.45 – Qualité des modèles de coût - tarification par sous-poste de soins

La pondération des RMSE obtenues par sous-postes de soins par leur poids dans la sinistralité totale conduit à une RMSE globale de 28,7, soit une hausse de 2% du biais de prédiction par rapport à la tarification 1.

3.4.4 Modèle de fréquence

Sélection des distributions adéquates

Pour chaque sous-poste, les distributions quasi-Poisson et binomiale négative ont été testées pour retenir la distribution permettant de minimiser l'erreur de prédiction pour chaque modèle de fréquence. Les résultats sont les suivants :

Sous-poste de soins	Distribution retenue
FSEJ	Quasi-poisson
CHBR	Quasi-poisson
AUTR	Quasi-poisson

FIGURE 3.46 – Choix de distribution par sous-poste de soins - modèle de fréquence

Sélection de variables

Les variables considérées pour chaque segment sont résumées ci-après :

Sous-poste de soins	Variables retenues
FSEJ	Age, Region, Sexe, CMU, (Region11, Region28, Region84):Age+(Region76, Region93):(Age50, Age60,Age70,Age80), (Region75,Region24):(Age70,Age80), Age:Sexe, Age:CMU
CHBR	Age, Region, Sexe, CMU, (Region11, Region24, Region27, Region44, Region52, Region75, Region76, Region84, Region93):Age, (Age60,Age70,Age80):CMU, Sexe:CMU
AUTR	Age, Region, Sexe, CMU, (Region11, Region93): (Age50, Age60, Age70, Age80), CMU: (Age50, Age70, Age80)

FIGURE 3.47 – Choix de variable par sous-poste de soins - modèle de fréquence

Analyse des coefficients

L'âge et la région sont les variables qui impactent le plus la fréquence de consommation des soins hospitaliers. Sur ces variables, les coefficients estimés par sous-poste de soins sont cohérents (à l'exception de quelques profils) avec les tendances observées. A contrario, sur les variables sexe et CMU, les modèles ne parviennent pas toujours à reproduire correctement les tendances univariées observées (notamment sur les FSEJ et les actes AUTR). Toutefois, les interactions introduites dans le modèle viennent corriger ce constat.

Ci-après une vision univariée des tendances des coefficients estimés par sous-poste de soins.

— **Sous-poste frais de séjour (FSEJ) :**

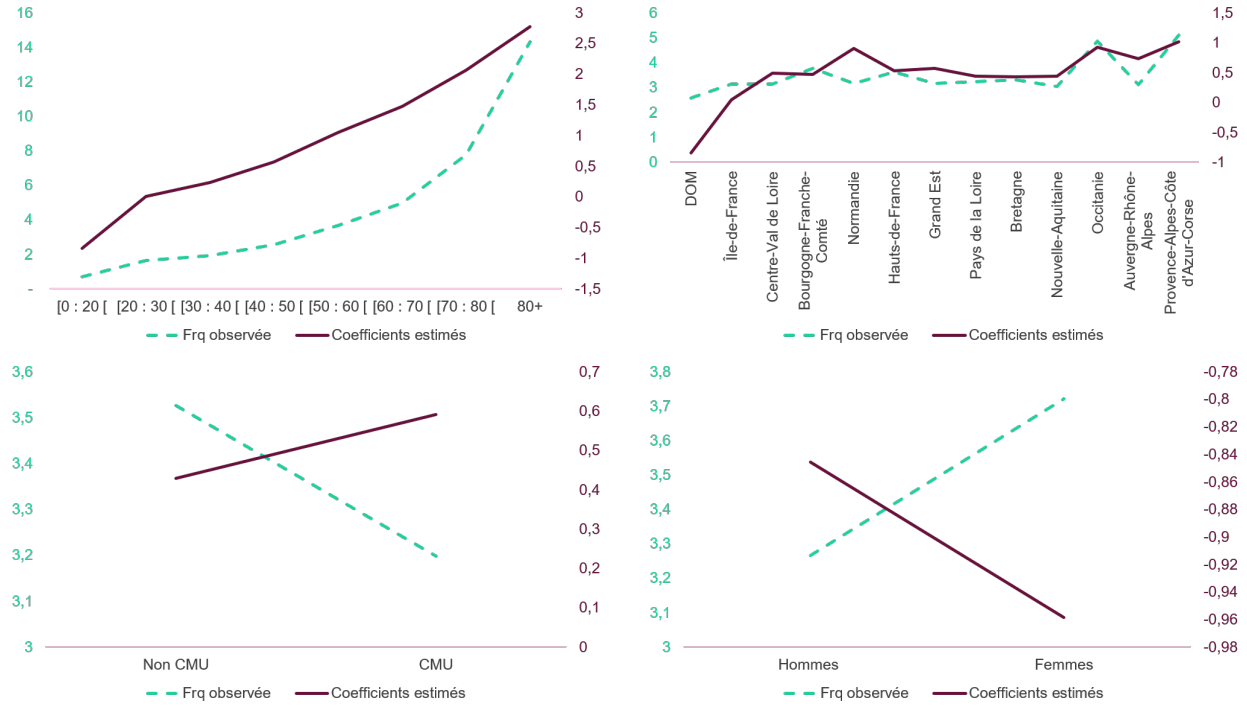


FIGURE 3.48 – Coefficients tarifaires FSEJ - modèle de fréquence

— **Sous-poste chambres (CHBR) :**

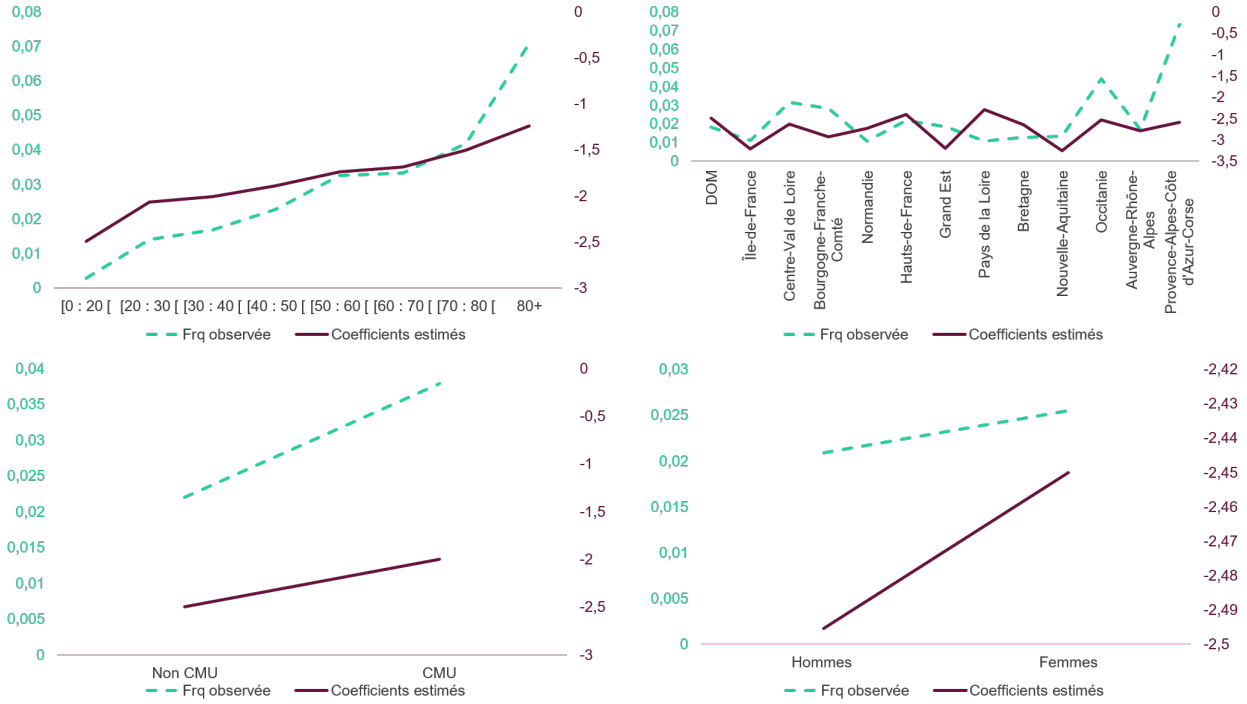


FIGURE 3.49 – Coefficients tarifaires CHBR - modèle de fréquence

— **Sous-poste autres (AUTR) :**

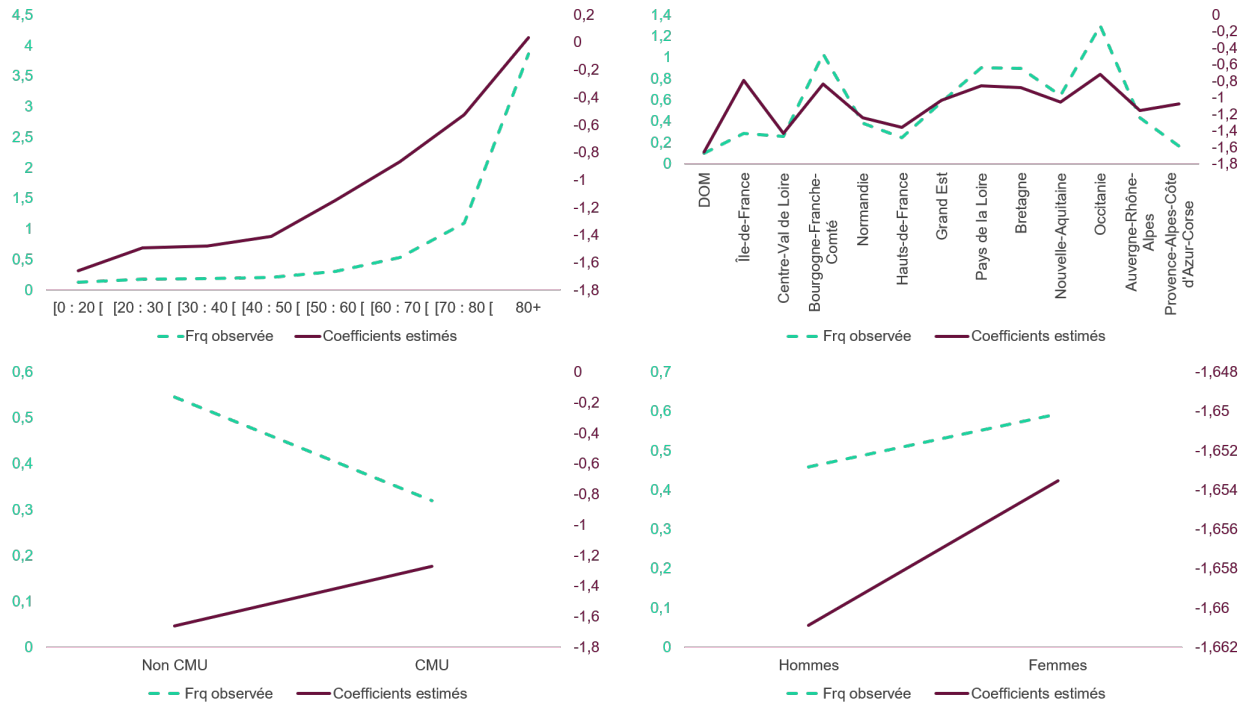


FIGURE 3.50 – Coefficients tarifaires AUTR - modèle de fréquence

Validation du modèle

L'analyse des résidus de chaque modèle conduit aux résultats suivants :

Sous-poste de soins	Moyenne nulle	Homoscédasticité	Indépendance
FSEJ	OK	OK	Non OK
CHBR	OK	OK	Non OK
AUTR	OK	OK	OK

FIGURE 3.51 – Bilan analyse des résidus - modèle de fréquence

Contrairement à la tarification 1, la segmentation par sous-poste de soins induit une perte d'indépendance des résidus sur les sous-postes de soins FSEJ et CHBR.

Qualité prédictive des modèles et comparaison au modèle de coût global

Le passage à une modélisation par sous-poste permet de réduire la RMSE globale [8] du modèle de fréquence de -20% par rapport à la tarification 1. Rapportée au nombre moyen de sinistres [9], cette baisse de RMSE traduit un gain de précision de 1,75%. Ce gain est porté principalement par les FSEJ.

	RMSE	RMSE / nombre moyen de sinistres
FSEJ	15 033	6,74%
CHBR	186	12,61%
AUTR	4 529	14,33%
Vision globale tarification 2	**13 536**	**6,90%**
Tarification 1	**16 957**	**8,65%**
Ecart	**- 3 421**	**-1,75%**

FIGURE 3.52 – Bilan qualité des modèles de fréquence

8. Obtenue en pondérant les RMSE de chaque sous-poste de soins par la part de sinistres enregistrés sur chaque segment.
9. Moyenne pondérée par l'exposition

3.4.5 Prime pure et performance globale du modèle

Pour chacun des sous-postes de soins ainsi modélisés, les primes pures par profil ont été calculées sur la **base de test**. Elles ont ensuite été sommées pour obtenir les primes pures globales par profil. Une pondération par l'exposition a enfin permis d'évaluer la prime pure moyenne prédite.

	FSEJ	CHBR	AUTR	Tarification 2	Tarification 1
Prime pure moyenne observée	460,94	0,85	5,03	466,82	466,82
Prime pure moyenne prédite	490,13	0,82	5,25	496,20	496,17
Ecart	6,334%	-2,886%	4,222%	6,294%	6,287%

FIGURE 3.53 – Résultat modèle de coût-fréquence - tarification par sous-poste

In fine, cette tarification n'induit pas d'amélioration des résultats obtenus par la modélisation globale. Un constat qui vient confirmer la nécessité de reconsidérer la segmentation par sous-poste de soins appliquée usuellement.

Chapitre 4

Apport d'une segmentation des actes par un algorithme de classification non supervisée

L'un des avantages des données *open source* réside dans le volume et l'exhaustivité des informations disponibles qui offrent de nombreuses possibilités en matière d'innovation. Ce fut donc le lieu de s'interroger sur la possibilité d'améliorer la tarification des soins hospitaliers en proposant une alternative à la segmentation actuelle des actes.

Cette partie vise donc à proposer une méthodologie de segmentation des actes hospitaliers par l'utilisation d'un algorithme de classification non supervisée. L'objectif visé par cette segmentation est d'accroitre la précision de la modélisation du risque et donc de la tarification obtenue.

Ainsi, les deux premières sections de cette partie présenteront la notion de classification non supervisée, les divers algorithmes de ce type et *in fine* celui retenu dans cette étude. L'application de cet algorithme et l'impact sur la tarification de la nouvelle segmentation en découlant feront ensuite l'objet de la dernière section.

4.1 La notion de classification non supervisée et de *clustering*

4.1.1 Présentation du principe

La classification non supervisée vise à partitionner une population donnée en sous-groupes homogènes, dits clusters, selon un certain nombre de caractéristiques. La population concernée peut être de diverses natures (individus, objets, actes médicaux...) mais l'objectif final reste le même : évaluer les similarités entre individus pour ensuite créer des catégories de telle sorte que les individus d'un même *cluster* aient des traits similaires entre eux et le plus distincts possible de ceux d'autres groupes.

Dans cet objectif, la notion d'absence de supervision réside dans le fait que la création des *clusters* se fait totalement à l'aveugle : le nombre de classes à constituer est inconnu et aucun échantillon d'apprentissage n'est disponible.

Pour formaliser le problème, il est nécessaire de disposer d'un ensemble de n observations décrites par p caractéristiques. Elles sont supposées ne pas appartenir à une même population homogène, mais plutôt à K groupes différents de populations. En classification non supervisée, K n'est pas

connue encore moins l'appartenance des observations à l'une ou l'autre des K populations. C'est cette segmentation qu'il s'agit de réaliser à partir des p variables disponibles. *A contrario*, la classification sera dite supervisée si le nombre K de classes est connu ainsi que l'appartenance des n observations aux différentes populations. L'objectif dans ce dernier cas s'assimile à une méthode de régression avec pour objectif de plutôt construire une règle de classification sur la base des p descripteurs pour prédire la classe d'appartenance de nouvelles observations.

La classification non supervisée est donc un processus exploratoire où la difficulté majeure est de déterminer les critères permettant de définir la « similitude » entre les points de données.

4.1.2 Mesures de similarité, dissimilarité et notion de distance

En l'absence de toute hypothèse a priori sur la distribution des données, la catégorisation des observations en classes nécessite de définir des mesures de proximité entre observations. Pour ce faire, des mesures de similarité et de dissimilarité ou encore des distances, différentes selon les algorithmes choisis, sont utilisées et se définissent comme suit :

— **Mesure de dissimilarité** : une fonction f est une mesure de dissimilarité si $\forall\ (x, y) \in \mathbb{R}^2$

$$\begin{cases} f(x,y) \geq 0 \\ f(x,y) = f(y,x) \\ f(x,y) = 0 \Leftrightarrow x = y \end{cases}$$

Plus deux observations seront différentes l'une de l'autre, plus cette mesure sera élevée. Ainsi, la distance euclidienne est un bon exemple de mesure de dissimilarité.

— **Mesure de similarité** : une fonction f est une mesure de similarité si $\forall\ (x, y) \in \mathbb{R}^2$

$$\begin{cases} f(x,y) \geq 0 \\ f(x,y) = f(y,x) \\ f(x,x) \geq f(x,y) \end{cases}$$

Contrairement à une mesure de dissimilarité, une telle mesure sera élevée en cas de forte ressemblance de deux observations. La valeur absolue du coefficient de corrélation en est un bon exemple.

— **Distance** : une distance est une mesure de dissimilarité vérifiant en plus la propriété d'inégalité triangulaire. Ainsi, une fonction f est une distance si $\forall\ (x, y, z) \in \mathbb{R}^3$

$$\begin{cases} f(x,y) \geq 0 \\ f(x,y) = f(y,x) \\ f(x,y) = 0 \Leftrightarrow x = y \\ f(x,y) \leq f(x,z) + f(z,y) \end{cases}$$

Ainsi, ces différentes mesures seront utilisées pour maximiser la similitude entre les différents éléments d'une même classe (intra-classe) et *a contrario* limiter cette similitude entre les éléments appartenant à deux classes différentes (inter-classe).

4.1.3 Les différentes méthodes de *clustering*

Après avoir défini le critère d'appréciation de proximité de deux observations données, plusieurs logiques de partitionnement peuvent être adoptée selon l'objectif à atteindre et les données disponibles. Les modèles en découlant se distinguent par leur organisation et leur type de relation. Les plus importants sont :

— **Modèles centroïdes** : ces modèles nécessitent de choisir des observations de référence, dits centroïdes, pour chaque classe à construire. L'appartenance d'une observation à une classe donnée est alors déterminée par le niveau de proximité qu'elle a avec ces différents centroïdes : elle sera affectée au *cluster* dont elle est le plus proche du centre. Ce centre n'est pas figé et devient à chaque affectation le barycentre des observations composant le *cluster* concerné.

La nécessité de fixer des centres de *cluster* exige une très bonne connaissance préalable des données et implique implicitement de définir d'avance le nombre de *clusters* à former, ce qui représente le principal inconvénient de ces algorithmes. De plus, ces derniers ont tendance à être rigides et à vouloir former des *clusters* de taille similaire et de forme elliptique induisant des segmentations incorrectes.

L'algorithme de *clustering K-Means* est un des algorithmes de type centroïdes les plus populaires.

