

Mémoire présenté le :
pour l'obtention du diplôme de
Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : **Alexandre Beaune**

Titre : **Création d'une table d'incidence en invalidité**

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

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

Membres présents du jury de la filière

Guillaume Biessy

Entreprise : AXA France 

Nom : Antoine Nguyen

Signature :



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

Jury IA 1

Jury IA 2

Jury IA 3

Jury IA 4

Jury IA 5

Jury IA 6

Directeur du mémoire en entreprise :

Nom : Antoine Nguyen

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

Secrétariat :



Signature du candidat

Bibliothèque :




Table des matières

Résumé	3
Abstract	3
Synthèse	4
Summary	8
Remerciements	12
1 Introduction	14
1.1 Présentation de l'entité	14
1.2 La prévoyance	14
1.3 L'invalidité	15
1.4 Contexte et problématique	16
2 Extraction et retraitements des données	17
2.1 Récupération et retraitements des sinistres invalidité	18
2.2 Récupération et retraitements des éléments de la DSN	19
2.3 Jointure des tables et derniers retraitements	20
3 Statistiques descriptives	22
3.1 Statistiques descriptives de la population en portefeuille	22
3.2 Statistiques descriptives de la population sinistrée	23
4 Estimation des taux bruts d'entrée en invalidité	26
4.1 Analyse de survie : Censure et troncature	26
4.2 L'estimateur de Kaplan-Meier	29
4.3 L'estimateur des moments de Hoem	32
4.4 Résultats	33
5 Lissage des taux	37
5.1 Lissage par splines cubiques	37
5.2 Lissage par Whittaker-Henderson	40
5.3 Lissage par moyennes mobiles	44
5.4 Lissage par la méthode des noyaux	47

6	Tests de lissage et validation de la table	51
6.1	Test du changement de signe	51
6.2	Test du SMR	53
7	Comparaison avec les lois d'incidence en invalidité existantes	57
8	Tarification	59
9	Lois d'incidence pour les cadres et les non-cadres	63
9.1	Quelques statistiques descriptives	64
9.2	Estimation des taux bruts d'entrée en invalidité	66
9.3	Lissage des taux	68
9.4	Tests de lissage et validation de la table	70
9.5	Tarification	72
9.6	Conclusion sur les lois cadres/non-cadres	74
10	Conclusion	75
	Références	76
	Annexes	79

Résumé

Ce mémoire a pour objectif la construction d'une table d'incidence en invalidité depuis l'extraction des données qui permettent de l'établir jusqu'à son utilité dans l'outil de tarification. Cette table est le fruit d'une succession d'étapes que les parties de ce mémoire abordent chronologiquement.

Il faut avant tout savoir que l'outil de tarification actuel de la garantie invalidité repose sur deux lois : une loi femmes et une loi hommes. Nous avons donc répondu dans un premier temps à ce besoin opérationnel de mise à jour de ces deux tables.

Pour cela, il a fallu tout d'abord définir notre périmètre d'étude et collecter les données en conséquence. En l'occurrence, la table a été construite à partir des données des DSN (Déclarations Sociales Nominatives) recoupées avec les données d'invalidité issues des tables internes à AXA sur une période s'étendant de janvier 2017 à janvier 2022.

A partir de ces données, nous avons extrait des variables essentielles à l'estimation des taux bruts d'entrée en invalidité comme les dates de début et de fin de contrat, les dates de survenance ou encore le sexe. Par ailleurs, cette estimation des taux bruts d'entrée en invalidité a été faite grâce à des modèles relativement traditionnels comme celui de Kaplan-Meier ou celui s'appuyant sur la méthode des moments de Hoem. Après analyse des taux obtenus, nous avons retenu le modèle de Hoem car il maximise davantage les taux et est donc considéré comme étant plus prudent. De plus, sur les âges élevés, il semble mieux percevoir le comportement des taux réels du fait que ce n'est pas un estimateur qui dépend des valeurs précédentes contrairement à l'estimateur de Kaplan-Meier.