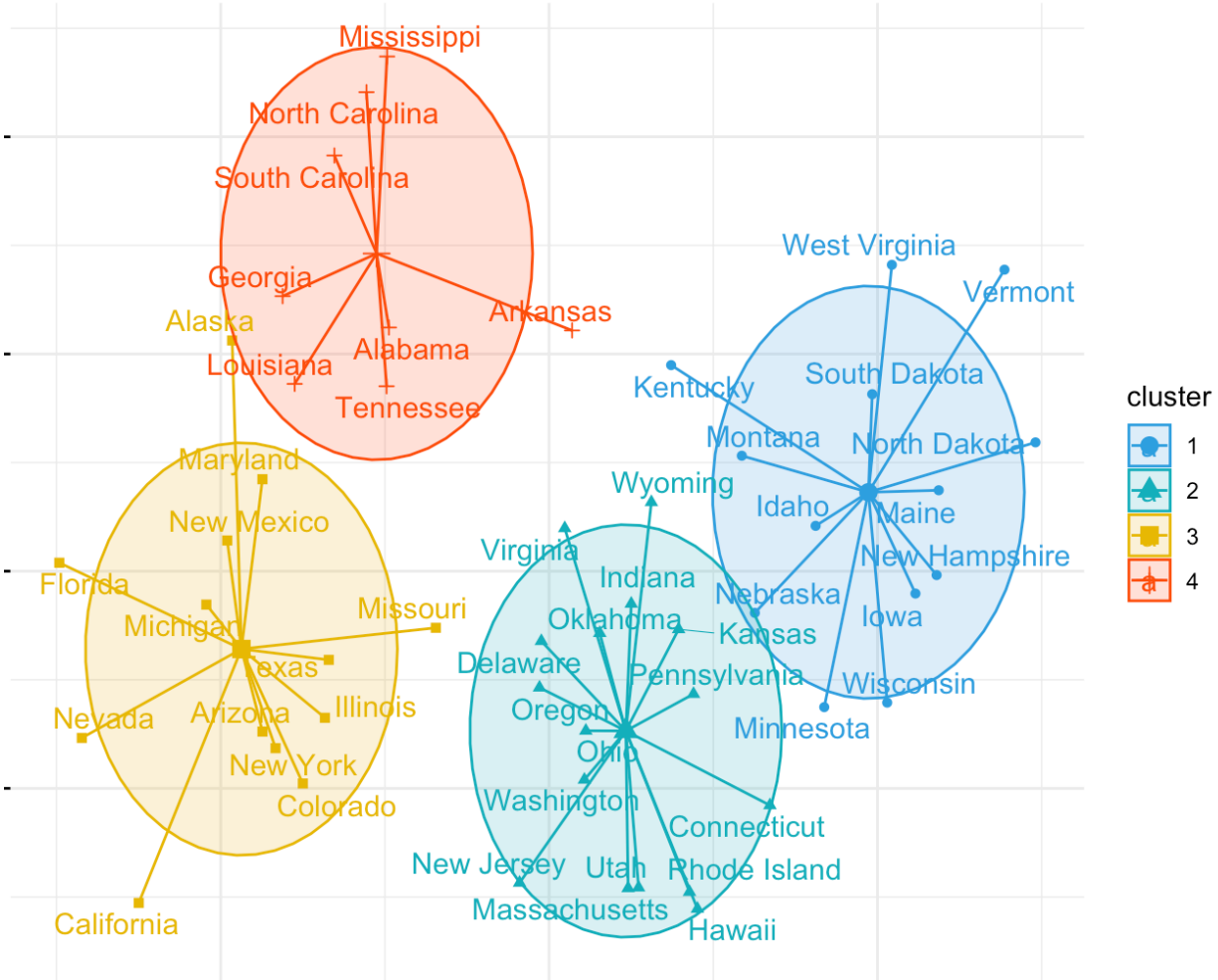


FIGURE 4.1 – Illustration de *clusters* découlant d'un algorithme centroïdes (source : datanovia.com)

— **Modèles de connectivité** : ces modèles suivent deux approches inverses présentées ci-après.

(i) Classification ascendante : chaque observation est considérée comme un *cluster* distinct, puis à chaque étape les *clusters* sont fusionnés en fonction de la distance qui les séparent

toujours avec le souci que cette distance soit la plus faible possible. Un arbre se forme ainsi avec un *cluster* unique à la fin du processus.

(ii) Classification descendante : contrairement à la classification ascendante, dans cette approche, toutes les observations sont considérées comme un tout compact qui est ensuite séparé en blocs de plus en plus fins pour obtenir des *clusters* distincts homogènes.

L'avantage de ces modèles réside dans le fait qu'ils sont faciles à interpréter. Toutefois, face à un grand jeu de données, ils fournissent des arbres particulièrement étendus et peu performants. De plus, ils ne fournissent pas des *clusters* figés mais une hiérarchie de regroupement basée sur la distance entre les points et laisse libre choix à l'utilisateur de former les *clusters* définitifs. Enfin, ils sont très sensibles aux valeurs aberrantes qu'ils considèrent très souvent comme des *clusters* supplémentaires et/ou qui conduisent à la fusion d'autres clusters.

Des exemples de ces modèles sont l'algorithme de *clustering* hiérarchique et ses variantes.

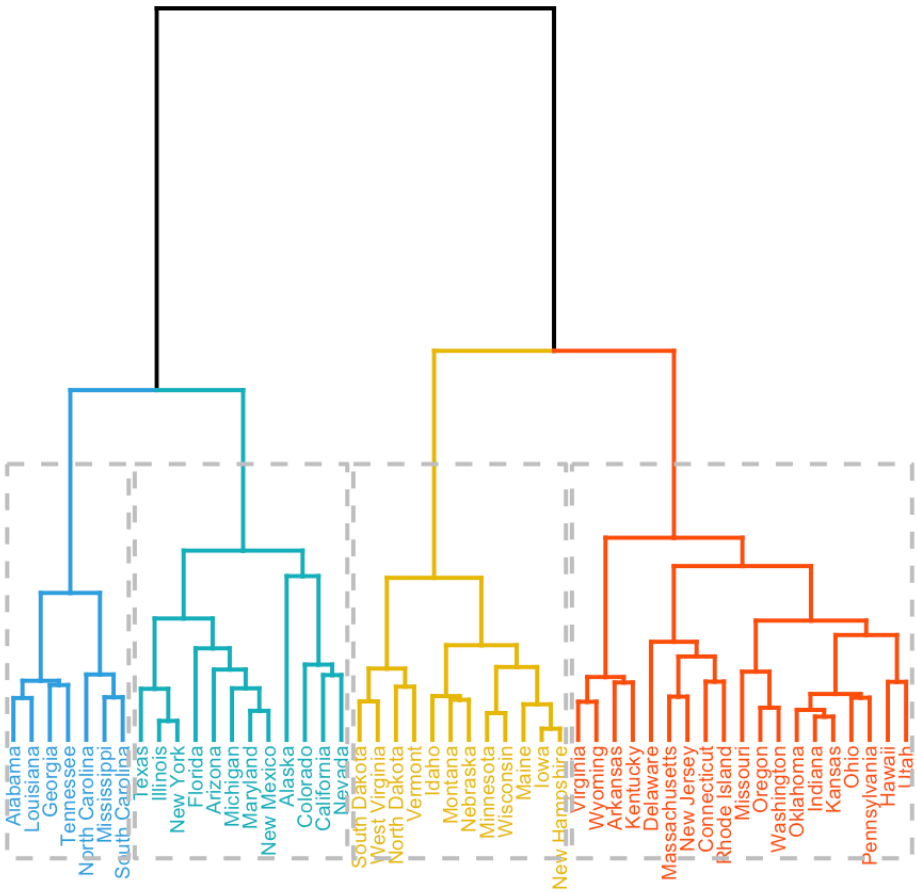


FIGURE 4.2 – Illustration de *clusters* découlant d'un algorithme de connectivité (source : datanovia.com)

— **Modèles de distribution** : l'objectif de ces modèles est de constituer des groupes d'observations suivant une loi statistique similaire. Ici, la distance est remplacée par la probabilité que tous les points d'un *cluster* appartiennent à la même distribution (Normale, Gamma,...). Le principe est de regrouper les observations qui permettent de maximiser la vraisemblance à une loi donnée.

Bien que ce modèle utilise des méthodes d'analyses statistiques usuelles, ils souffrent très souvent de sur-ajustement. De plus, sans contraintes imposées, ils découlent sur des systèmes d'équations très complexes. Un exemple courant de ces modèles est l'algorithme de maximisation des attentes qui utilise des distributions normales multivariées.

— **Modèles basé sur la densité de points** : dans ces modèles, les *clusters* coïncident avec les zones à fortes densité de données. Autrement dit, dans un espace défini par les variables caractérisant les observations à classer, les observations semblables se regrouperont en *cluster* . Ici, les valeurs aberrantes ou qui ne seraient pas proche d'une zone à forte densité ne sont pas considérées comme de nouveaux *clusters* mais plutôt comme du bruit et des points frontières entre clusters.

L'avantage majeur de ces modèles est leur flexibilité sur la forme des *clusters* à constituer. Les *clusters* sont ainsi détectés automatiquement en se basant purement sur la localisation géométrique des points. Ainsi, ces modèles correspondent parfaitement à l'objectif de classification des actes hospitaliers visés dans cette étude. Le modèle retenu sera donc de ce type.

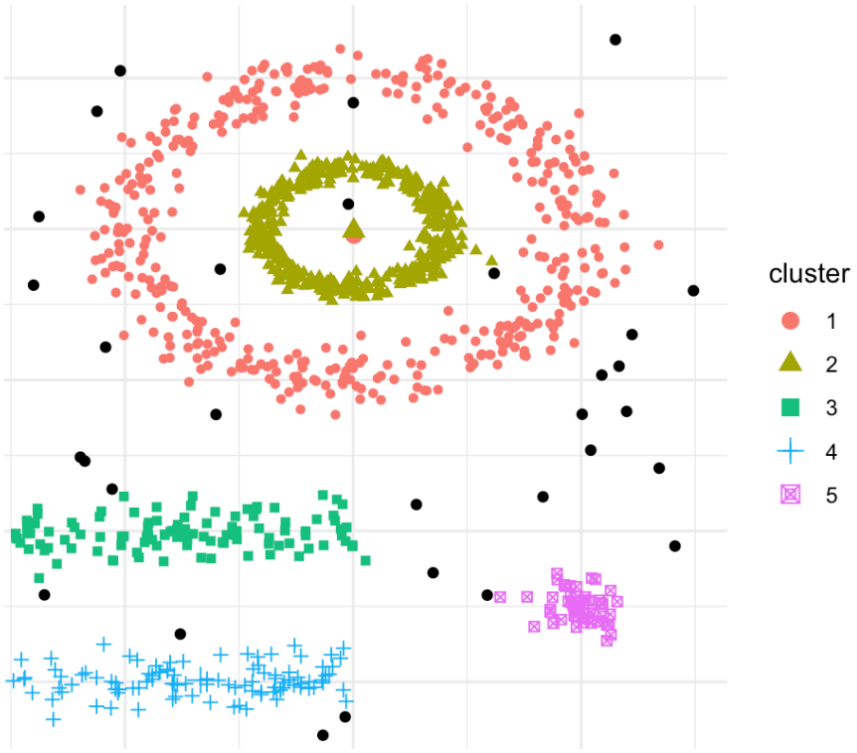


FIGURE 4.3 – Illustration de *clusters* découlant d'un algorithme basé sur la densité (source : datanovia.com)

4.2 Choix de l'algorithme de segmentation

Comme susmentionné, la méthode de segmentation qui correspond le mieux à l'objectif visé par cette étude est celle basée sur les densités. En effet, l'idée est d'obtenir une segmentation naturelle des actes hospitaliers à partir de la similitude des niveaux de coûts moyens et de fréquence de consommation qu'ils présentent.

Ci-après, les principaux algorithmes de *clustering* basés sur la densité des points et celui retenu pour les présents travaux sont présentés.

4.2.1 Présentation des principaux algorithmes de *clustering* basé sur la densité

Il existe plusieurs algorithmes de *clustering* basés sur la densité dont les principaux sont :

— ***Density-based spatial clustering of applications with noise*** **(DBSCAN)** : c'est un algorithme couramment utilisé dans l'exploration de données et l'apprentissage automatique. Il utilise généralement la distance euclidienne comme mesure de dissimilarité et nécessite essentiellement 2 paramètres :

  — ϵ : distance maximale devant séparer deux points pour qu'ils soient considérés comme proches et pouvoir appartenir au même *cluster* ;

  — $minPts$: le nombre minimum de points que doit contenir un regroupement pour être considéré comme un cluster.

  Avec ces paramètres renseignés, l'algorithme part d'un point arbitrairement choisi et évalue tous les points à proximité : s'il y a au minimum $minPts$ points proches de ce point au sens de ϵ alors un *cluster* est formé, sinon, ce point est considéré temporairement comme du bruit, et il le restera si la répétition du processus avec d'autres observations ne conduit pas à l'intégrer dans un cluster.

  Le principal inconvénient de cet algorithme réside dans le choix des paramètres dont les niveaux conditionnent totalement les résultats. L'utilisateur doit donc faire ce réglage de manière experte en maîtrisant parfaitement ses données. De plus, considérer des paramètres identiques pour tous les *clusters* à construire introduit une rigidité dans l'algorithme qui sera peu performant si les classes présentent des densités hétérogènes. Enfin, la difficulté de définir ces paramètres est croissante avec le volume de données à traiter.

— ***Ordering points to identify the clustering structure*** **(OPTICS)** : cet algorithme est basé sur le même principe que DBSCAN mais en élimine ses principales insuffisances : l'obligation de faire un choix judicieux du ϵ et l'impossibilité de détecter des *clusters* de densités différentes dû à ce paramètre. En effet, dans OPTICS ce paramètre est optionnel car l'algorithme se base sur de nouvelles mesures de dissimilarités.

  L'algorithme fonctionne de manière itérative en considérant chaque point comme de possibles centres de *clusters* en calculant pour chaque cluster deux distances :

  — Distance centrale : elle représente la distance qui sépare le centre du *cluster* du point le plus éloigné parmi les $minPts$ qui l'entourent.

  — Distance d'accessibilité mutuelle : pour deux points p1 et p2, c'est la valeur maximale entre leurs distances centrales respectives et la distance qui les sépare l'un de l'autre.

Au début de la procédure, une distance centrale est attribuée à chaque observations puis un point p est choisi comme potentiel centre de cluster. Les distances d'accessibilité sont alors calculées pour les points environnants en traitant chaque point une seule fois. Le *cluster* ayant pour centre le point p sera composé par les minPts ayant les distances d'accessibilité les plus faibles. Le prochain point choisi pour ce traitement sera celui qui a la distance d'accessibilité la plus proche. S'en suit alors une succession de *clusters* avec des centres différents hiérarchisés selon leurs distances centrales tel un dendrogramme. La formation des *clusters* définitifs est donc soumise au libre arbitre de l'utilisateur introduisant ainsi de la subjectivité dans les conclusions du modèle.
Par ailleurs, dans ce processus, bien que le paramètre epsilon ne soit pas nécessaire, le préciser permet de limiter le rayon d'évaluation des distances d'accessibilité. Autrement, à chaque itération c'est la base de données entière qui est parcourue, entraînant un long temps de calcul et un coût en mémoire considérable.

— ***Hierarchical DBSCAN*** : cet algorithme est une version améliorée de DBSCAN qui ne nécessite plus de préciser un seuil de distance minimale de proximité mais uniquement une taille minimale des *clusters* à former. En effet, l'algorithme réalise plusieurs fois la procédure de DBSCAN en faisant varier les valeurs d'epsilon afin d'obtenir des *clusters* avec la meilleure stabilité sur epsilon. Il permet ainsi de trouver des *clusters* de densités variables.
Comparativement aux deux autres algorithmes, HDBSCAN est de fait l'algorithme le plus axé sur les données en nécessitant le moins d'informations provenant de l'utilisateur. De plus, contrairement à l'algorithme OPTICS, il ne sollicite pas d'interprétation subjective des résultats et nécessite moins de temps de calcul.
C'est cet algorithme qui est retenu pour les présents travaux et qui sera détaillé dans la partie suivante.

4.2.2 Présentation du fonctionnement de HDBSCAN

HDBSCAN est un algorithme de *clustering* développé par Ricardo CAMPELLO, Davoud MOULAVI et Joerg SANDER du département des sciences informatiques de l'Université de l'Alberta, Edmonton au Canada. Il transforme DBSCAN en un algorithme de *clustering* hiérarchique dont il est ensuite extrait des *clusters* plats selon un critère de stabilité. Pour y parvenir, HDBSCAN suit un processus en 6 grandes étapes [1] :

— **Etape 1 : construction d'une nouvelle métrique d'évaluation des zones denses et des zones clairsemées**
Afin de détecter les zones à forte densité et les points éloignés, une nouvelle mesure d'éloignement et de proximité des données est adoptée. Elle est nommée dans la littérature distance d'accessibilité mutuelle ($d_{acc}(.,.)$) et a pour objectif d'accentuer l'écart dans l'espace entre les points appartenant à une zone dense et ceux appartenant à une zone clairsemée. Plus formellement, elle se construit à partir de la distance centrale ($d_{cent}(.)$) correspondant à la distance maximale entre un point et les $minPts$ points qui l'entoure. En effet, pour deux points a et b donnés, la distance d'accessibilité est le maximum entre leurs distances centrales respectives et la distance d'origine ($d(.,.)$) qui les séparent.

$$d_{acc}(a,b) = max(d_{cent}(a), d_{cent}(b), d(a,b))$$

Ainsi, les points appartenant à des zones clairsemées sont de fait encore plus marqués afin de clairement les distinguer des points denses.
Le schéma ci-après montre un exemple de transformation des distances initiales en distances d'accessibilité mutuelle pour $minPts = 5$.

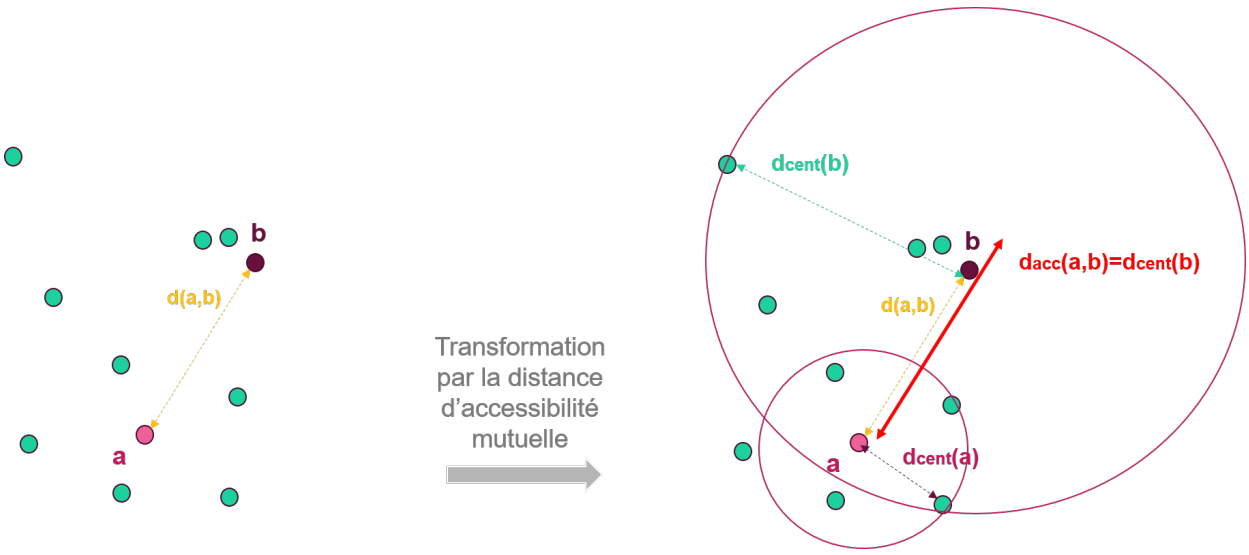


FIGURE 4.4 – Illustration d'une transformation de l'espace par la distance d'accessibilité mutuelle

1. Les graphiques des étapes 2 à 5 sont issus de : `https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html`

— **Etape 2 : construction d'arbres minimums couvrants**
Après avoir redéfini la distance entre les points, la seconde étape consiste à créer des groupes de points minimisant le chemin les reliant. Des grappes de connectivité sont alors construites illustrant les différentes possibilités de liaisons.

L'ampleur d'une grappe est conditionnée par un seuil sur la longueur des branches à maintenir dans la grappe. Pour avoir des grappes homogènes, la logique voudrait que ce seuil soit mis à un niveau élevé puis réduit progressivement afin de former des grappes de plus en plus compactes. Toutefois, ce jeu de variation de seuil demande un temps de calcul important surtout que l'objectif est d'arriver à un ensemble minimal tel que la suppression d'une arrête supplémentaire conduise à la dislocation de toute la grappe.

Par conséquent, pour y parvenir, l'algorithme utilise la théorie des graphes pour obtenir des arbres couvrants minimums de chaque grappe mère préalablement constituées. Le principe est de former des grappes dont la somme des poids des arêtes est minimale : à un point donné, on rattachera le point qui a la distance d'accessibilité mutuelle la plus faible et ainsi de suite.

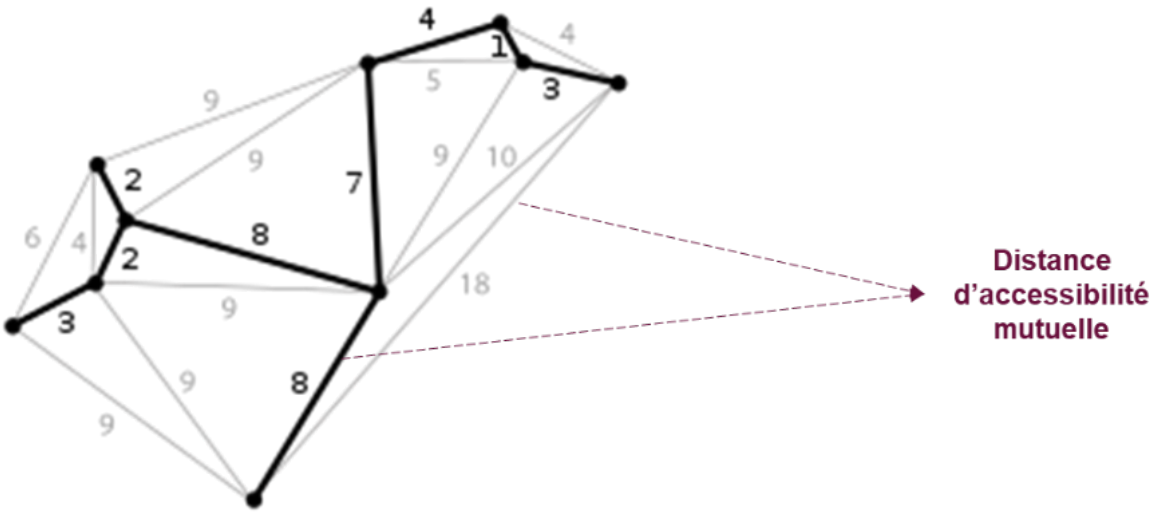


FIGURE 4.5 – Illustration d'un arbre minimum couvrant

— **Etape 3 : hiérarchisation des *clusters***
En fonction de la longueur des branches reliant les points des arbres minimums couvrants obtenus, une arborescence se créée et permet d'introduire une hiérarchie dans les *clusters* qui peut être visualisée sous forme de dendrogramme.

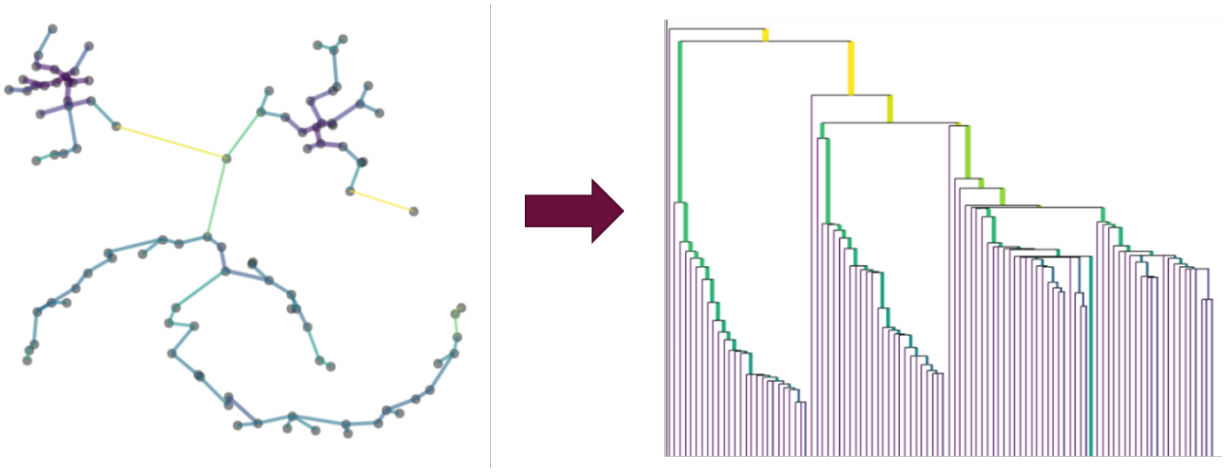


FIGURE 4.6 – Illustration du passage d'un arbre de liaison à un dendrogramme

— **Etape 4 : condensation de l'arborescence des clusters**
Cette étape vise à former des *clusters* contenants un nombre minimal choisi de points. Ainsi, toute l'arborescence hiérarchisée est parcourue et à chaque nouvelle proposition de séparation du dendrogramme (proposition de formation d'un nouveau cluster), si le nouveau *cluster* proposé ne correspond pas à la taille minimale exigée, la segmentation est supprimée et le *cluster* à former est considéré comme une partie du *cluster* précédent. *A contrario*, la segmentation est maintenue et un nouveau *cluster* est créé si celui-ci contient au moins minPts points. L'arbre initial se retrouve ainsi réduit avec des grappes de densité différente.

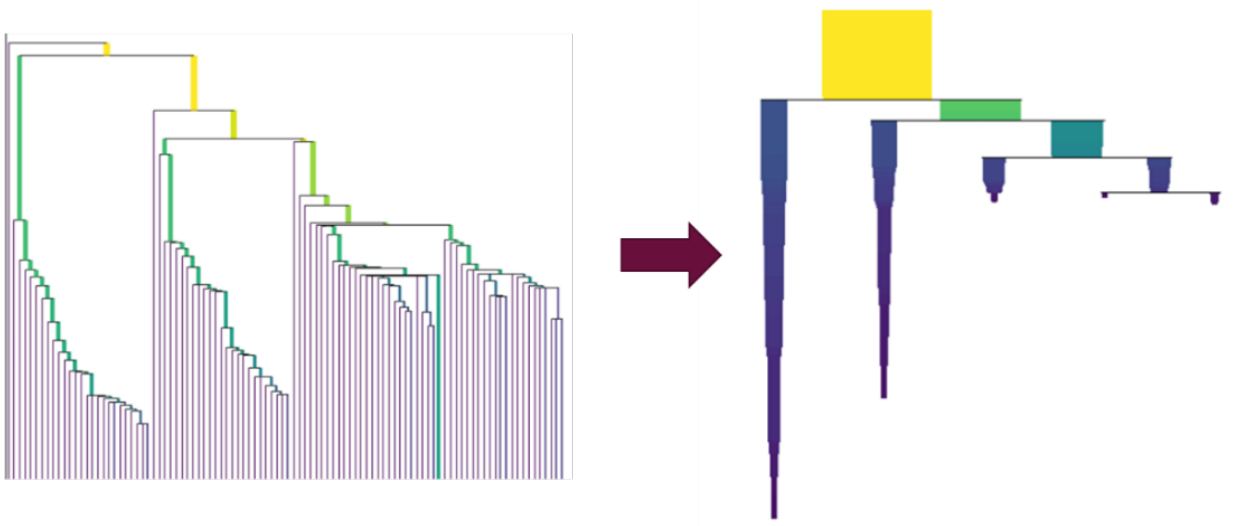


FIGURE 4.7 – Illustration du processus de condensation d'un dendrogramme

— **Etape 5 : extraction des *clusters* plats**
L'objectif de l'algorithme n'est pas de fournir une hiérarchie des *clusters* mais les *clusters* définitifs à adopter. Ainsi une mesure permettant d'évaluer l'importance de chaque *cluster* est considérée :

$$\lambda = \frac{1}{(\text{distance d'accessibilité mutuelle entre les points du noeud de segmentation})}$$

Sur chaque nœud de segmentation une comparaison est faite entre le *cluster* initial et ses *clusters* « fils » en calculant pour chacun d'eux trois indicateurs :

$$\begin{cases} \lambda_{\text{birth}} = \lambda \text{ au noeud initial de création du } \textit{cluster} \text{ concerné} \\ \lambda_{\text{death}} = \lambda \text{ au moment de la segmentation éventuelle du } \textit{cluster} \text{ concerné} \\ s = \sum_{p \in \text{cluster}} \lambda_p - \lambda_{\text{birth}} \\ \text{avec} \begin{cases} p \text{ indicant un point du } \textit{cluster} \text{ considéré} \\ \lambda_p = \lambda \text{ au moment où le point p quitte le } \textit{cluster} \text{ considéré dû à une segmentation} \end{cases} \end{cases}$$

Ainsi, pour un *cluster* donné, s'il présente une stabilité (noté "s") inférieure à la somme des stabilités de ses *clusters* « fils », la segmentation est retenue en l'état, dans le cas contraire les *clusters* « fils » ne sont pas considérés comme de nouveaux *clusters* à retenir.

Les *clusters* plats recherchés sont ainsi obtenus et tout point n'appartenant pas à un *cluster* sélectionné est considéré comme un bruit.

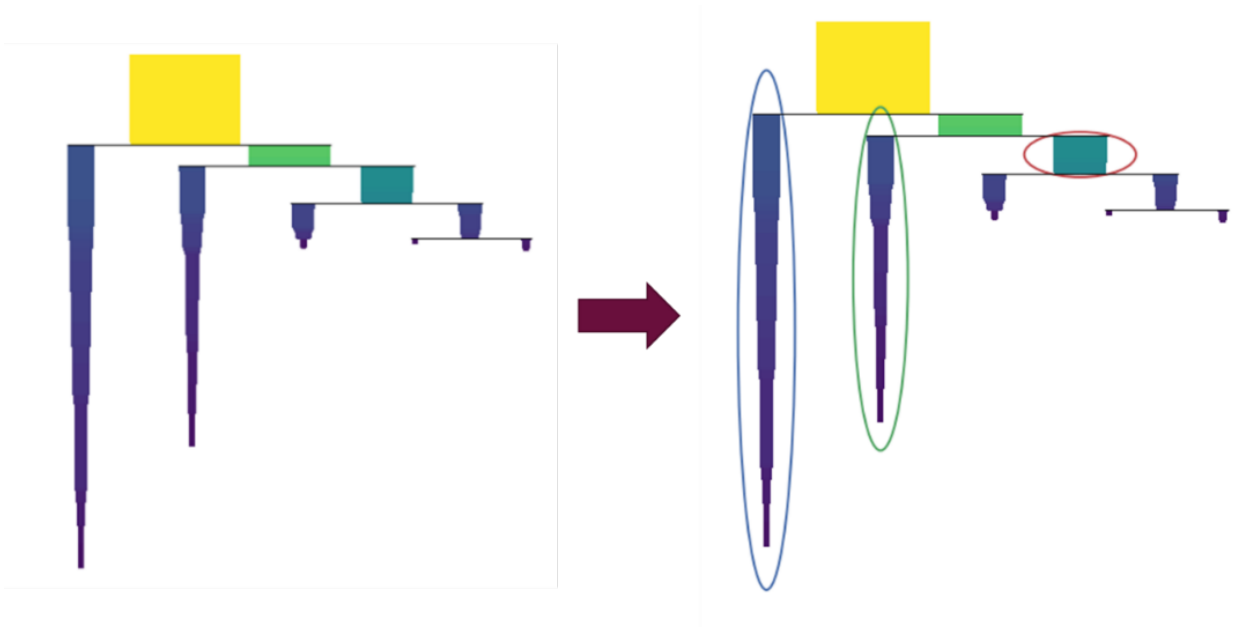


FIGURE 4.8 – Illustration du choix des *clusters* plats

— **Etape 6 : étiquetage des observations**
L'algorithme fournit comme *output* un étiquetage des données (*cluster* d'appartenance) et les λ_p normalisés (variant entre 0 et 1) décrivant ainsi la force d'appartenance de chaque point à un *cluster* donné.

4.2.3 Présentation des *outputs* fournis par HDBSCAN

En plus de fournir une segmentation des points, différents indicateurs sont disponibles après l'application de l'algorithme :

- le **nombre de points considérés comme du bruit** : ces points n'appartiennent à aucun *cluster* et présentent des caractéristiques qui s'écartent des principales tendances moyennes. Le bruit a la particularité de contenir tant les valeurs extrêmes hautes que faibles. Par ailleurs, il reste fortement influencé par le nombre minimum de points à considérer pour constituer un cluster.
- le **nombre de points par clusters** créés : ce nombre varie fortement avec le nombre minimum de points à considérer pour constituer un cluster.
- la **probabilité d'appartenance** de chaque point aux *clusters* créés : c'est la force avec laquelle un point est lié au *cluster* auquel il a été assigné. Un point bruit aura une probabilité d'appartenance de 0. Celle des autres points dépend fortement de la persistance du *cluster* considéré au fil des variations du paramètre $minPts$.
- la **stabilité des clusters** : elle correspond à la somme des scores de stabilité (s) pour chaque cluster. Une valeur qui permet de comparer le niveau de stabilité de différents *clusters* : plus elle est élevée, plus le *cluster* est stable.

Ce sont ces différents indicateurs qui serviront de base d'évaluation des options de *clustering* des soins hospitaliers.

4.3 Application et résultats

L'utilisation de HDBSCAN dans le processus de tarification requiert plusieurs étapes préliminaires avant d'aboutir à une classification optimale des actes et à la réalisation de la modélisation de la prime pure. Ces différentes étapes sont :

— sélection des variables de segmentation cohérentes avec l'objectif recherché d'homogénéité des distributions intra *cluster* ;

— détermination du nombre minimum optimal d'actes médicaux devant constituer un *cluster* ;

— construction des *clusters* finaux cohérents avec les besoins de l'étude à partir des *clusters* bruts obtenus ;

— modélisation tarifaire sur la base des *clusters* finaux retenus.

Ces travaux ont été réalisés respectivement sur le coût moyen des sinistres et sur leur fréquence d'occurence, avant d'évaluer la prime pure.

4.3.1 Sélection des variables de segmentation

Comme susmentionné dans la description de HDBSCAN, l'algorithme analyse les distances entre les points afin d'évaluer les *clusters* à créer. Il est donc nécessaire de caractériser un point par différentes variables. L'objectif étant ici de parvenir à des *clusters* de distributions homogènes, les variables considérées sont les indicateurs classiques caractérisant une densité, à savoir :

— moyenne ;

— écart-type ;

— minimum ;

— maximum ;

— médiane ;

— coefficient de variation ;

— kurtosis ;

— skewness.

L'évaluation de ces indicateurs a été réalisée pour chaque acte sur la base annuelle globale. Les valeurs brutes obtenues ont ensuite été normalisées pour éviter les écarts importants de niveaux de valeurs pouvant être une source de biais lors de la classification. La méthode de normalisation qui a été appliquée pour un indicateur X donné est la suivante :

$$X_{normal} = \frac{(X - \bar{X})}{\sigma_X} \text{ avec } \bar{X} \text{ moyenne de } X \text{ et } \sigma_X \text{ son écart-type}$$

HDBSCAN n'impose aucune contrainte sur le nombre maximum de variables de segmentation à considérer. Toutefois, afin d'éviter que la correlation existante entre certains de ces indicateurs ne réduise la performance de l'algorithme, une sélection de la combinaison la plus pertinente de ces variables a été réalisée.

Cette sélection s'est déroulée selon un procédé *forward* avec pour chaque nouveau jeu de variables, le calcul de plusieurs indicateurs d'évaluation de l'optimalité dudit jeu selon le procédé suivant :

- Application de HDBSCAN au jeu de donné considéré, et ce, pour chaque valeur de $mintPts$ possible ;
- Pour chaque $minPts$ calcul/récupération des valeurs suivantes :
  - nombre de points bruits ;
  - nombre de *clusters* ;
  - probabilité moyenne d'appartenance des actes aux *clusters* formés. Cette moyenne a été pondérée par le poids de chaque acte dans le coût total des sinistres afin de pénaliser les segmentations qui conduiraient à une mauvaise affectation des actes les plus importants en matière de sinistralité.
  - stabilité moyenne des *clusters* formés (moyenne arithmétique).
- Evaluation des niveaux moyens des indicateurs précédents tous $minPts$ confondus :
  - nombre moyen de points bruits (pondéré par le nombre de *clusters* formés pour chaque $minPts$) ;
  - nombre moyen de *clusters* (moyenne arithmétique) ;
  - probabilité moyenne d'appartenance (moyenne arithmétique) ;
  - stabilité moyenne des *clusters* (moyenne arithmétique).

Le schéma ci-après résume le procédé explicité pour un jeu donné de variables :

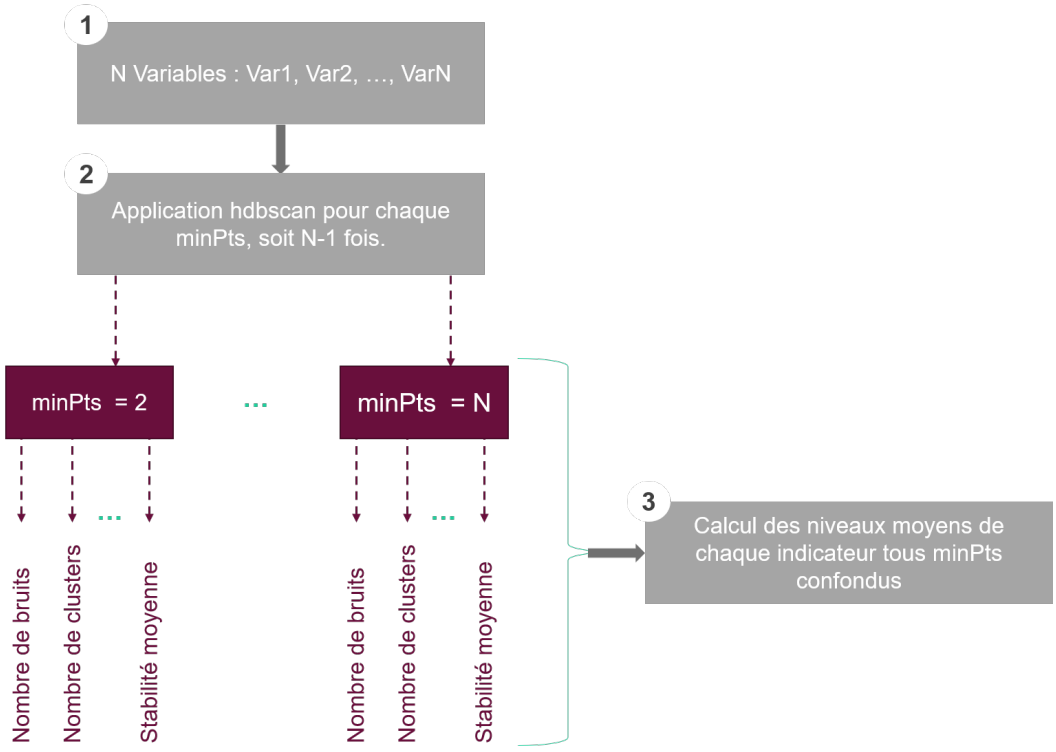


FIGURE 4.9 – Méthodologie d'évaluation des indicateurs de sélection des variables de segmentation

L'application de cette méthodologie aux variables caractérisant les distributions du coût moyenet du nombre de sinistres, a conduit aux résultats présentés ci-après. Ils découlent de l'utilisation de la distance euclidienne comme distance itiniale fournis à HDBSCAN.

Variables	Nombre moyen de bruits	Nombre moyen de clusters	Probabilité moyenne d'appartenance aux clusters	Stabilité moyenne des clusters
Moyenne	63,21	10,10	27,80 %	769 015,79
Moyenne, écart-type	119,45	7,82	24,18 %	486 464,07
Moyenne, écart-type, minimum	105,73	14,82	17,93 %	189 888,48
Moyenne, écart-type, minimum, maximum	117,60	15,22	17,77 %	143 811,78
Moyenne, écart-type, minimum, maximum, coefficient de variation	112,84	7,19	21,33 %	3 883,43
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis	114,34	14,80	16,38 %	1 334,95
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis, médiane	116,31	16,00	15,56 %	998,94
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis, médiane, skewness	111,81	12,73	13,97 %	1 472,89

FIGURE 4.10 – Indicateurs sélection de variable - *clustering* coût moyen

Variables	Nombre moyen de bruits	Nombre moyen de clusters	Probabilité moyenne d'appartenance aux clusters	Stabilité moyenne des clusters
Moyenne	48,02	25,50	32,96 %	8 815 638,44
Moyenne, écart-type	81,80	21,67	18,83 %	2 482 247,43
Moyenne, écart-type, minimum	119,29	11,58	16,97 %	27 145,15
Moyenne, écart-type, minimum, maximum	126,55	11,25	14,98 %	26 174,06
Moyenne, écart-type, minimum, maximum, coefficient de variation	108,02	10,75	16,46 %	2 593,28
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis	107,90	9,63	16,31 %	2 763,52
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis, médiane	110,30	9,69	16,42 %	2 662,41
Moyenne, écart-type, minimum, maximum, coefficient de variation, kurtosis, médiane, skewness	129,22	5,08	12,79 %	1 085,99

FIGURE 4.11 – Indicateurs sélection de variable - *clustering* fréquence

Le choix du jeu de donné adéquat s'est basé sur un objectif de minimisation du bruit et du nombre de *clusters* à former, tout en cherchant à obtenir des niveaux élevés de probabilité d'appartenance aux *clusters* et de stabilité.

Ainsi, les variables de segmentation retenues sur le coût moyen sont la moyenne et l'écart-type. Sur la fréquence, bien que la moyenne puisse être utilisée comme unique variable de segmentation, l'écart-type fut également retenu afin d'avoir un second critère de classification et un référentiel de segmentation similaire à celui du coût moyen.

4.3.2 Détermination du nombre minimal de points par cluster

Le choix du nombre minimal de points par *cluster* a été déterminé en analysant, les résutats des deux premières étapes réalisées selon le schéma de la figure 4.9 pour le jeu de variable composé de la moyenne et de l'écart-type. Une analyse graphique des *clusters* proposés a également été réalisée.

— **Recherche du *minPts* optimal - coût moyen**
Sur le coût moyen, les *minPts* pouvant être considérés sont compris entre 2 et 18. Au delà, aucun regroupement n'est possible et tous les points sont considérés comme du bruit. Parmi les *minPts* testés, celui égale à 9 se démarque des autres en permettant de limiter le nombre de *clusters* et le nombre de points bruits, tout en conduisant à la probabilité moyenne d'appartenance la plus élevée.