Une fois ces taux bruts obtenus, nous les avons lissés via différentes méthodes (Whittaker-Henderson, splines cubiques, noyaux, moyennes mobiles). Après avoir appliqué des tests statistiques (changement de signe, SMR), il a été choisi de retenir celui de Whittaker-Henderson avec des paramètres du lissage qui diffèrent selon les populations étudiées.

Suite à l'obtention des taux lissés, nous avons pu alors calculer un tarif par âge avant de le comparer dans un dernier temps à un tarif prenant en compte l'âge actuariel. Une fois cette démarche accomplie, nous avons donc répondu aux contraintes opérationnelles de l'outil de tarification. Toutefois, nous avons effectué le même processus mais cette fois-ci avec une segmentation cadres/non-cadres qui nous a semblé plus appropriée à la garantie invalidité. Nous avons également établi deux lois distinctes utilisant une estimation des taux bruts par Hoem et un lissage par Whittaker-Henderson avant de détailler les tarifs obtenus.

Mots clés : prévoyance collective, invalidité, déclarations sociales nominatives, Kaplan-Meier, méthode des moments de Hoem, lissage, moyennes mobiles, méthode des noyaux, Whittaker-Henderson, splines cubiques, SMR, test du changement de signes, tarification.

Abstract

The objective of this dissertation is to build a disability incidence table from the extraction of the data used to establish it to its use in the pricing tool. This table is the result of a succession of steps that the parts of this brief address chronologically.

First of all, it is important to know that the current rating tool for disability coverage is based on two laws : a female law and a male law. We have therefore first responded to the operational need to update these two tables. First of all, we simply had to define our study perimeter and collect the data accordingly. In this case, the table was built from DSN (Nominative Social Declarations) data cross-referenced with disability data from AXA's internal tables over a period extending from January 2017 to January 2022.

From these data, we extracted variables that are essential to the estimation of gross disability entry rates, such as contract start and end dates, occurrence dates and gender. Moreover, this estimation of the gross disability entry rates was done using relatively traditional models such as the Kaplan-Meier model or the one based on the Hoem method of moments. After analyzing the rates obtained, we chose the Hoem model because it maximizes the rates more and is therefore considered to be more conservative. Moreover, at high ages, it seems to better perceive the behavior of real rates because it is not an estimator that depends on previous values, unlike the Kaplan-Meier estimator.

Once these raw rates were obtained, we smoothed them using different methods (Whittaker-Henderson, cubic splines, kernels, moving averages). After applying statistical tests (change of sign, SMR), we chose to use the Whittaker-Henderson method with smoothing parameters that differed according to the populations studied.

Once the smoothed rates were obtained, we were able to calculate a rate per age and then compare it to a rate that takes into account the actuarial age. Once this process was completed, we were able to meet the operational constraints of the pricing tool. However, we performed the same process but this time with a management/non-management segmentation that we felt was more appropriate for disability coverage. We also established two separate laws using a Hoem estimation of the gross rates and a Whittaker-Henderson smoothing before detailing the obtained rates.

Keywords : group insurance, disability, nominative social declarations, Kaplan-Meier, method of moments of Hoem, smoothing, moving averages, method of kernels, Whittaker-Henderson, cubic splines, SMR, sign change test, pricing.

Synthèse

Ce mémoire a pour but de décrire la méthode de construction d'une table d'incidence en invalidité. Le véritable apport actuariel de cette étude est l'utilisation des DSN, obligatoires pour les organismes sociaux depuis 2017 et qui fournissent mensuellement de nombreuses données sur les personnes salariées. Toutefois, dans le cadre d'un arrêt de travail, les DSN ne mentionnent pas si la personne est en incapacité ou bien en invalidité. Pour le déterminer, nous devons regarder dans les tables propres à AXA qui répertorient les invalidités. Nous avons ensuite réalisé des jointures entre ces sources de données grâce à des variables clés.

Notre étude a donc été réalisée entre le 1er janvier 2017 et le 1er janvier 2022, soit une plage de 5 ans.