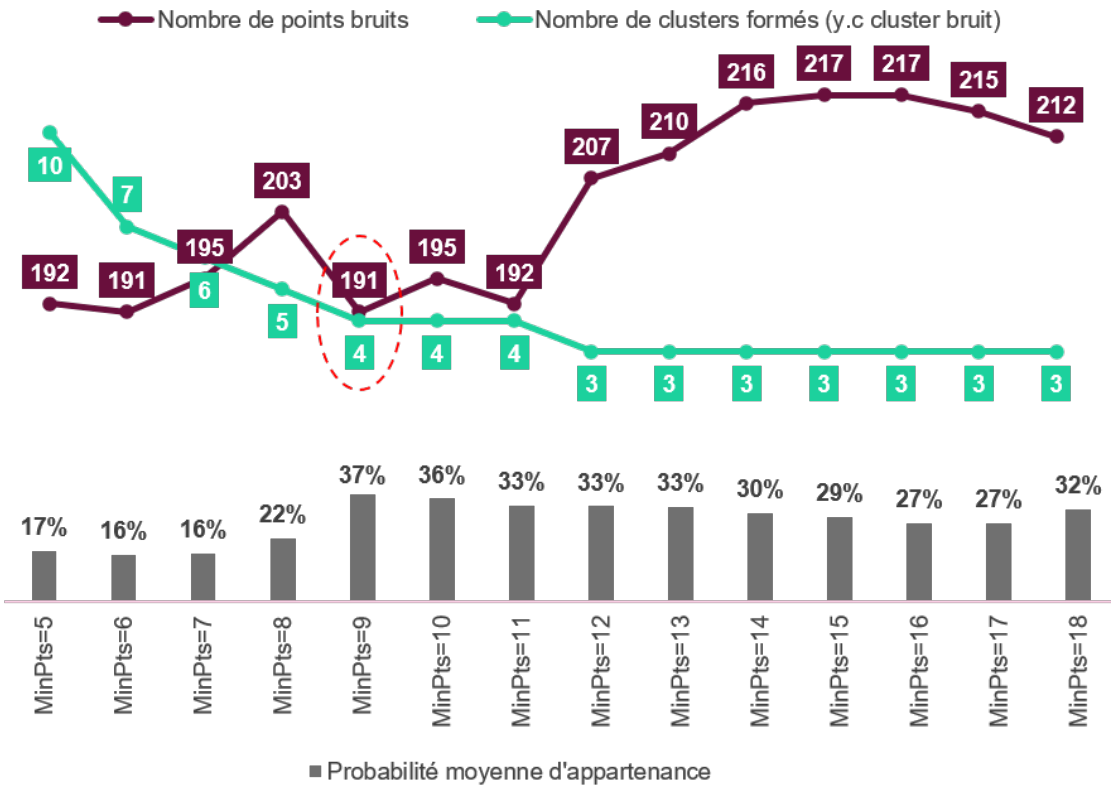


FIGURE 4.12 – Indicateurs sélection du *minPts* optimal - *clustering* coût moyen

Par ailleurs, une analyse graphique des jeux de *clusters* induits par les différents $minPts$ met en exergue la persistance, à quelques points près, de la segmentation découlant d'un $minPts = 9$ pour un nombre de *clusters* similaires. La segmentation qui en découle et les densités de chaque *cluster* (hors *cluster* bruit qui en constitue le *cluster* 1) sont présentées ci-après.

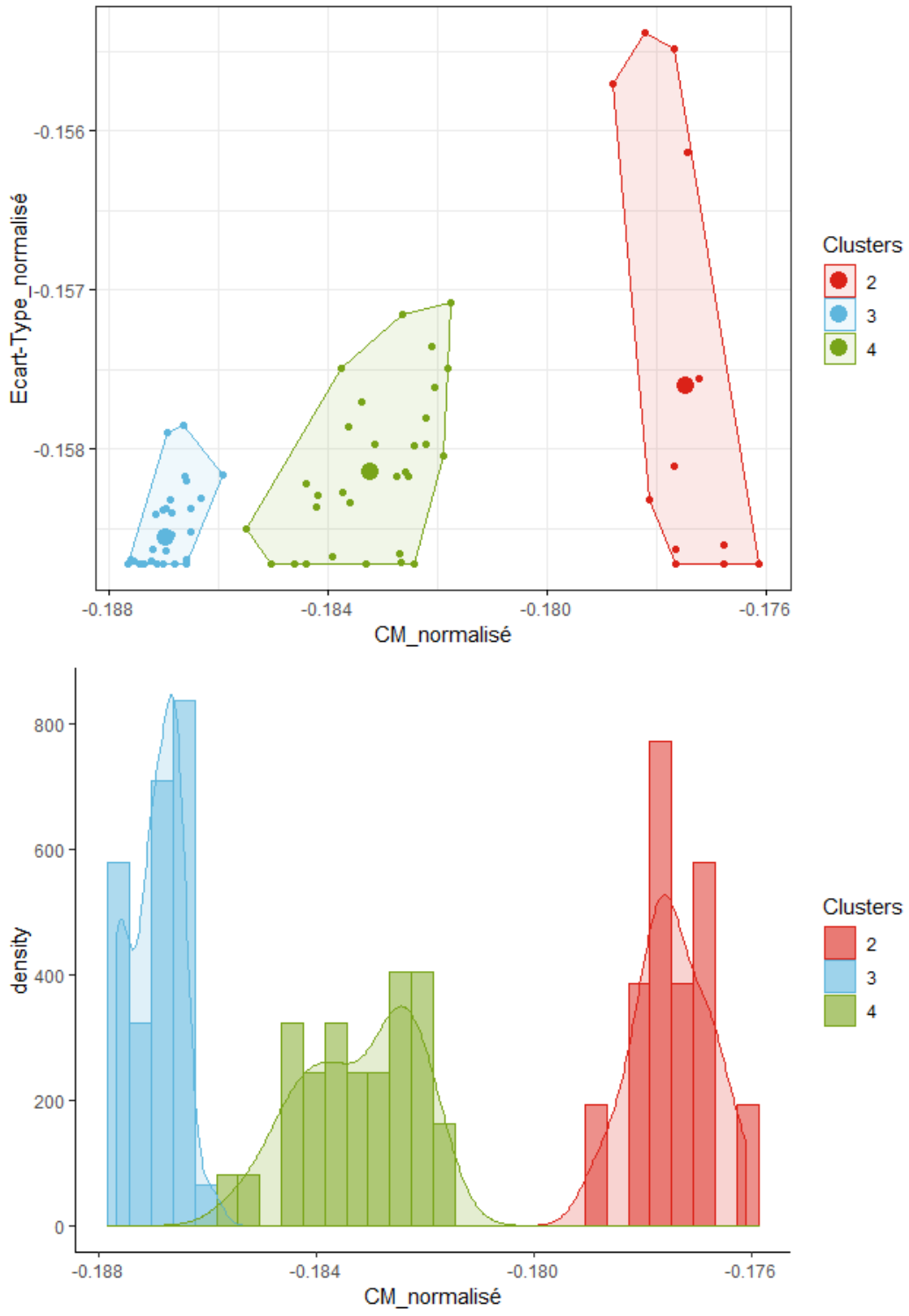


FIGURE 4.13 – Segmentation $minPts = 9$ - *clustering* coût moyen

Ainsi, les actes sont segmentés entre ceux ayant un coût moyen faible, moyen et élevé. Les valeurs extrêmes basses et hautes constituent le *cluster* bruit (non représenté : voir annexe D).

— **Recherche du *minPts* optimal - fréquence**
Sur la fréquence, c'est le $minPts = 7$ qui est le plus adéquat. En effet, cette valeur de minPts permet à la fois de minimiser les nombres de *clusters* et de points bruits tout en maximisant la probabilité d'appartenance moyenne.

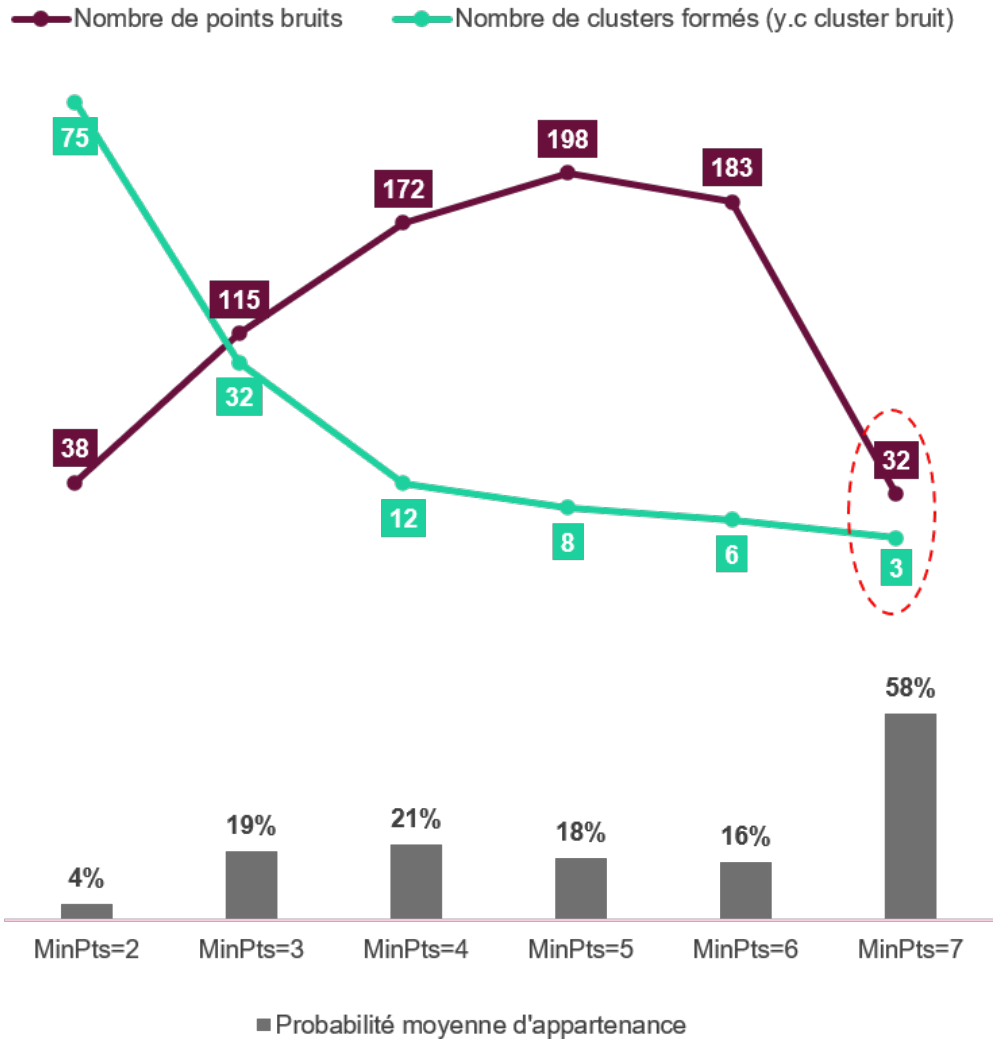


FIGURE 4.14 – Indicateurs sélection du *minPts* optimal - *clustering* fréquence

La représentation graphique des *clusters* obtenus (hors bruit) et de leurs densités est la suivante :

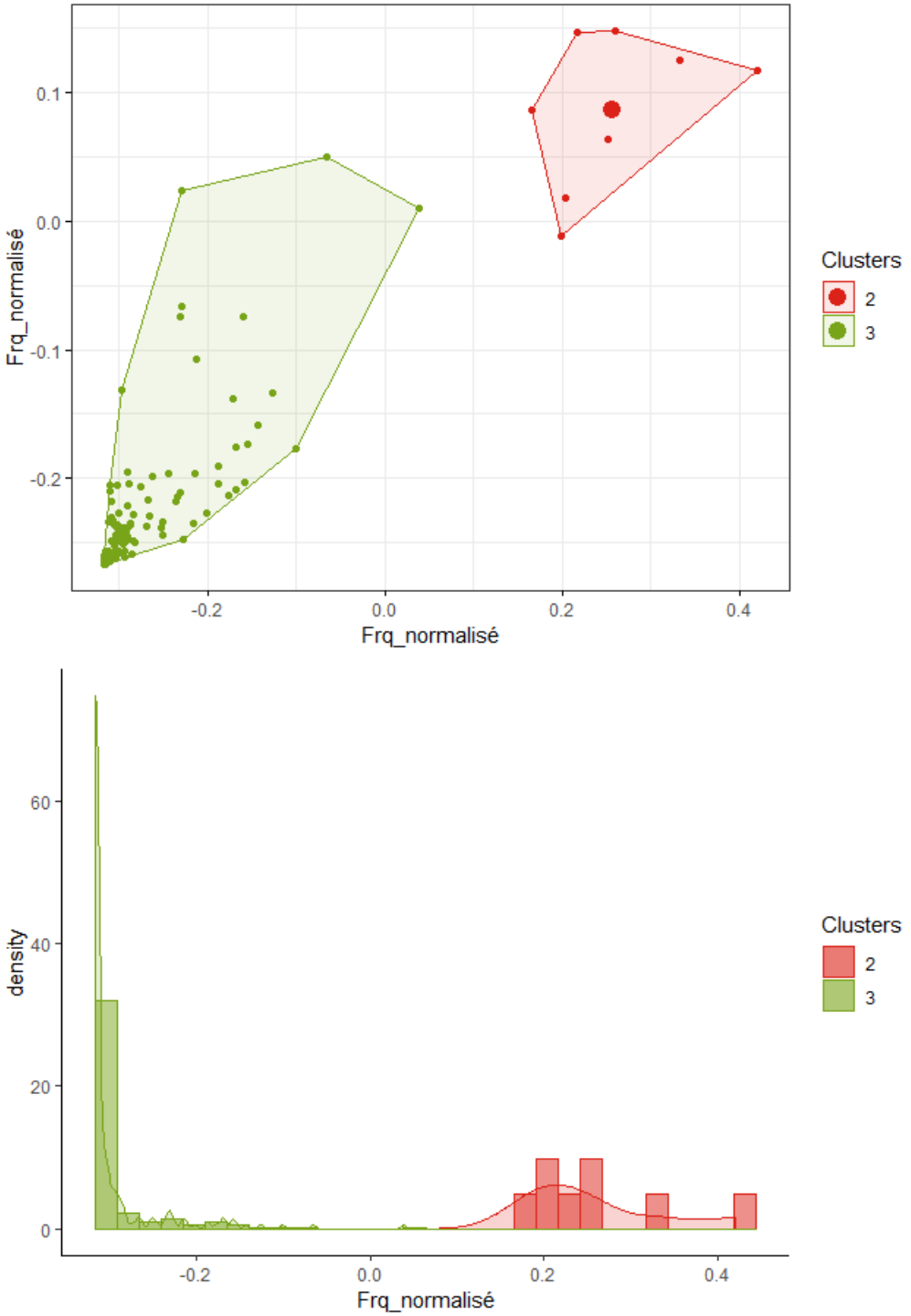


FIGURE 4.15 – Segmentation $minPts = 7$ - *clustering* fréquence

Le *cluster* 3 représente les actes avec des fréquences de consommations les plus faibles et le *cluster* 2, les actes qui ont une fréquence moyenne. Le *cluster* bruit quant lui (non représenté : voir annexe D) est principalement composé des valeurs extrêmes hautes.

4.3.3 Construction des *clusters* adaptés à l'étude

L'ojectif final de ce processus de *clustering* est la réalisation d'une tarification plus fine de la garantie hospitalisation que celle basée sur de la segmentation représentée ci-après :

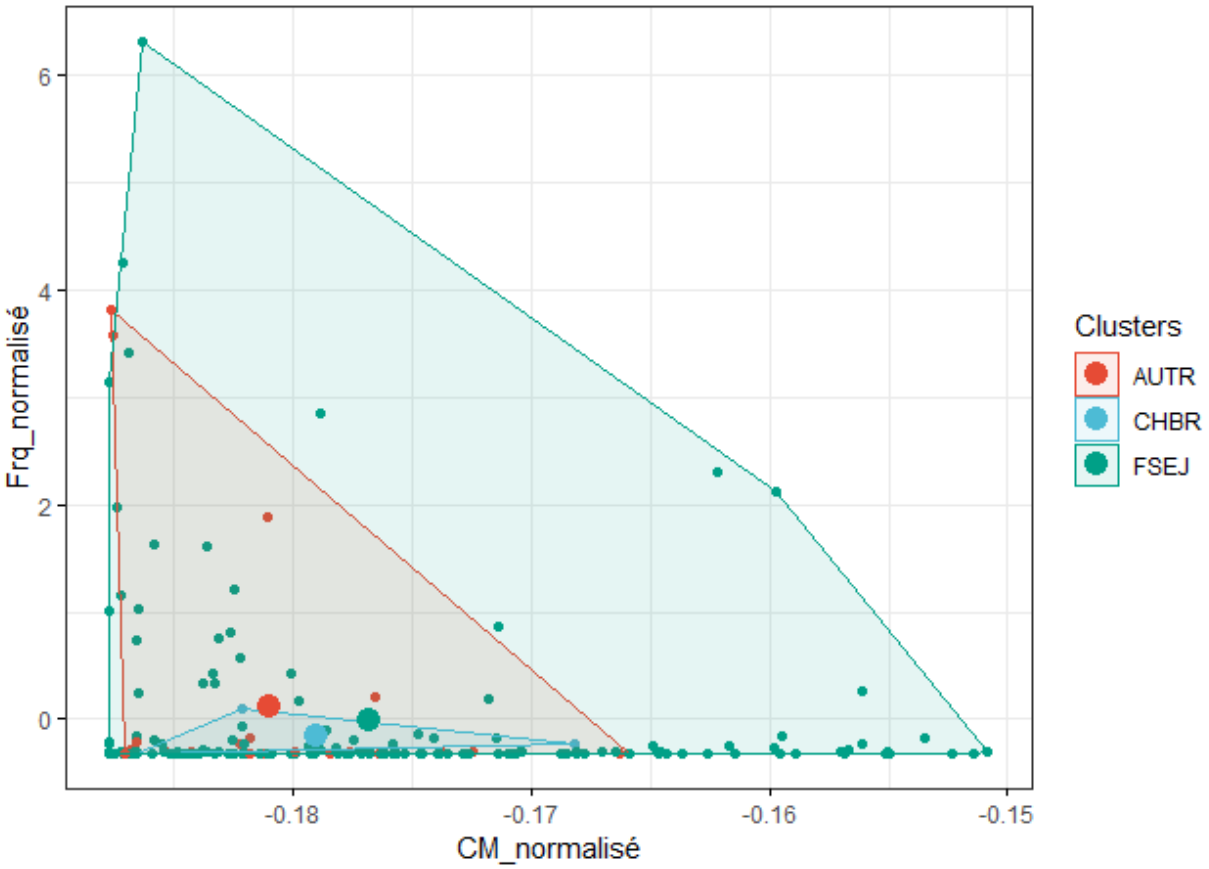


FIGURE 4.16 – Segmentation usuelle assureur (tarification 2)

La création de *clusters* différents sur le coût moyen et la fréquence pose la problématique de la méthodologie d'application du modèle croisé coût-fréquence. En effet, le modèle nécessite que pour chaque acte un lien soit fait entre son modèle de coût et son modèle de fréquence correspondant. Ainsi, la modélisation tarifaire ne peut se faire en l'état sur les *clusters* formés.

Pour résoudre cette problématique, plusieurs options ont été envisagées :

— **Option 1** : réaliser les modèles de coût et de fréquence distinctement sur les *clusters* de coût et ceux de fréquence. Ventiler ensuite les résultats obtenus entre les actes composant les différents *clusters* en fonction d'une clé de répartition à définir (poids de l'acte dans la sinistralité du *cluster* par exemple) avant d'effectuer le croisement des modèles. Cette possibilité a été écartée car elle n'est pas robuste et introduit du biais dans le processus de tarification. En effet, la clé de répartition n'est pas stable puisqu'elle dépend de la base considérée.

— **Option 2** : redéfinir des *clusters* qui coincideraient à l'intersection des *clusters* coût moyen et des *clusters* fréquence. L'objectif est d'obtenir des *clusters* composés d'actes homogènes tant sur les coûts que sur la fréquence. Cette option nous replace dans le même schéma de tarification que celui de la tarification 2. Elle a tout de même comme inconvénient de conduire à un nombre élevé de *clusters* finaux.

— **Option 3** : procéder à du *clustering* imbriqué. La procédure consiste à réaliser une première segmentation sur la fréquence (par exemple), puis à réaliser la segmentation du modèle de coût moyen sur les premiers *clusters* obtenus. Ainsi, chaque modèle de coût sur les *clusters* finaux peut être relié sans ambiguité à un modèle de fréquence pour réaliser les croisements. Cette option a également l'inconvénient d'accroître le nombre de modèles finaux à réaliser.

Dans le cadre de cette étude, c'est l'option 2 qui est retenue. Ainsi, l'intersection des *clusters* de coût et de fréquence obtenus précédemment conduit à considérer dix *clusters* représentés ci-après (les valeurs très élevées de coût moyen ont été tronquées) :

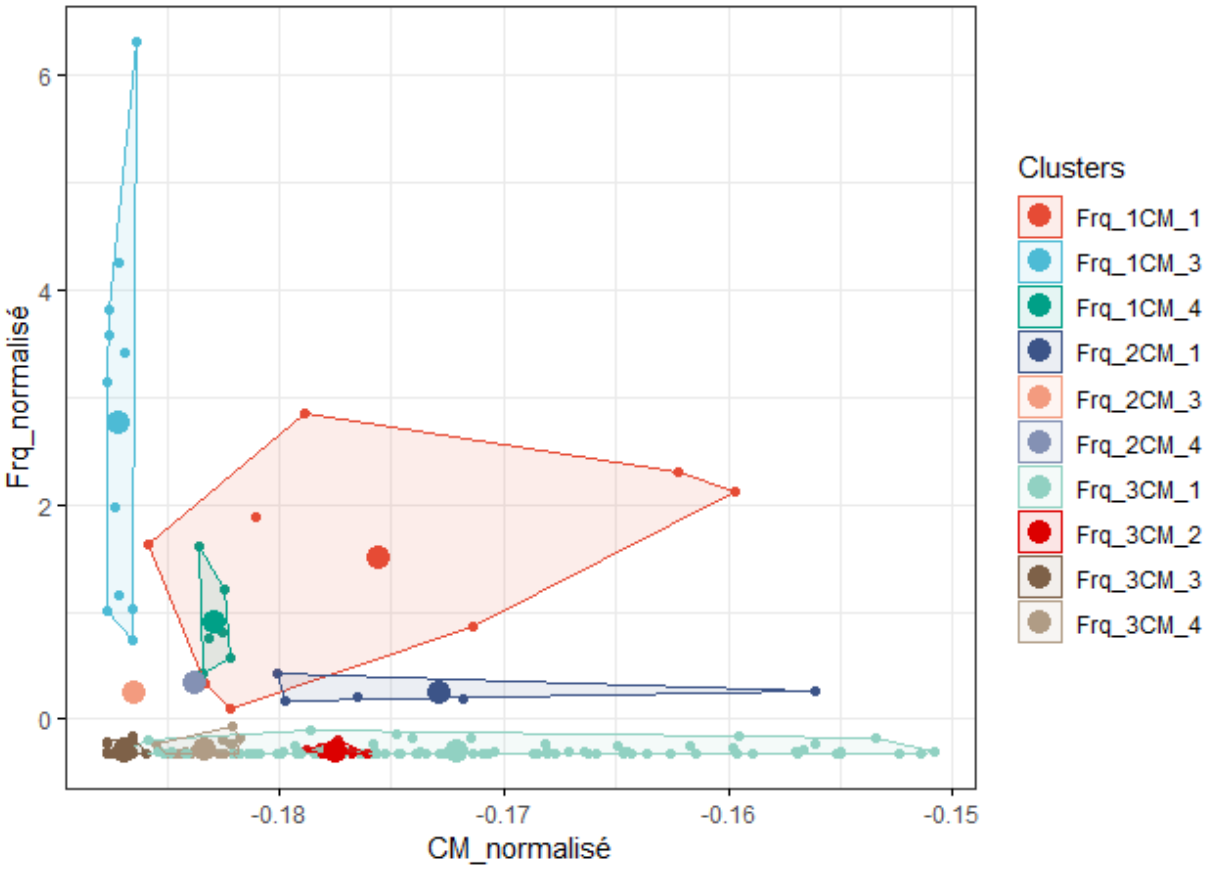


FIGURE 4.17 – Résultat croisement des *clusters* coût moyen et fréquence

Lecture : Frq1-CM3 représente l'intersection entre le *cluster* 1 du modèle de fréquence et le *cluster* 3 du modèle de coût. Pour rappel, les *clusters* 1 regroupent les points bruits des modèles.

Plusieurs *clusters* se chevauchent et certains ne contiennent qu'un seul acte. Afin de corriger ces constats, des fusions de *clusters* ont été réalisée. Ces fusions ont été menées tout en gardant l'objectif d'uniformité des caractéristiques de coût moyen et de fréquence.

La représentation des *clusters* finaux en découlant sont présentées ci-après (les valeurs très élevées de coût moyen ont été tronquées).

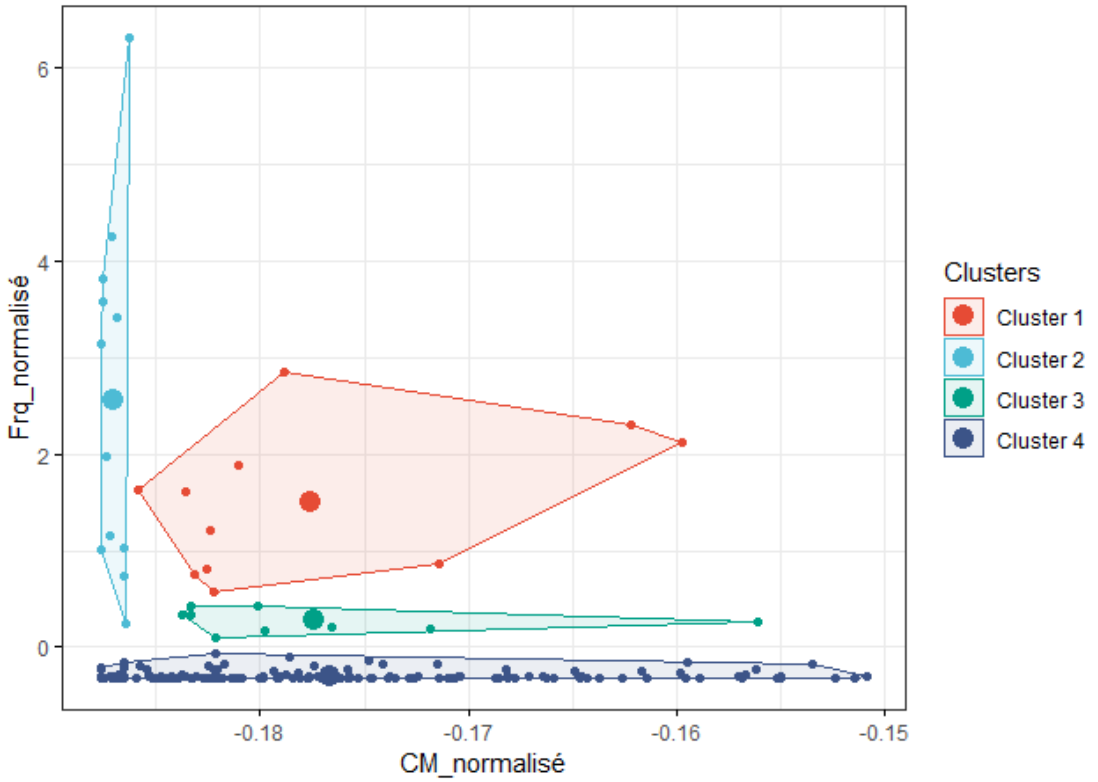


FIGURE 4.18 – *clusters* finaux de modélisation tarifaire

Comparativement à la segmentation assureur, cette segmentation met en exergue quatre profils d'actes :

— les actes avec un coût moyen très élevé et une fréquence élevée (*cluster* 1) ;
— les actes avec un coût moyen faible mais une fréquence élevée (*cluster* 2) ;
— les actes avec un coût moyen peu élevé mais une fréquence faible (*cluster* 3) ;
— les actes avec un coût moyen élevé mais une fréquence faible (*cluster* 4).

Les caractéristiques de chaque *cluster* sont précisées dans le tableau récapitulatif suivant (les indicateurs de sinistralité sont évalués sur la base sinistres annuelle) :

Clusters	Nombre d'actes	Poids dans la dépense totale	Poids dans le volume de sinistres	Coût moyen	Fréquence moyenne
Cluster 1	18	88,2%	47,9%	222,96	1,94
Cluster 2	12	0,9%	39,8%	2,60	1,61
Cluster 3	10	3,6%	6,7%	64,41	0,27
Cluster 4	234	7,4%	5,6%	159,03	0,23

FIGURE 4.19 – Caractéristiques *clusters* finaux de modélisation tarifaire

Les distributions de chaque *cluster* sur la population d'assurés ainsi que celles des lois de modélisation théoriques candidates ajustées aux données sont présentées ci-après. Le constat est que ces distributions ne sont pas très adaptées à l'application des GLM. Toutefois, la contrainte d'utilisation de ces modèles demeure afin de pourvoir réaliser une stricte comparaison avec les pratiques usuelles.

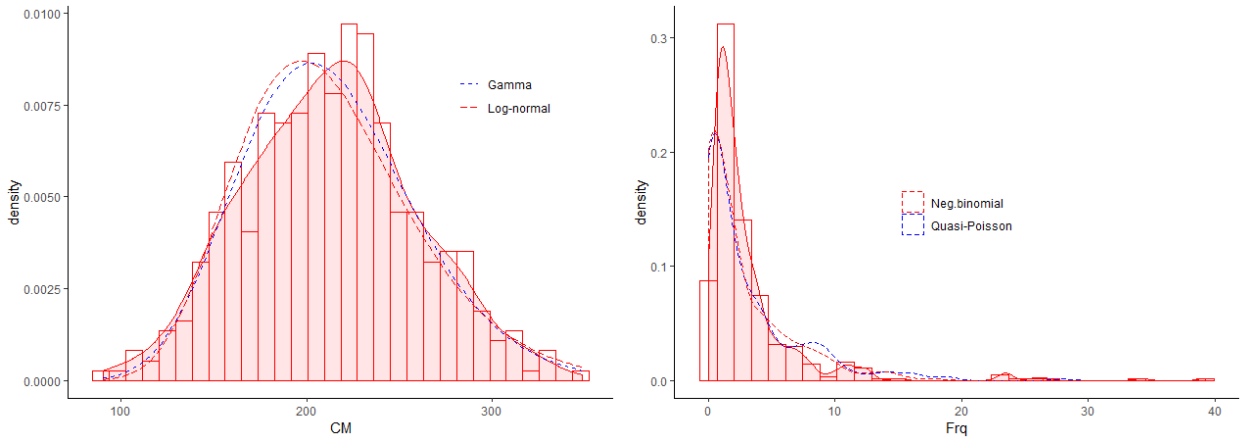


FIGURE 4.20 – Distribution coût moyen (CM) et fréquence (Frq) - *cluster* 1

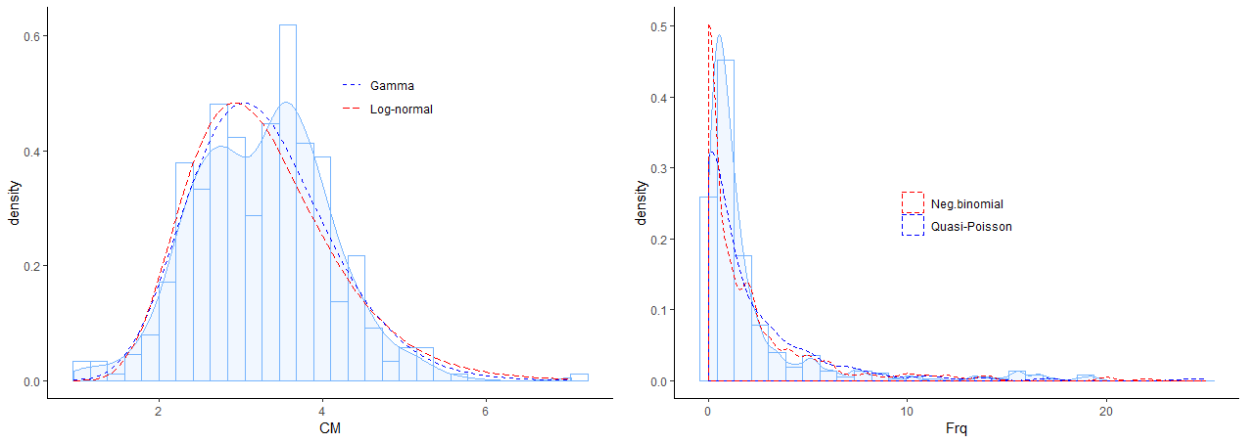


FIGURE 4.21 – Distribution coût moyen (CM) et fréquence (Frq) - *cluster* 2

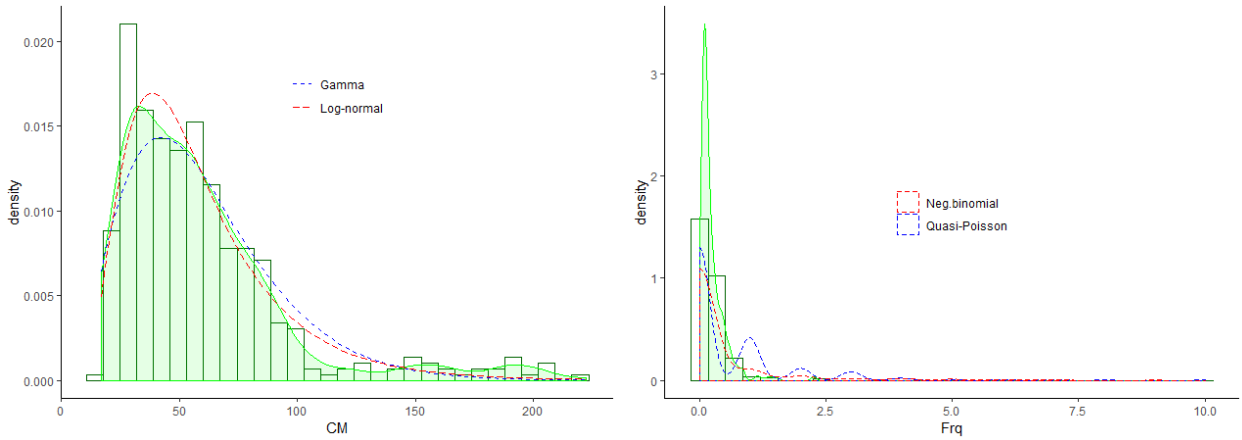


FIGURE 4.22 – Distribution coût moyen (CM) et fréquence (Frq) - *cluster* 3

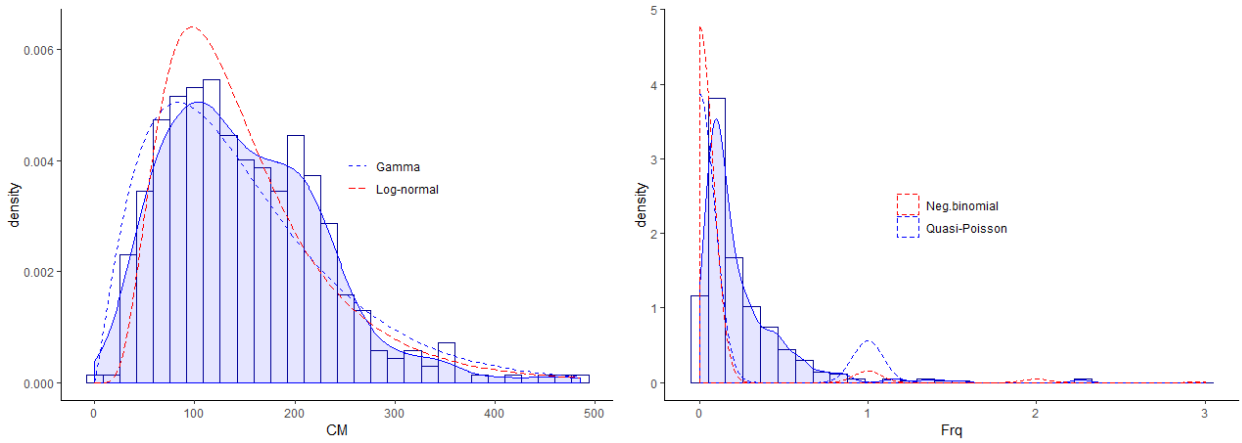


FIGURE 4.23 – Distribution coût moyen (CM) et fréquence (Frq) - *cluster* 4

4.3.4 Impact de la segmentation sur la qualité de la tarification

Une tarification a été réalisée à partir de la nouvelle segmentation obtenue afin d'évaluer son impact sur la qualité de prédiction de la dépense engagée.

Le processus de tarification appliqué a été le même que sur les deux premières tarifications. Ainsi, 4 modèles de coût et 4 modèles de fréquence ont été réalisés avant d'aboutir à 4 modèles croisés puis à l'estimation de la dépense totale. Les lois retenues pour chaque modèle et les résultats de prédictions en découlant sont détaillés ci-après.

Modèle de coût

A l'issue de la modélisation du coût moyen, le constat est que la nouvelle segmentation n'a pas permis d'aboutir à une amélioration des prédictions comparativement aux tarifications précédentes.

	Loi	Résidus centrés	Indépendance	homocédasticité	RMSE	Tarification 2	Tarification 1
Cluster 1	Gamma	OK	OK	OK	38,05		
Cluster 2	Gamma	OK	Non OK	OK	0,72		
Cluster 3	Log-normal	OK	OK	OK	21,05		
Cluster 4	Gamma	OK	OK	OK	91,83		
Modèle global					**38,43**	**28,70**	**26,81**

FIGURE 4.24 – Résultats prédictions tarification 3 - modèle de coût

Modèle de fréquence

Contrairement au modèle de coût, un gain en précision est enregistré : la RMSE est réduite de -13% par rapport à la tarification 2 et de -30% par rapport à la tarification 1.

	Loi	Résidus centrés	Indépendance	homocédasticité	RMSE	Tarification 2	Tarification 1
Cluster 1	Quasi-poisson	OK	OK	OK	6 393		
Cluster 2	Quasi-poisson	OK	OK	OK	21 524		
Cluster 3	Quasi-poisson	OK	OK	OK	1 907		
Cluster 4	Quasi-poisson	OK	OK	OK	1 519		
Modèle global					**11 797**	**13 536**	**16 957**

FIGURE 4.25 – Résultats prédictions tarification 3 - modèle de fréquence

Modèle coût-fréquence

La réalisation des différents croisements des modèles précédents permet, in fine, d'enregistrer une légère amélioration de la prédiction de la dépense engagée totale. Les niveaux prédits de dépenses et de primes pures sont précisés dans le tableau ci-après.

	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	Total tarification 3	Total tarification 2	Total tarification 3
Dépense totale prédite	4 876 107 396	47 763 514	196 527 594	412 719 351	5 533 117 855	5 539 092 764	5 538 699 125
Dépense totale observée	4 584 859 121	45 867 631	191 201 486	389 154 902	5 211 083 140	5 211 083 140	5 211 083 140
Ecart à l'observé	6,35%	4,13%	2,79%	6,06%	**6,181%**	**6,294%**	**6,287%**
Prime pure moyenne	436,81	4,28	17,61	36,98	**495,67**	**496,2**	**496,19**

FIGURE 4.26 – Résultats prédictions tarification 3 - modèle coût-fréquence

Si le gain global semble marginal, une vision plus fine par profil permet d'en comprendre tout l'apport. En effet, une analyse des ratio sinistres à primes (S/P) met en exergue une baisse radicale des profils sur lesquels il était enregistré des niveaux de S/P très élevés (S/P>200%) lors des tarifications précédentes. Le graphique suivant présente clairement ce phénomène avec une concentration plus forte des S/P de cette dernière tarfication autour de 100%.

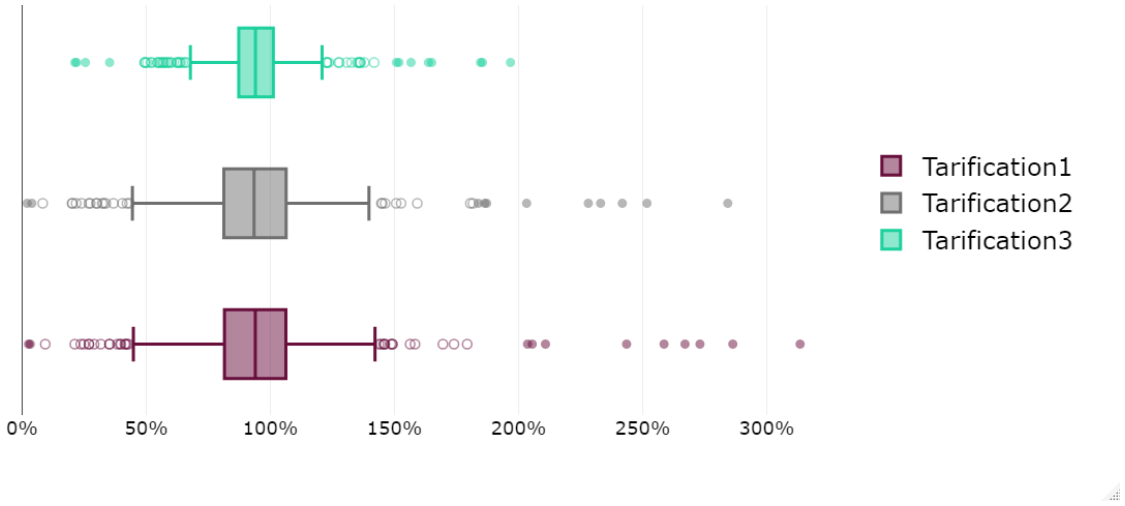


FIGURE 4.27 – Ratio sinistres à primes selon tarification

Ainsi, la nouvelle segmentation proposée permet de réduire l'écart moyen [2] des S/P [3] au niveau d'équilibre (100%) à 9,6 points contre 12,2 points et 12,7 points respectivement pour la tarification 2 et la tarification 1. De plus, le nombre de profils avec des S/P supérieurs à 100 passe de 37% à 28%, ce, avec des niveaux de dépassements moindres que dans les deux précédentes tarifications. Dans le cadre de ces dernières, les dépassements représentent plus de 30% de la dépense engagée contre 17% après application de la nouvelle segmentation proposée, soit une réduction de 45% de la perte enregistrée.