Une fois nos données retraitées intégralement, nous obtenons une table où chaque ligne correspond à un individu muni d'un numéro de régime professionnel de prévoyance (RPP). Ceci est indispensable car, comme l'étude est réalisée pour les assurances collectives, la tarification se fait à l'échelle d'une entreprise (à l'échelle d'un numéro de RPP). Il est donc possible de compter deux fois un même individu qui travaille dans deux entreprises différentes mais nous le comptons aussi deux fois dans notre exposition ce qui ne pose donc pas de souci au final.

Avant d'évoquer l'élaboration des taux bruts, il faut savoir que l'outil de tarification actuel est basé sur une loi d'incidence femmes et une loi d'incidence hommes. Bien qu'intuitivement une telle segmentation paraisse peu adaptée à la garantie invalidité (par rapport à une segmentation cadres/non-cadres par exemple), nous devons dans un premier temps répondre aux contraintes opérationnelles de l'outil et mettre à jour ces deux lois en profitant de toutes les données fournies par les DSN.

Pour estimer les taux bruts, nous avons utilisé des méthodes traditionnelles d'analyse de survie, tenant donc compte des censures et troncatures, telles que l'estimateur de Kaplan-Meier ou celui de Hoem. Du fait de taux légèrement plus élevés (et donc plus prudents) et d'une fiabilité un peu plus aiguisée sur les âges supérieurs à 55 ans, l'estimateur de Hoem a été préféré à celui de Kaplan-Meier. Bien que le volume conséquent de données aboutisse à des courbes des taux bruts déjà relativement régulières, nous leur avons appliqué différentes méthodes de lissages, à savoir le lissage par splines cubiques, le lissage par Whittaker-Henderson, le lissage par moyennes mobiles et le lissage par la méthode des noyaux. Nous avons par ailleurs fait varier les paramètres des lissages pour voir leur impact et ajuster au mieux les taux lissés.

Suite à cela, nous avons effectué des tests statistiques tels que le test du changement de signes ou le test du SMR. Le lissage par noyaux et celui par moyennes mobiles présentant l'in-

convénient de ne pas lisser suffisamment de valeurs, celui par splines cubiques ne validant pas le test du changement de signes, nous avons opté pour un lissage par Whittaker-Henderson qui, après sélection des paramètres de lissage optimaux, nous donne les résultats suivants :

Pour la population femmes :

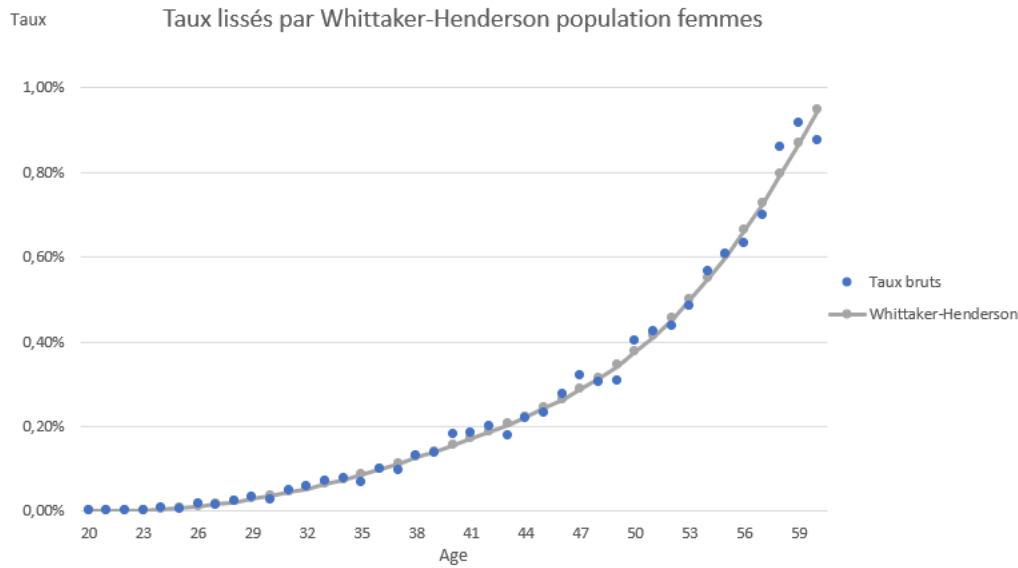


FIGURE 1 – Taux lissés définitifs pour la population femmes

et pour la population hommes :

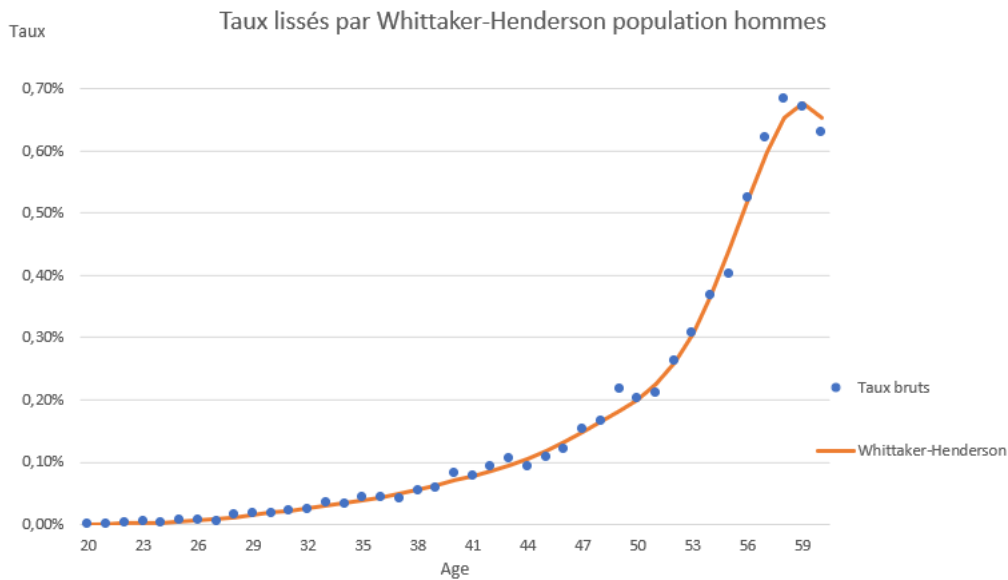


FIGURE 2 – Taux lissés définitifs pour la population hommes

Nous avons alors comparé les lois obtenues à celles résultant de la précédente étude (qui avait été réalisée en 2015). Nous en concluons l'apport considérable des DSN et une sous-estimation globale des taux d'incidence en invalidité auparavant.

Enfin, grâce à ces lois, nous déterminons un tarif "individuel" par âge résultant de formules classiques d'assurance vie que nous comparons à un tarif "groupe" prenant en compte l'âge actuariel c'est à dire la moyenne des âges d'un groupe au regard des risques assurés. Cette dernière étape vient conclure notre étude et répondre par la même occasion au besoin de mise à jour des lois.

Néanmoins, comme évoqué plus haut, une segmentation cadres/non-cadres voire 4 lois distinctes (femmes, hommes, cadres, non-cadres) nous semblant plus judicieuse, nous avons refait l'intégralité du processus décrit. Une fois encore, les taux bruts ont été estimés par la méthode des moments de Hoem, sur lesquels nous avons appliqué un lissage de Whittaker-Henderson nous fournissant les résultats suivants :

Pour la population cadres :