	Part profils avec S/P > 100%	Dépense engagée pour les profils avec S/P >100%		Ecart moyen des S/P à 100%	S/P maximal
		Total (en euro)	Part de la dépense engagée totale (%)		
Tarification 1	37,7%	1 847 894 409	35,46%	12,74%	313,44%
Tarification 2	37,0%	1 652 856 821	31,72%	12,18%	284,31%
Tarification 3	28,1%	908 104 601	17,43%	9,60%	196,74%

FIGURE 4.28 – Comparaison des résultats selon tarification

Ces constats sont particulièrement importants car ils touchent au pilotage technique de la sinistralité du portefeuille. En effet,cela assure une tarification plus juste selon les profils, évitant ainsi l'anti-sélection.

La méthodologie présentée au travers de cette troisième tarification permet d'améliorer l'adéquation technique du tarif, et ce en contrepartie d'un faible coût opérationnel de mise en place.

2. moyenne pondérée par l'exposition de la valeur absolue des écarts à 100%
3. dépense engagée observée / dépense engagée prédite

Conclusion

Le faible positionnement des complémentaires santé dans le cadre de la prise en charge du risque hospitalier pose la problématique de la connaissance réelle du risque couvert. Ainsi, la mise à disposition par la Sécurité sociale de données nationales du secteur privé sur cette garantie, a suscité un intérêt à évaluer l'optimalité des pratiques actuelles de tarification.

Ce mémoire s'est donc interessé à l'efficacité de la segmentation du poste hospitalier en sous-postes de soins pour réaliser la tarification.

Dans cette optique, sur la base des GLM et du modèle coût-fréquence, la tarification par sous-poste de soins a été comparée à celle modélisant tous les actes sans distinction. Il en ressort que la segmentation par sous-poste de soins, nécessitant la réalisation de six modèles, n'induit pas de plus-value, en matière de prédiction, par rapport à une approche directe en deux modèles.

Ce fut donc le lieu de s'affranchir de la logique médicale afin de construire une nouvelle segmentation des actes hospitaliers répondant à un objectif de création de groupes d'actes homogènes. Ainsi, sur la base de la moyenne et de l'écart-type, quatre nouveaux groupes d'actes ont été crées par l'intermédiaire de l'algorithme HDBSCAN.

La tarification selon ces nouveaux groupes permet de réduire de -25% la volatilité du résultat par rapport à l'adéquation technique. En effet, l'écart moyen (tous profils confondus) des S/P à l'équilibre est passé de 12,2 points, dans le cadre de la tarification par sous-poste de soins, à 9,6 points. Par ailleurs, cette nouvelle segmentation permet une réduction de 45% de la dépense engagée liée aux profils sous-tarifés (S/P>100%), soit une économie de 745 millons d'euros.

Bien que satisfaisants, ces résultats souffrent toutefois de quelques limites :

— le niveau de granularité des données (base en *model points*), couplé aux faibles nombre de variables disponibles a indéniablement engendré une perte d'informations individuelles donc de précision dans les travaux ;

— les distributions de fréquence et de coût moyen observées sur cette garantie ne se prêtent pas toutes à l'utilisation d'un modèle paramétrique tel que le GLM.

Ainsi, l'application de cette méthodologie sur des bases non aggrégées avec des modèles tarifaires plus adaptés, pourrait siginificativement améliorer les premiers résultats obtenus. En outre, l'introduction de variables d'adéquation aux lois usuelles comme variables de segmentation pourrait être une réponse à la contrainte d'utilisation des GLM. Enfin, les performances de l'algorithme pourraient être améliorées en initiant une évaluation de la distance optimale à utiliser, afin de challenger la distance euclidienne implémentée par défaut dans HDBSCAN.

Table des figures

Annexe A

Liste des actes hospitaliers retenus pour l'étude

Code actes (Open DAMIR)	Sous-poste de soins	Libellé	Code actes (Open DAMIR)	Sous-poste de soins	Libellé
1096	FSEJ	TELECONSULTATION MEDECIN TRAITANT AVEC EHPAD	1121	FSEJ	HONORAIRE DE SURVEILLANCE
1098	FSEJ	CONSULTATION CCMU 3	1122	FSEJ	EXAMEN SPECIAL (PROTOCOLE)
1099	FSEJ	CONSULTATION CCMU 4 ET 5	1123	FSEJ	SUITE D EXAMEN DE SANTE
1101	AUTR	AVIS PONCTUEL DE CONSULTANT PUPH	1125	FSEJ	MAJORATION DE COORDINATION DES GENERALISTES
1102	AUTR	AVIS PONCTUEL DE CONSULTANT PSYCHIATRE	1126	FSEJ	MAJORATION DE COORDINATION SPECIALISTES
1103	AUTR	AVIS PONCTUEL DE CONSULTANT DU MEDECIN	1127	FSEJ	MAJORATION DE COORDINATION CARDIOLOGUES
1104	AUTR	CONSULTATION OBLIGATOIRE ENFANT	1129	FSEJ	MAJORATION FORFAITAIRE TRANSITOIRE (POUR LES MOINS DE 16 ANS)
1105	FSEJ	CONSULTATION COMPLEXE	1130	FSEJ	FORFAIT MEDECIN TRAITANT
1106	FSEJ	MAJORATION CONSULTATION COMPLEXE	1132	AUTR	MAJORATION CONSULTATION ENDOCRINO
1107	FSEJ	CONSULTATION TRES COMPLEXE ENFANT	1133	FSEJ	MAJORATION GENERALISTE ENFANT
1108	FSEJ	MAJORATION CONSULTATION TRES COMPLEXE	1140	AUTR	CONSULTATION SPECIFIQUE DE DEPISTAGE
1109	AUTR	CONSULTATION SPECIALISTE MEDECINE GENERALE	1141	AUTR	MAJORATION PEDIATRE ENFANT
1110	AUTR	CONSULTATION MEDECINE GENERALE	1148	AUTR	REMUNERATION FORFAITAIRE PAR CONSULTATION POUR LE SUIVI DES PERSONNES AGEES
1111	FSEJ	CONSULTATION COTEE C	1149	CHBR	CONTESTATION REMUNERATION SUIVI PERSONNES AGEES
1112	FSEJ	CONSULTATION COTEE CS	1158	FSEJ	Acte de télé expertise
1113	FSEJ	CONSULTATION COTEE CNP	1164	FSEJ	TÉLÉ CONSULTATION - ALD ET / OU EHPAD
1114	FSEJ	CONSULTATION SPECIFIQUE CARDIOLOGIE	1165	FSEJ	TÉLÉ EXPERTISE - ALD ET/OU EHPAD
1116	FSEJ	MAJORATION FORFAITAIRE TRANSITOIRE	1169	AUTR	REMUNERATION POUR CERTIFICAT DE DECES
1117	FSEJ	CONSULTATION DES SPECIALISTES COTEE C2	1170	FSEJ	PEC EXCEPTIONNELLE DÉPASSEMENT HONORAIRE
1118	FSEJ	CONSULTATION DES PSYCHIATRES COTEE C2,5	1171	FSEJ	RÉMUNÉRATION MÉDECIN TRAITANT CENTRES DE SANTÉ

Code actes (Open DAMIR)	Sous-poste de soins	Libellé	Code actes (Open DAMIR)	Sous-poste de soins	Libellé
1172	FSEJ	TELESURVEILLANCE : PS EFFECTUANT L'ACCOMPAGNEMENT	1331	FSEJ	ACTES DE RADIOLOGIE
1173	FSEJ	FORFAIT PATIENTELE MEDECIN TRAITANT	1361	FSEJ	VIDEOCAPSULE
1174	FSEJ	TELESURVEILLANCE : MEDECIN TELESURVEILLANT	1431	FSEJ	ACTES EN D (ET OCC POUR LA CRPCEN)
1175	FSEJ	PEC EXCEPTIONNELLE DEPASSEMENT HONORAIRE TP	1437	AUTR	MAJORATION SPÉCIFIQUE PDS CLINIQUE DENTISTE
1191	FSEJ	TELECONSULTATION TOUTES SPECIALITES	1522	FSEJ	MAJORATION ASTREINTE
1192	FSEJ	TELECONSULTATION GENERALISTE	1811	AUTR	IK PLAINE
1193	FSEJ	TELE EXPERTISE DE NIVEAU 1	1812	AUTR	IK MONTAGNE
1194	FSEJ	TELE EXPERTISE DE NIVEAU 2	1813	AUTR	IK PIED SKI
1210	AUTR	VISITE MEDECINE GENERALE	1821	FSEJ	ID PARIS LYON MARSEILLE, +100.000 HA, -100.000 HA
1211	FSEJ	VISITE COTEE V	1841	FSEJ	INDEMNITES FORFAITAIRES DE DEPLACEMENT
1222	FSEJ	VISITE URGENCE VU/MU	1842	FSEJ	INDEMNITES FORFAITAIRES DE DEPLACEMENT DES AUXILIAIRES MEDICAUX ET ASSIMILES
1224	FSEJ	MD (CRITERES MEDICAUX)	1903	AUTR	MAJORATION ENFANT GENERALISTE
1226	FSEJ	MD DE NUIT	1904	AUTR	MAJORATION ENFANT PEDIATRE
1228	FSEJ	MD DE DIMANCHE ET JOUR FERIES	1905	AUTR	NOUVEAU FORFAIT ENFANT
1311	FSEJ	ACTES EN K CHIRURGICAL	1906	FSEJ	NOUVEAU FORFAIT PEDIATRIQUE
1312	FSEJ	ACTES DE SPECIALITE EN K	1907	FSEJ	REMUNERATION DES SOINS DE PROXIMITE
1316	FSEJ	ACTES DE DIAGNOSTIC COTES KE	1910	AUTR	PLAN PERSONNALISÉ DE SANTÉ
1322	FSEJ	ACTE D'OBSTETRIQUE CCAM	1911	FSEJ	ACTES DES SAGES-FEMMES
1323	FSEJ	ACTE D'ANESTHESIE CCAM	1912	FSEJ	HONORAIRES NON VENTILABLES INDIVIDUALISES
1324	FSEJ	ACTE D'ECHOGRAPHIE CCAM	1913	FSEJ	MAJORATION MILIEU DE NUIT

FIGURE A.1 – Liste des actes hospitaliers retenus pour l'étude (1/4)

Code actes (Open DAMIR)	Sous-poste de soins	Libellé		Code actes (Open DAMIR)	Sous-poste de soins	Libellé
1914	AUTR	FORFAIT PEDIATRIQUE		1981	FSEJ	FORFAIT IVG HONORAIRES DE VILLE
1918	FSEJ	MAJORATION D'URGENCE		1990	FSEJ	FORFAIT D'INTERVENTION PAR SORTIE SUR DEMANDE DE LA REGULATION
1931	FSEJ	MAJORATION NOURRISSON PEDIATRE		1991	FSEJ	REMUNERATION REGULATION
1932	FSEJ	MAJORATION NOURRISSON GENERALISTE		1992	FSEJ	PERMANENCE REMUNERATION DE NUIT
1933	AUTR	MAJORATION CONSULTATION REGULEE DE NUIT		1993	FSEJ	PERMANENCE REMUNERATION MILIEU DE NUIT
1934	AUTR	MAJORATION CONSULTATION REGULEE MILIEU DE NUIT		1994	FSEJ	PERMANENCE REMUNERATION DIMANCHE ET FERIE
1935	FSEJ	MAJORATION CONSULTATION REGULEE DIMANCHE, JOURS FERIES ET ASSIMILES		1995	FSEJ	PERMANENCE REMUNERATION TOTAL
1939	FSEJ	MAJORATION SAGE-FEMME		1996	FSEJ	PERMANENCE REMUNERATION SAMEDI MATIN
1943	FSEJ	MAJORATION URGENCE MT		1997	FSEJ	PERMANENCE REMUNERATION SAMEDI APRES MIDI
1944	FSEJ	MAJORATION CORRESPONDANT URGENCE		1998	FSEJ	ASTREINTE DE JOUR CORRESPONDANT SAMU
1945	FSEJ	MAJORATION MEDECIN TRAITANT REGULATION		1999	FSEJ	ASTREINTE DE NUIT CORRESPONDANT SAMU
1951	FSEJ	PARTICIPATION FORFAITAIRE HORS TIERS PAYANT		2103	FSEJ	SUPPLEMENT TRANSPORT PERMISSION THERAPEUTIQUE (PSY/SSR)
1952	FSEJ	PARTICIPATION FORFAITAIRE TIERS PAYANT		2104	FSEJ	SUPPLEMENT TRANSPORT PROVISOIRE (PSY/SSR)
1954	FSEJ	PARTICIPATION ASSURE CONSULTATIONS ET SOINS EXTERNES (CMU + AME)		2105	FSEJ	SUPPLEMENT TRANSPORT DEFINITIF (PSY/SSR)
1956	FSEJ	PARTICIPATION ASSURE EN AMBULATOIRE		2106	FSEJ	TRANSPORT DEFINITIF DIALYSE
1957	FSEJ	MAJORATION HORS PARCOURS DE SOINS		2107	FSEJ	TRANSPORT SEANCE DIALYSE
1961	FSEJ	SUPPLEMENT DEROGATOIRE SG SUR ACTE PROFESSIONNEL REMBOURSABLE (CNMSS)		2108	FSEJ	SUPPLEMENT TRANSPORT 2
1974	FSEJ	FRANCHISE TIERS PAYANT SUR TRANSPORT		2109	FSEJ	SUPPLEMENT TRANSPORT SEANCES
1975	FSEJ	FRANCHISE HORS TIERS PAYANT ACTE D'AUXILIAIRE MEDICAUX		2111	FSEJ	FRAIS D HEBERGEMENT ET ENVIRONNEMENT EN GHS
1976	FSEJ	FRANCHISE TIERS PAYANT ACTE D'AUXILIAIRE MEDICAUX		2112	FSEJ	FRAIS DE SEJOUR SUPPLEMENTAIRE AU GHS

Code actes (Open DAMIR)	Sous-poste de soins	Libellé		Code actes (Open DAMIR)	Sous-poste de soins	Libellé
2113	FSEJ	GROUPE HOMOGENE DE TARIFS		2151	FSEJ	SUPPLEMENT REANIMATION
2116	FSEJ	SUPPLEMENT NEONATOLOGIE 1		2152	FSEJ	SUPPLEMENT SURVEILLANCE CONTINUE
2117	FSEJ	SUPPLEMENT NEONATOLOGIE 2		2153	FSEJ	FORFAIT SOINS INTENSIFS
2119	FSEJ	SUPPLEMENT DEFIBRILLATEUR		2155	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 1
2120	FSEJ	DIFFERENTIEL TARIFAIRE CLINIQUE		2156	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 2
2132	FSEJ	FORFAIT D HEMODIALYSE EN UNITE DE DIALYSE MEDICALISEE		2157	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 3
2133	FSEJ	DIALYSE TIERCE PERSONNE		2158	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 4
2134	FSEJ	FORFAIT D AUTODIALYSE SIMPLE		2159	FSEJ	FORFAIT DE SECURITE DERMATOLOGIQUE
2135	FSEJ	FORFAIT D AUTODIALYSE ASSISTEE		2163	FSEJ	SUPPLEMENT JOURNALIER DIALYSE PERITONEALE
2136	FSEJ	FORFAIT D HEMODIALYSE A DOMICILE		2164	FSEJ	ADMINISTRATION DE PRODUITS ET PRESTATIONS EN ENVIRONNEMENT HOSPITALIER
2137	FSEJ	FORFAIT DE DIALYSE PERITONEALE AUTOMATISEE (DPA)		2165	FSEJ	FORFAIT PRESTATION INTERMEDIAIRE
2138	FSEJ	FORFAIT DE DIALYSE PERITONEALE CONTINUE AMBULATOIRE (DPCA)		2167	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 5
2141	CHBR	FRAIS DE CHAMBRE PARTICULIERE POUR CONVENANCE PERSONNELLE		2168	FSEJ	FORFAIT ENVIRONNEMENT HOSPITALIER 6
2142	FSEJ	FORFAIT D ENTRAINEMENT A DIALYSE PERITONEALE CONTINUE AMBULATOIRE		2173	FSEJ	DIFFERENTIEL PSY REGLEMENTAIRE
2143	FSEJ	FF D ENTRAINEMENT A LA DIALYSE PERITONEALE AUTOMATISEE A DOMICILE		2181	FSEJ	PRELEVEMENT D ORGANE 1
2144	FSEJ	FF D ENTRAINEMENT A LA DIALYSE PERITONEALE CONTINUE AMBULATOIRE A DOMICILE		2182	FSEJ	PRELEVEMENT D ORGANE 2
2145	FSEJ	FORFAIT DE DIALYSE PERITONEALE AUTOMATISE POUR HOSPITALISATION DE 3 A 6 JOURS		2184	FSEJ	COORDINATION PRELEVEMENT D ORGANES
2146	FSEJ	FORFAIT DE DIALYSE PERITONEALE CONTINUE AMBULATOIRE POUR HOSPITALISATION DE 3 A 6 JOURS		2186	FSEJ	PRELEVEMENT D'ORGANE 5
2147	FSEJ	FORFAIT D'ENTRAINEMENT A L'HEMODIALYSE EN UNITE DE DIALYSE MEDICALISEE		2187	FSEJ	PRELEVEMENT D'ORGANE 6
2150	FSEJ	VIDEOCAPSULE		2188	FSEJ	PRELEVEMENT D'ORGANE 7

FIGURE A.2 – Liste des actes hospitaliers retenus pour l'étude (2/4)

Code actes (Open DAMIR)	Sous-poste de soins	Libellé
2189	FSEJ	PRELEVEMENT D'ORGANE 8
2190	FSEJ	PRELEVEMENT D'ORGANE 9
2191	FSEJ	PRELEVEMENT D'ORGANE
2211	FSEJ	FRAIS DE SEJOUR
2212	FSEJ	MAJORATION PMSI
2213	FSEJ	FRAIS DE SEJOUR IME
2215	FSEJ	SUPPLEMENT DEROGATOIRE SG SUR PRESTATION SEJOUR REMBOURSABLE (CNMSS)
2221	CHBR	SUPPLEMENT CHAMBRE PARTICULIERE
2222	FSEJ	SUPPLEMENT POUR SURVEILLANCE DU MALADE
2231	FSEJ	FORFAIT PHARMACEUTIQUE
2234	FSEJ	FORFAIT D ENTREE
2237	FSEJ	PART COMPLEMENTAIRE AIDE MEDICALE ETAT (REGULARISATION CMU COMPLEMENTAIRE)
2238	FSEJ	FORFAIT D ACCUEIL ET DE TRAITEMENT DES URGENCES
2241	FSEJ	FRAIS DE SALLE D OPERATION
2242	FSEJ	FRAIS D ANESTHESIE ET REANIMATION
2247	FSEJ	FORFAIT PSYCHIATRIE DE SECURITE - HOSPITALISATION AVEC HEBERGEMENT
2250	FSEJ	FORFAIT JOURNALIER AIDE MEDICALE (REGULARISATION CMU COMPLEMENTAIRE)
2251	FSEJ	FORFAIT JOURNALIER
2252	FSEJ	FORFAIT JOURNALIER DE SORTIE
2258	FSEJ	FORFAIT DE SOINS JOURNALIER
2259	AUTR	SAISIE MANUELLE DES SEJOURS POUR LEQUELS LE FJ EST SUPERIEUR AU TM
2260	AUTR	FORFAIT HOPITAL PROXIMITE COMPLEMENTAIRE
2282	FSEJ	PARTICIPATION ASSURE TRANSITOIRE
2283	FSEJ	PARTICIPATION ASSURE HOSPITALISATION PUBLIQUE (CMU + AME)
2284	FSEJ	PARTICIPATION ASSURE HOSPITALISATION PUBLIQUE (REGIME LOCAL)
2285	FSEJ	PARTICIPATION ASSURE SUR SEJOUR
2321	FSEJ	FORFAIT - LONG SEJOUR PERSONNES AGEES
2331	FSEJ	RADIUMTHERAPIE ET CHIMIOTHERAPIE
2332	FSEJ	READAPTATION FONCTIONNELLE
2333	FSEJ	REEDUCATION PROFESSIONNELLE
2334	FSEJ	SEANCE D HEMODIALYSE
2336	FSEJ	FORFAIT POUR CONSULTATION EN CENTRE MEDICO-PSYCHO PEDAGOGIQUE
2337	FSEJ	SEANCE DE DIAGNOSTIC
2338	FSEJ	FORFAIT PETIT MATERIEL
2339	FSEJ	AUTRES FORFAITS DIVERS (Y COMPRIS NUTRITION ENTERALE A DOMICILE)
2342	FSEJ	FORFAIT POUR GARDE DE DEBUT DE NUIT EN ETABLIS. PRIVE
2343	FSEJ	FORFAIT POUR GARDE DE NUIT OU SAMEDI APRES MIDI EN ETABLIS. PRIVE
2344	FSEJ	FORFAIT DE GARDE NUIT ET FERIE EN ETABLIS. PRIVE
2345	FSEJ	FORFAIT POUR ASTREINTE DE DEBUT DE NUIT EN ETABLIS. PRIVE
2346	FSEJ	FORFAIT POUR ASTREINTE DE NUIT OU SAMEDI APRES MIDI EN ETABLIS. PRIVE

Code actes (Open DAMIR)	Sous-poste de soins	Libellé
2347	FSEJ	FORFAIT D'ASTREINTE NUIT ET FERIE EN ETABLIS. PRIVE
2351	FSEJ	FORFAIT TECHNIQUE NORMAL IRMN -SCANNERS
2352	FSEJ	FORFAIT TECHNIQUE REDUIT IRMN -SCANNERS
2353	FSEJ	FORFAIT TECHNIQUE SCANNER (SPP expo amiante)
2354	FSEJ	FORFAIT TECHNIQUE TOMOGRAPHIE
2380	FSEJ	FORFAIT PSYCHIATRIE SEANCE COLL, 1 INTERVENANT 3 à 4H
2381	FSEJ	FORFAIT PSYCHIATRIE SEANCE IND, 1 INTERVENANT 3 à 4H
2382	FSEJ	FORFAIT PSYCHIATRIE SEANCE COLL, 2 INTERVENANTS 3 à 4H
2383	FSEJ	FORFAIT PSYCHIATRIE SEANCE IND. 2 INTERVENANTS 3 à 4H
2384	FSEJ	FORFAIT PSYCHIATRIE SEANCE COLL, 1 INTERVENANT 6 à 8H
2385	FSEJ	FORFAIT PSYCHIATRIE SEANCE IND. 1 INTERVENANT 6 à 8H
2386	FSEJ	FORFAIT PSYCHIATRIE SEANCE COLL, 2 INTERVENANT 6 à 8H
2387	FSEJ	FORFAIT PSYCHIATRIE SEANCE IND. 2 INTERVENANTS 6 à 8H
2388	FSEJ	FORFAIT PSYCHIATRIE DE SECURITE HOSPITALISATION SANS HEBERGEMENT
2389	FSEJ	PRISE EN CHARGE DE NUIT POUR UNE DUREE ENTRE 8 ET 12H
2391	FSEJ	FORFAIT TECHNIQUE TARIF REDUIT N°2
2392	FSEJ	FORFAIT TECHNIQUE TARIF REDUIT N°3
2411	FSEJ	INTERVENTION IVG
2412	FSEJ	ANESTHESIE GENERALE
2413	FSEJ	INVESTIGATIONS BIOLOGIQUES
2417	FSEJ	VERIFICATION BIOLOGIQUE
2418	FSEJ	VERIFICATION ECHOGRAPHIQUE
2419	FSEJ	FORFAIT INTERVENTION AMBULATOIRE
2420	FSEJ	FORFAIT INTERVENTION DUREE < OU = 12 H PRIVE MEDIC
2421	FSEJ	INTERVENTION + ANESTHESIE AMBULATOIRE
2422	FSEJ	FORFAIT POUR IVG MEDICAMENTEUSE
2423	FSEJ	FORFAIT INTERVENTION AVEC NUITEE
2428	FSEJ	ECHO PRE IVG
3110	AUTR	CONTRAT INCITATIF INFIRMIER
3111	FSEJ	ACTES EN AMI
3113	FSEJ	ACTES INFIRMIERS DES SAGES-FEMMES (SFI)
3115	AUTR	DEMARCHE INFIRMIER
3116	FSEJ	MAJORATION POUR ACTE UNIQUE
3121	FSEJ	ACTES AMC
3122	FSEJ	ACTES EN AMK
3125	FSEJ	ACTES DE KINESITHERAPIE OSTEO-ARTICULAIRE
3132	FSEJ	ACTES DES ORTHOPHONISTES
3133	FSEJ	ACTES DES ORTHOPTISTES
3211	FSEJ	ACTES DE BIOLOGIE
3213	FSEJ	FORFAIT PREALABLE BIOLOGIE IVG VILLE

FIGURE A.3 – Liste des actes hospitaliers retenus pour l'étude (3/4)

Code actes (Open DAMIR)	Sous-poste de soins	Libellé		Code actes (Open DAMIR)	Sous-poste de soins	Libellé
3216	FSEJ	FORFAIT ULTERIEUR BIOLOGIE IVG VILLE		3511	FSEJ	APPAREILS D ASSISTANCE RESPIRATOIRE,OXYGENOTHERAPIE A DOMICILE
3221	FSEJ	PRELEVEMENT AUTRE QUE SANGUIN PAR UN DIRECTEUR DE LABORATOIRE		3512	FSEJ	AUTRES MATERIELS POUR TRAITEMENTS A DOMICILE (CHAP. 1)
3222	FSEJ	PRELEVEMENT SANGUIN PAR UN DIRECTEUR DE LABORATOIRE		3513	FSEJ	MATERIELS ET APPAREILS DE CONTENTION ET DE MAINTIEN (CHAP. 2)
3223	FSEJ	PRELEVEMENT SANGUIN PAR UN TECHNICIEN DE LABORATOIRE		3514	FSEJ	MATERIELS ET APPAREILS POUR TRAITEMENTS DIVERS (CHAP. 3)
3225	FSEJ	PRELEVEMENT PAR PONCTION VEINEUSE DIRECTE POUR UN MEDECIN BIOLOGISTE		3515	FSEJ	ARTICLES DE PANSEMENTS (CHAP. 4)
3311	AUTR	PHARMACIE 100%		3517	FSEJ	ALIMENTS DESTINES A DES FINS MEDICALES
3315	FSEJ	MEDICAMENTS ANTIRETROVIRAUX		3521	FSEJ	ORTHESES (PETIT APPAREILLAGE) (CHAP. 1)
3318	FSEJ	PHARMACIE HOSPITALIERE A 65%		3522	FSEJ	DIVERS ORTHESES
3320	FSEJ	PHARMACIE HOSPITALIERE EN SUS DU GHS		3549	FSEJ	PROCESSEUR POUR IMPLANT OSTE-INTEGRE
3328	AUTR	MEDICAMENTS HOMEOPATHIQUES UNITAIRES		3551	FSEJ	IMPLANT INTERNE (CHAP. 1, 2 ET 3)
3329	FSEJ	FORFAIT MEDICAMENT IVG VILLE		3552	FSEJ	IMPLANT MU PAR ELECTRICITE (CHAP. 4)
3330	AUTR	ECART INDEMNISABLE RETROCESSION		3572	AUTR	ECART TIPS INDEMNISABLE
3331	FSEJ	VACCIN ANTI-GRIPPE		3574	FSEJ	DISPOSITIF MEDICAL (PRISE EN CHARGE EXCEPTIONNELLE)
3332	FSEJ	VACCIN ROR		3594	FSEJ	PEC EXCEPTIONNELLE DEPASSEMENT LPP TP
3337	FSEJ	FORFAIT FAUSSE COUCHE VILLE				
3338	FSEJ	FORFAIT FAUSSE COUCHE VILLE SANS ÉCHOGRAPHIE				
3339	FSEJ	FORFAIT FAUSSE COUCHE ETABLISSEMENT AVEC ECHOGRAPHIE				
3340	FSEJ	FORFAIT FAUSSE COUCHE ETABLISSEMENT SANS ECHOGRAPHIE				
3341	FSEJ	PHARMACIE 15%				
3386	AUTR	HONO DISP 7				

FIGURE A.4 – Liste des actes hospitaliers retenus pour l'étude (4/4)

Annexe B

Tests d'adéquation graphiques - modèle de coût - Tarification par sous-poste de soins