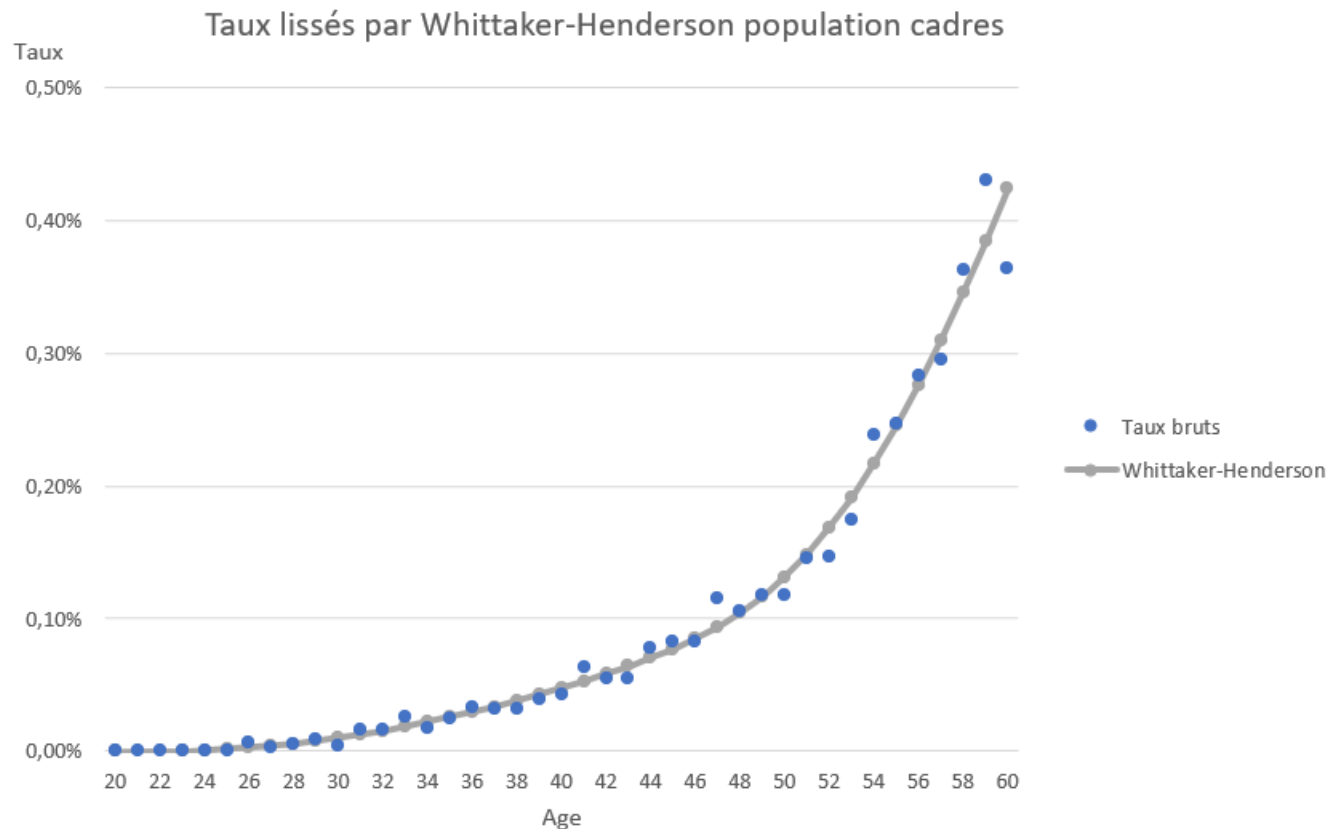


FIGURE 3 – Taux lissés définitifs pour la population femmes

et pour la population non-cadres :

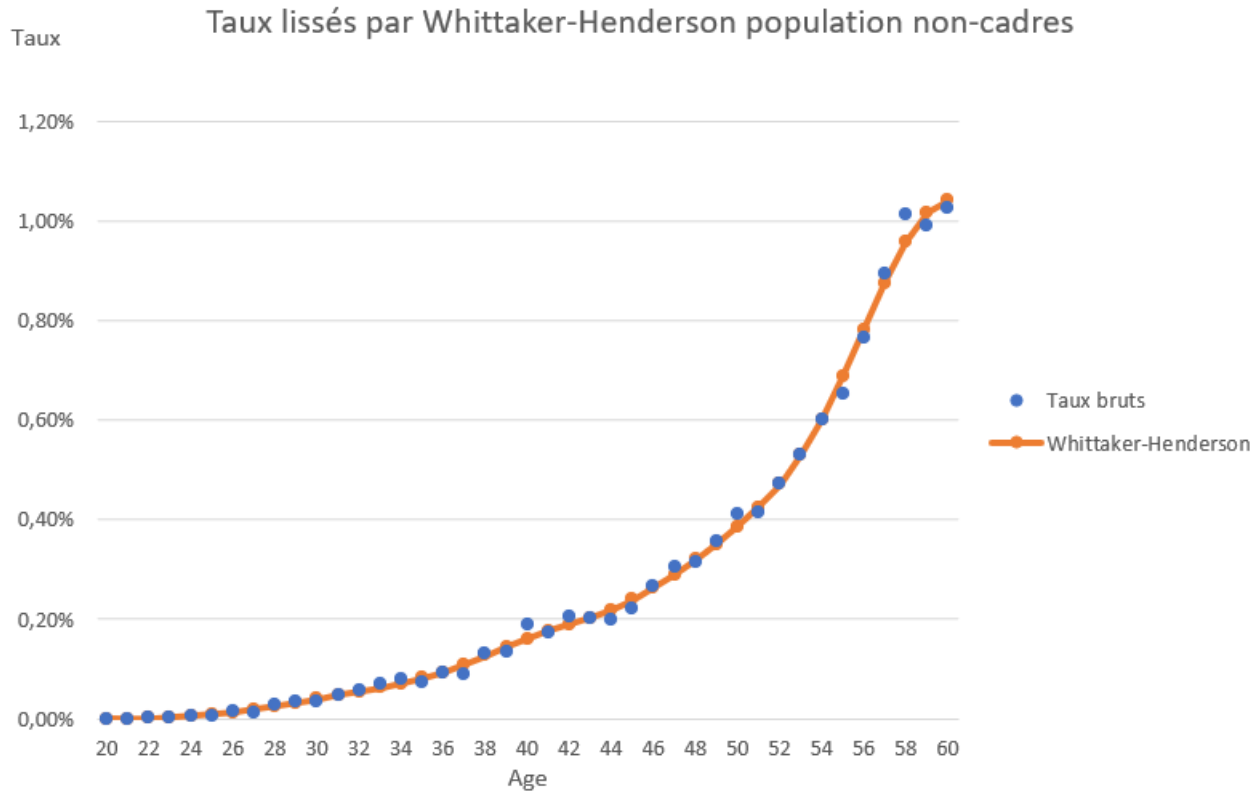


FIGURE 4 – Taux lissés définitifs pour la population hommes

Au vu de l'écart important des taux entre les deux populations, une implémentation de cette segmentation en complément de la segmentation femmes/hommes nous semble appropriée car elle permettrait d'ajuster au mieux le tarif aux sous-populations concernées. Nous avons suite à cela indiqué les tarifs engendrés d'une telle segmentation selon le même processus que celui mentionné pour les femmes et les hommes.

Pour conclure, ce travail a permis d'une part de voir la construction d'une loi d'incidence depuis la collecte des données jusqu'à son utilisation finale à savoir la tarification. D'autre part, nous avons pu comprendre les enjeux et les caractéristiques propres au monde des assurances collectives. Enfin, nous avons su utiliser nos connaissances statistiques pour répondre à des contraintes opérationnelles dans un premier temps puis aller plus loin en suggérant d'affiner la segmentation de la tarification avec une justification statistique à l'appui dans un second temps.

Summary

The purpose of this dissertation is to describe the method for constructing a disability incidence table. The real actuarial contribution of this study is the use of the DSNs, which have been mandatory for social organizations since 2017 and which provide a lot of data on employees on a monthly basis. However, in the context of a work stoppage, the DSNs do not mention whether the person is on disability or incapacity. To determine this, we have to look in AXA's own tables that list disabilities. We then made joins between these data sources using key variables.

Our study was therefore conducted between January 1, 2017 and January 1, 2022, i.e., a five-year period.