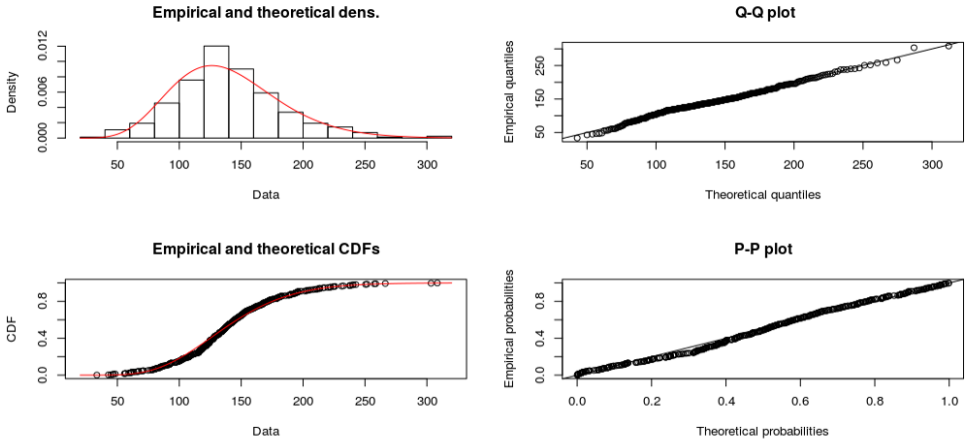


FIGURE B.1 – Adéquation loi Gamma - FSEJ

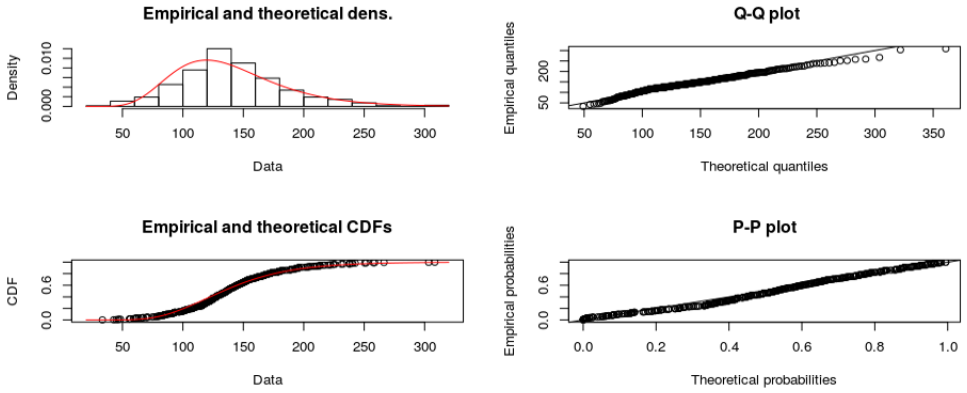


FIGURE B.2 – Adéquation loi log-normal - FSEJ

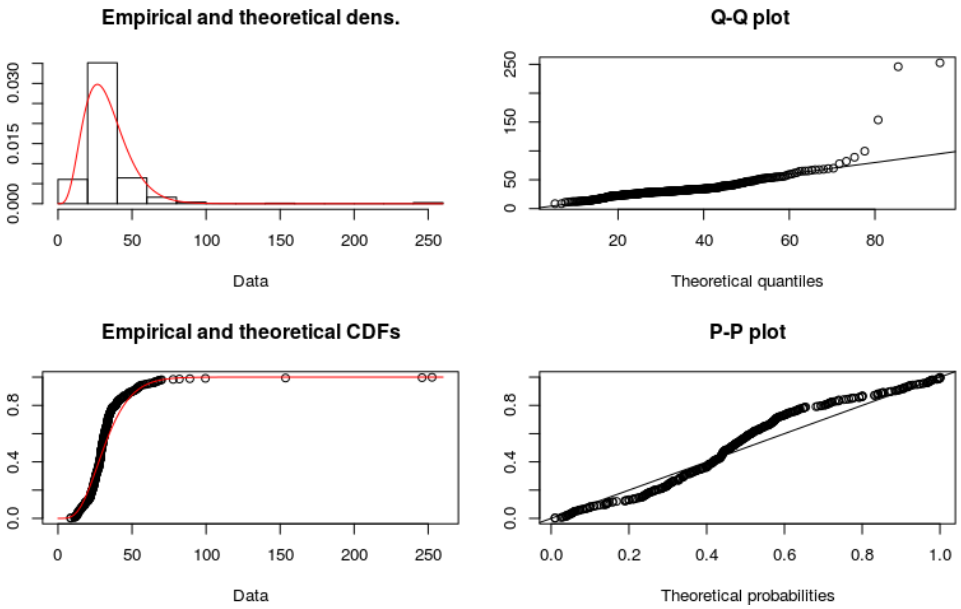


FIGURE B.3 – Adéquation loi Gamma - CHBR

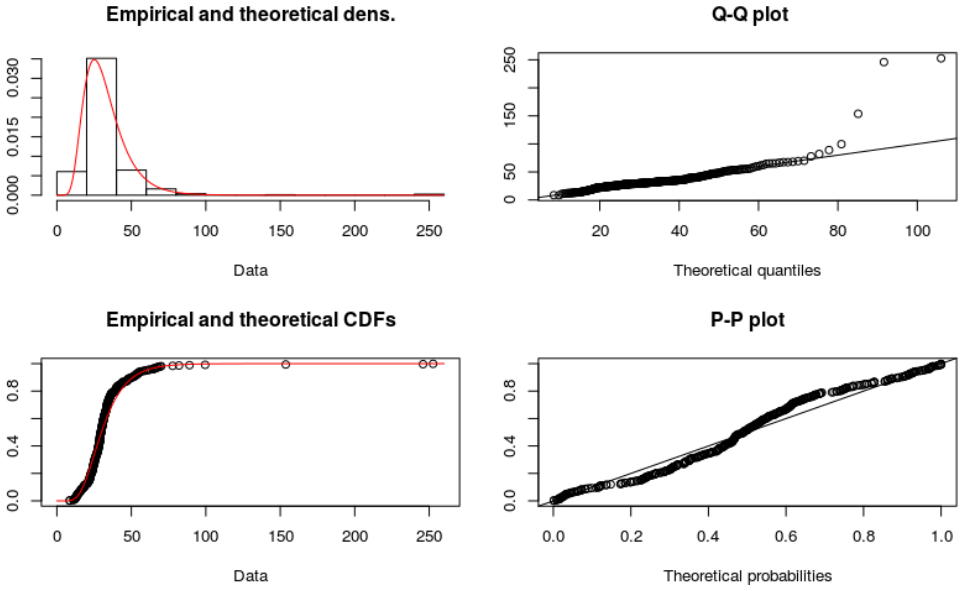


FIGURE B.4 – Adéquation loi log-normal - CHBR

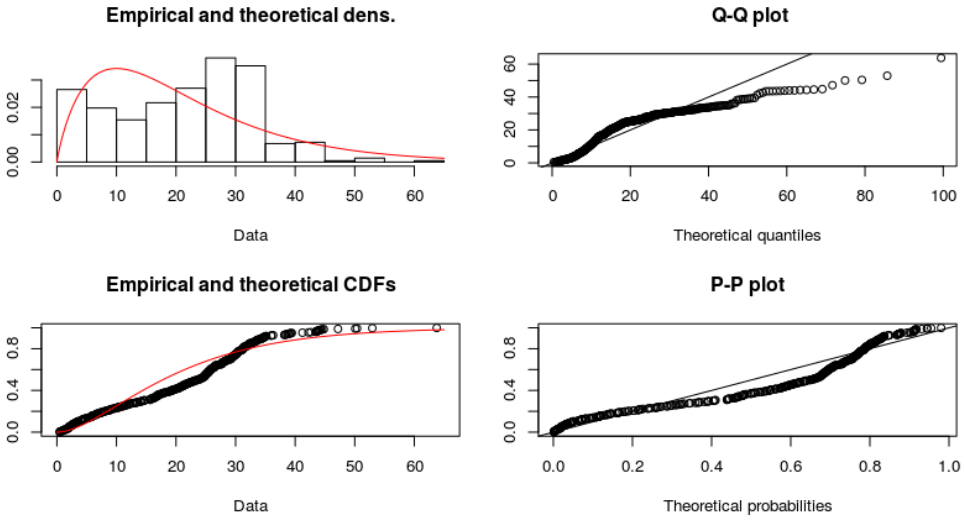


FIGURE B.5 – Adéquation loi Gamma - AUTR

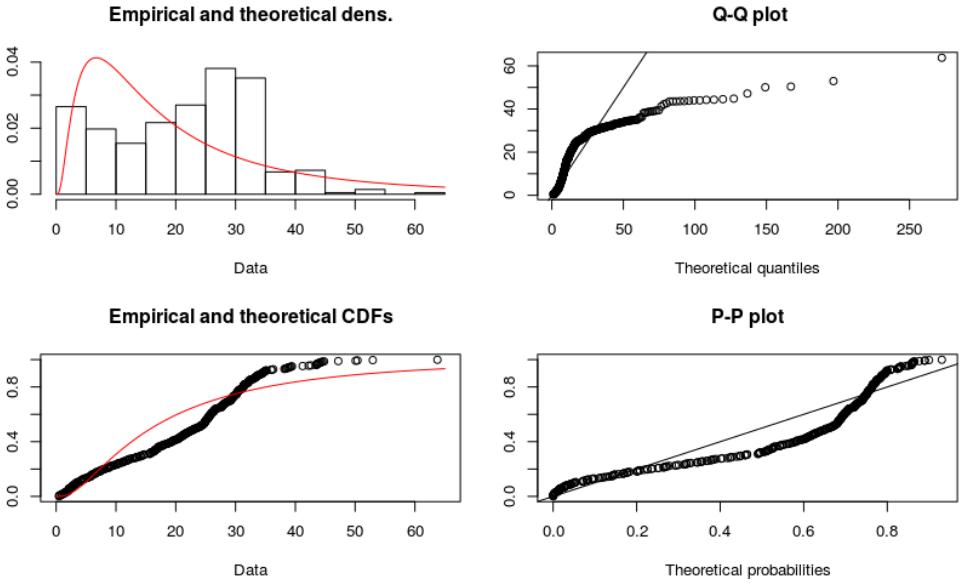


FIGURE B.6 – Adéquation loi log-normal - AUTR

Annexe C

Tests d'adéquation graphiques - modèle de coût - Tarification HDBSCAN

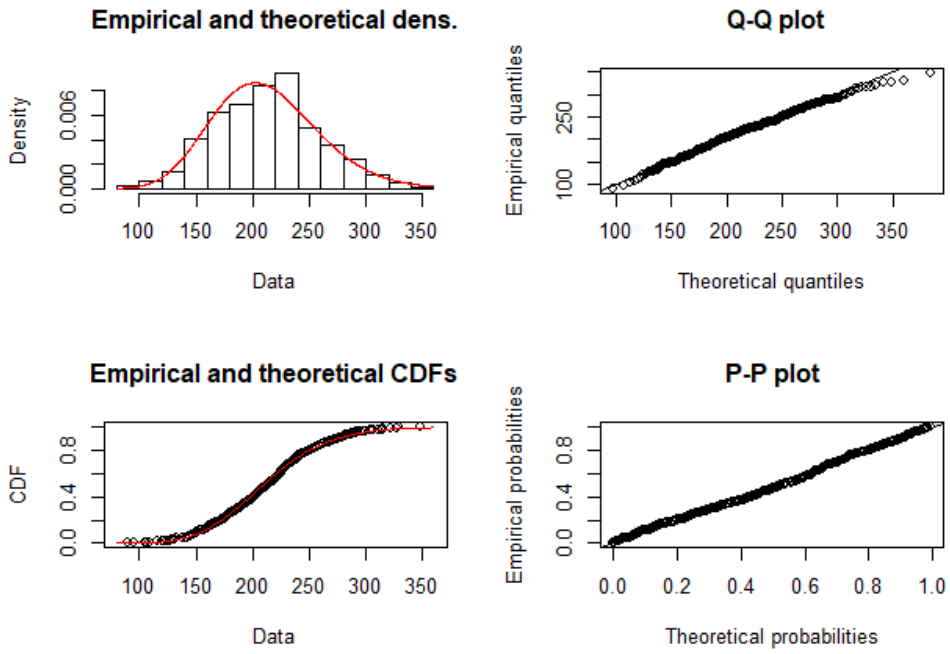


FIGURE C.1 – Adéquation loi Gamma - Cluster 1

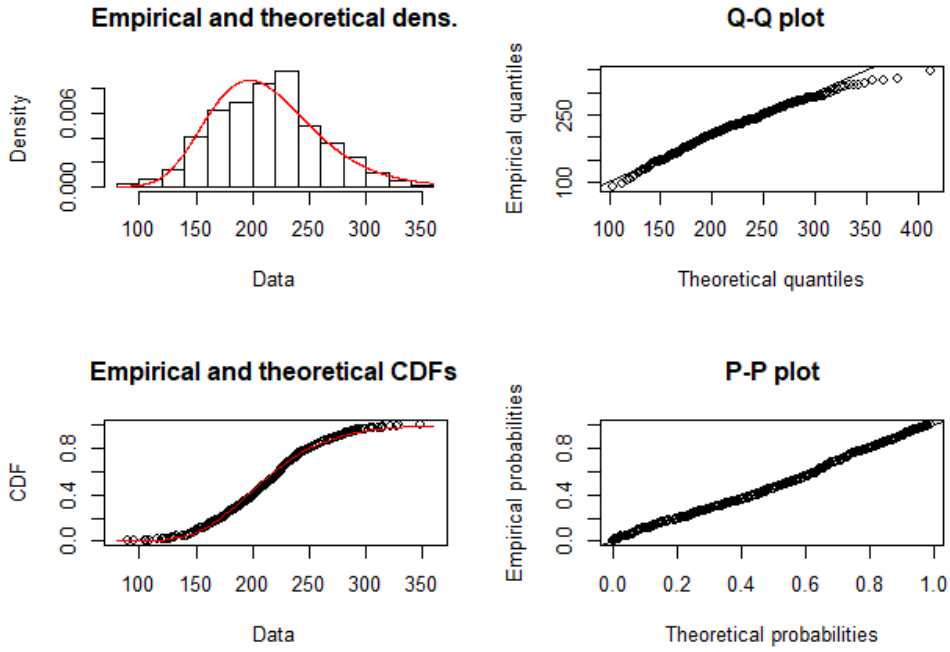


FIGURE C.2 – Adéquation loi log-Normal - Cluster 1

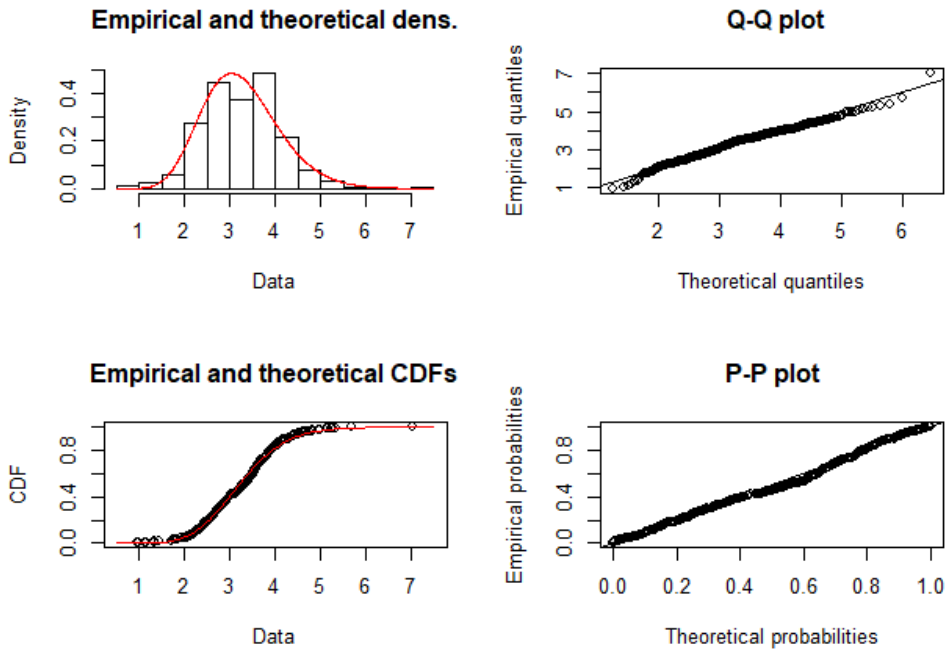


FIGURE C.3 – Adéquation loi Gamma - Cluster 2

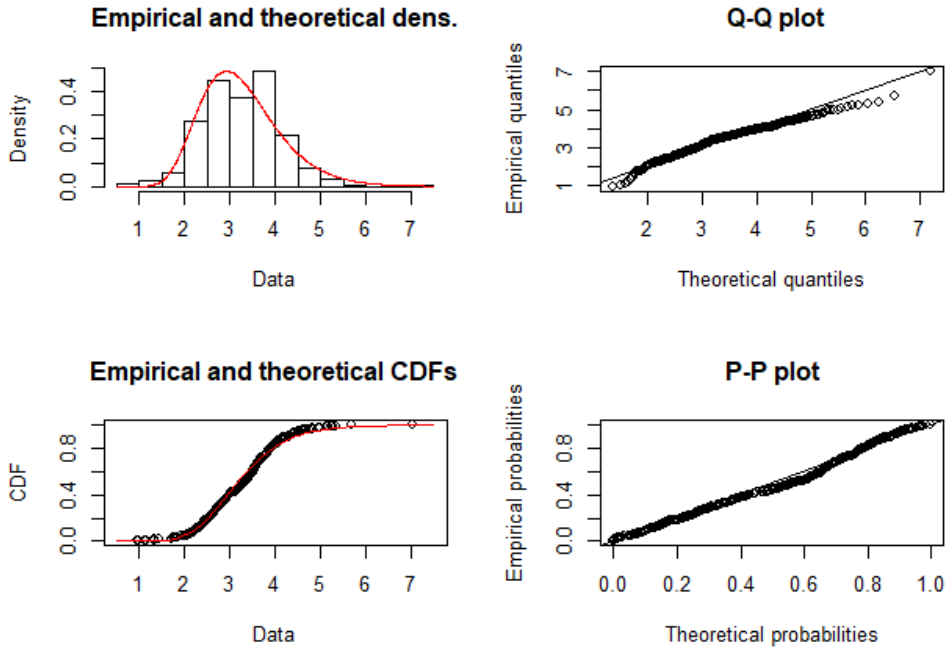


FIGURE C.4 – Adéquation loi log-Normal - Cluster 2

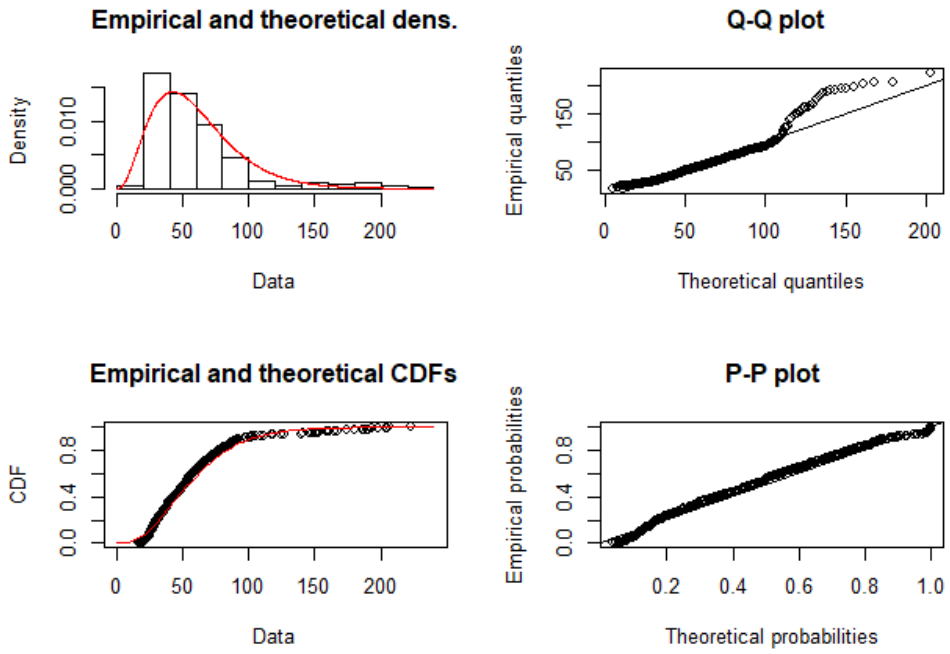


FIGURE C.5 – Adéquation loi Gamma - Cluster 3

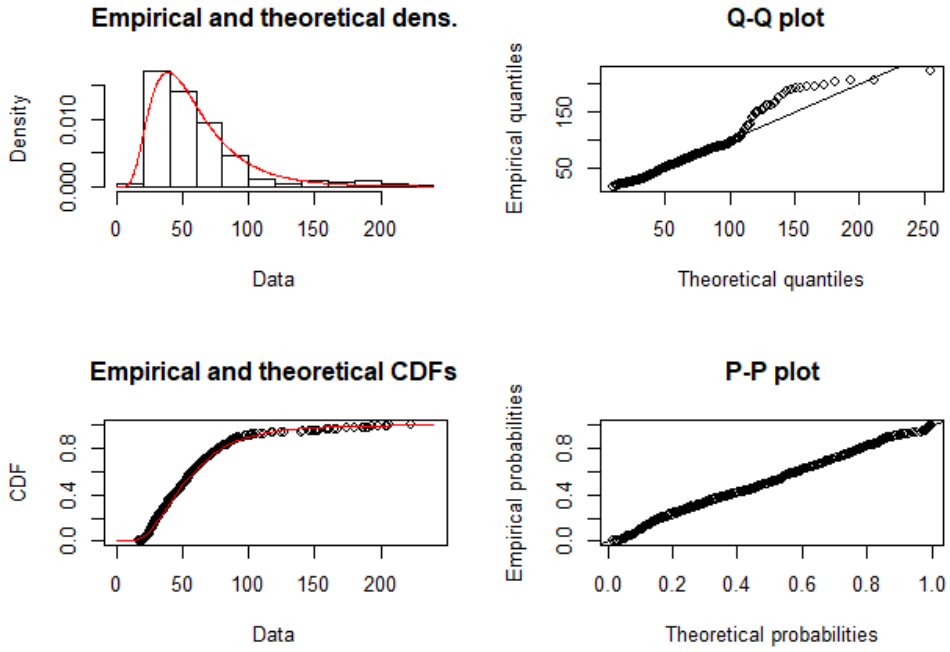


FIGURE C.6 – Adéquation loi log-Normal - Cluster 3

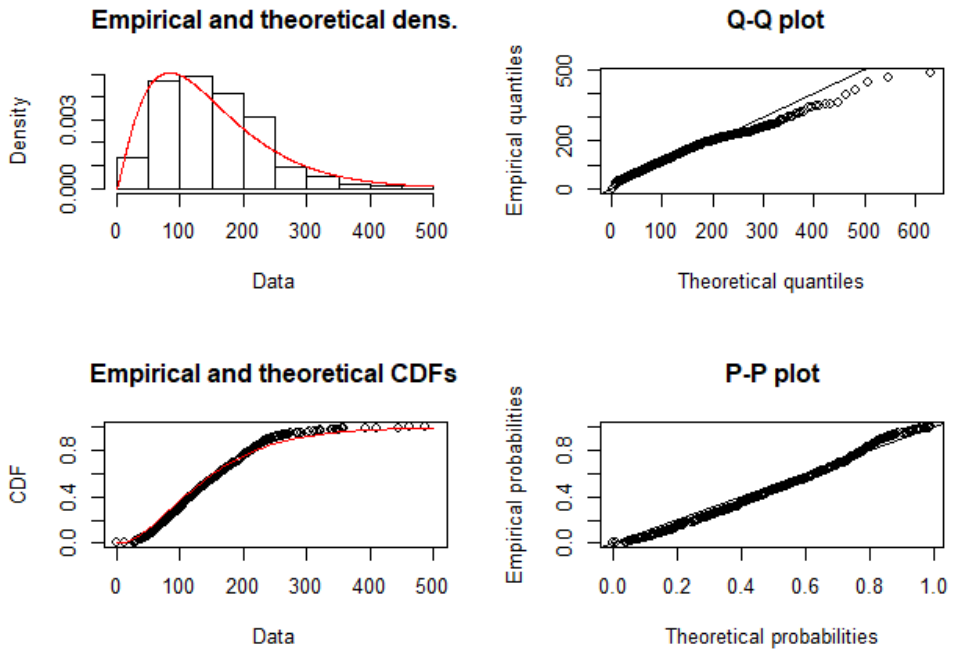


FIGURE C.7 – Adéquation loi Gamma - Cluster 4

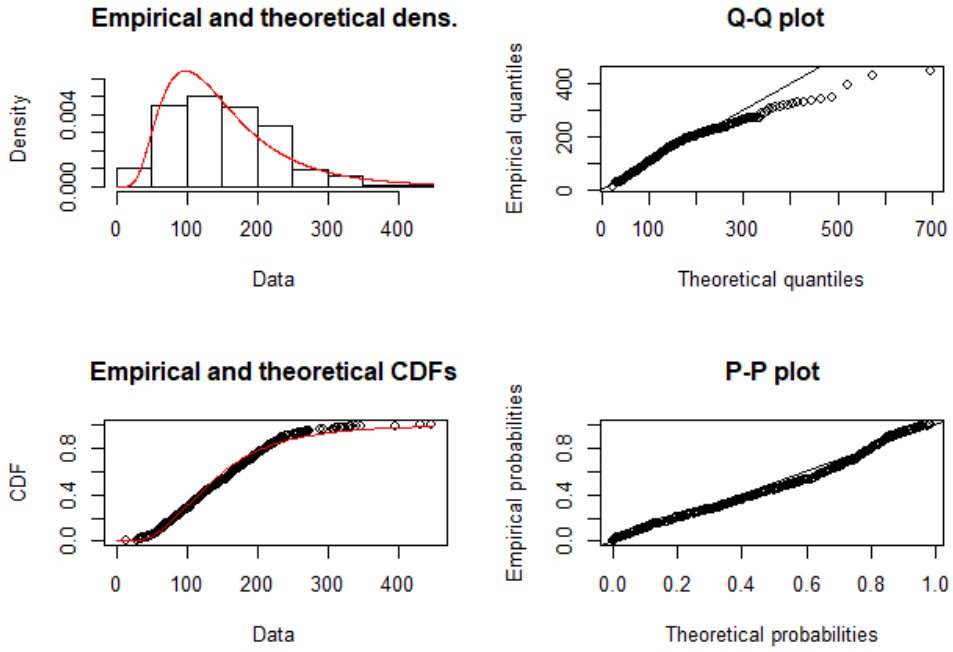


FIGURE C.8 – Adéquation loi log-Normal - Cluster 4

Annexe D

Représentation graphique des clusters bruits

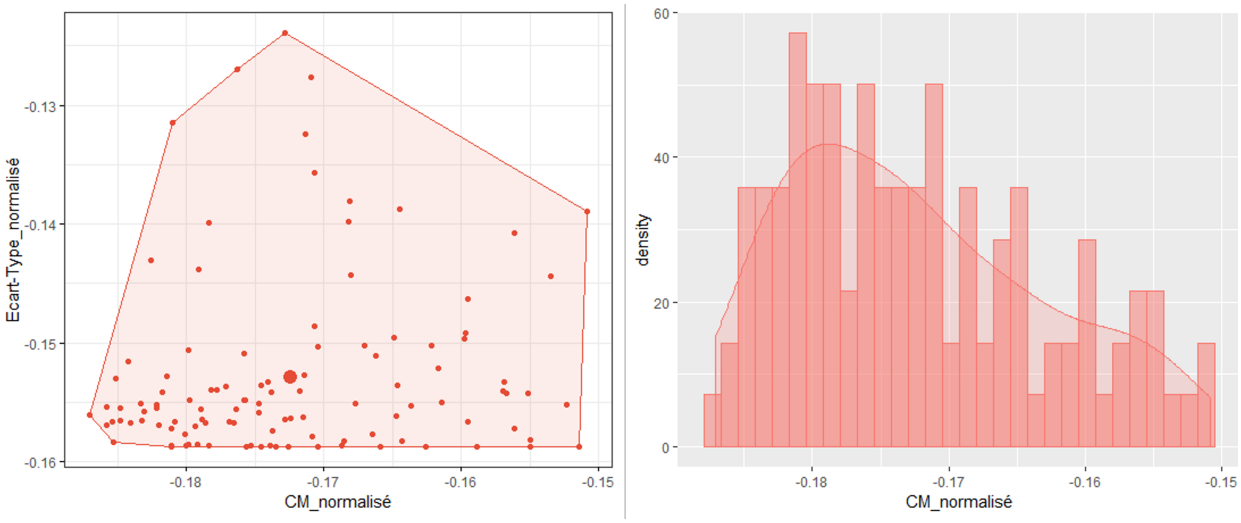


FIGURE D.1 – Cluster bruit - segmentation initiale selon coût moyen

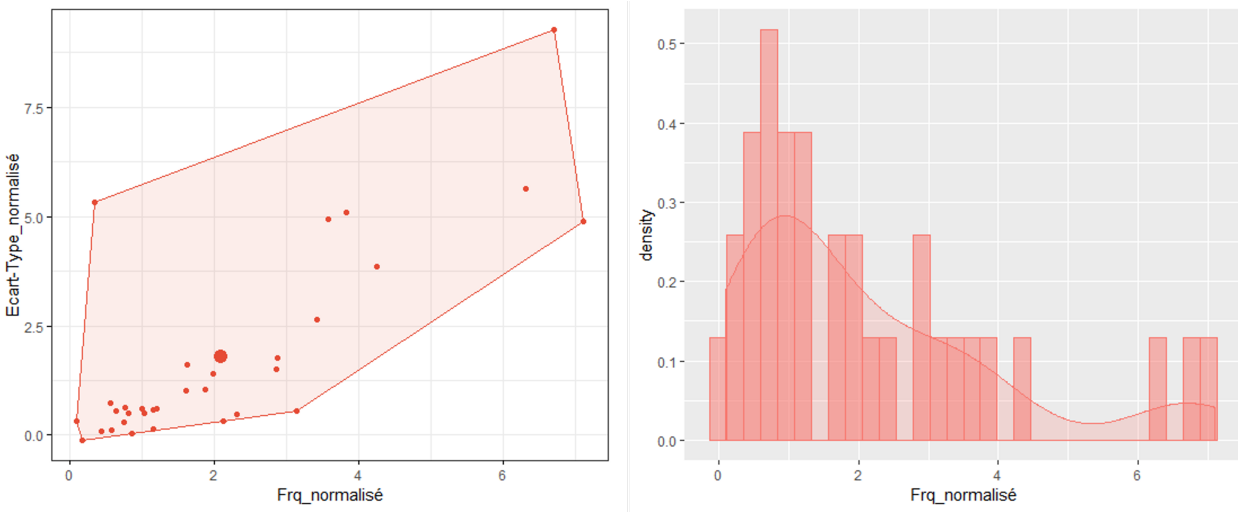


FIGURE D.2 – Cluster bruit - segmentation initiale selon fréquence

Bibliographie

[ABDOLLAHI(2017)] ABDOLLAHI, F. 2017, *Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité*, Mémoire d'actuariat, ISUP.

[AILLIOT(2018)] AILLIOT, P. 2018, *Modèles linéaires*, Cours master 1, EURIA.

[ANDRIEU(2015)] ANDRIEU, B. 2015, *Tarification et modélisation de la prime pure pour les garanties hospitalisation, soins courants et bien-être en santé individuelle*, Mémoire d'actuariat, EURIA.

[ATIF()] ATIF, J. *Machine Learning et Data Mining : Clustering, Groupement, Segmentation*, Cours master, Dauphine.

[CAROLO(2014)] CAROLO, P. 2014, *Le risque hospitalisation pour les complémentaires santé*, Mémoire d'actuariat, ISFA.

[CHARPENTIER et DENUIT(2005)] CHARPENTIER, A. et M. DENUIT. 2005, *Mathématiques de l'assurance non-vie, tome II : tarification et provisionnement*, Economica.

[CHARPENTIER et DUTANG(2012)] CHARPENTIER, A. et C. DUTANG. 2012, *L'Actuariat avec R*, thèse de doctorat.

[CHESNEAU(2017)] CHESNEAU, C. 2017, *Modèles de régression*, Cours master, Université de Caen.

[COURJAULT-RADE(2018)] COURJAULT-RADE, V. 2018, *Ballstering : un algorithme de clustering dédié à de grands échantillons*, Thèse, Université Toulouse 3 Paul Sabatier.

[GU(2012)] GU, R. 2012, *Sinistralité en assurance santé : modélisation, estimation et application*, Mémoire d'actuariat, ISFA.

[KARATEKIN(2014)] KARATEKIN, O. 2014, *Tarification et mesure de l'antisélection en assurance santé collective*, Mémoire d'actuariat, DUAS.

[KASSAMBARA()] KASSAMBARA, A. «Data mining and statistics for decision support», https ://www.datanovia.com/en/.

[LAGOS(2018)] LAGOS, T. 2018, *Étude de méthodes innovantes de machine learning permettant la tarification de produits Santé en mobilité internationale*, Mémoire d'actuariat, ISFA.

[LAILY(2019)] LAILY, R. 2019, *Application en tarification non-vie sous R*, Cours master 2, EURIA.

[MCINNES et collab.(2019)MCINNES, HEALY et ASTELS] MCINNES, L., J. HEALY et S. ASTELS. 2019, «*The hdbscan Clustering Library*», .

[MORIN(2012)] MORIN, J.-B. 2012, *La tarification en santé*, Mémoire d'actuariat, ISFA.

[NGUYEN(2013)] NGUYEN, N. T. P. 2013, *Construction de bases de tarification pour des contrats complémentaires santé collectifs par le Modèle Linéaire Généralisé*, Mémoire d'actuariat, ISFA.

[POSTEAU(2015)] POSTEAU, R. 2015, *Personnalisation du processus de tarification santé*, Mémoire d'actuariat, ISFA.