Once our data has been fully reprocessed, we obtain a table where each row corresponds to an individual with a professional pension plan (PPP) number. This is necessary because, as the study is conducted for group insurance, the pricing is done at the company level (at the level of an OPP number). It is therefore possible to count the same individual working in two different companies twice, but we also count him twice in our exposure, so this is not a problem in the end. Before discussing the elaboration of the gross rates, it is important to know that the current pricing tool is based on a female incidence law and a male incidence law. Although intuitively such a segmentation may not seem very adapted to disability coverage (compared to a management/non-management segmentation, for example), we must first respond to the operational constraints of the tool and update these two laws by taking advantage of all the data provided by the DSN.

To estimate the crude rates, we used traditional survival analysis methods, thus taking into account censoring and truncation, such as the Kaplan-Meier or Hoem estimators. Because of slightly higher (and therefore more conservative) rates and slightly higher reliability for ages above 55 years, the Hoem estimator was preferred to the Kaplan-Meier estimator. Although the large volume of data results in crude rate curves that are already relatively regular, we applied different smoothing methods, namely cubic spline smoothing, Whittaker-Henderson smoothing, moving average smoothing and kernel smoothing. We also varied the parameters of the smoothings to see their impact and to adjust the smoothed rates as best as possible.

Following this, we performed statistical tests such as the sign change test or the SMR test. Since the kernel and moving average smoothing methods have the disadvantage of not smoothing enough values, and since the cubic spline method does not validate the sign change test, we opted for a Whittaker-Henderson smoothing method which, after selecting the optimal smoothing parameters, gives us the following results For the female population :

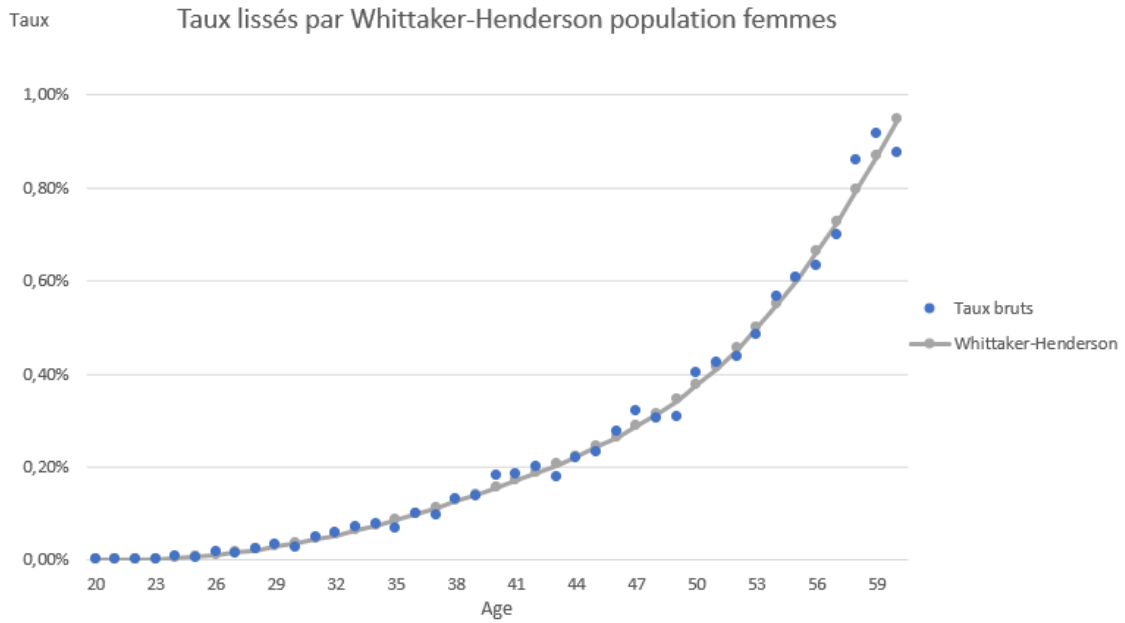


FIGURE 5 – Final smoothed rates for the female population

and for the male population :

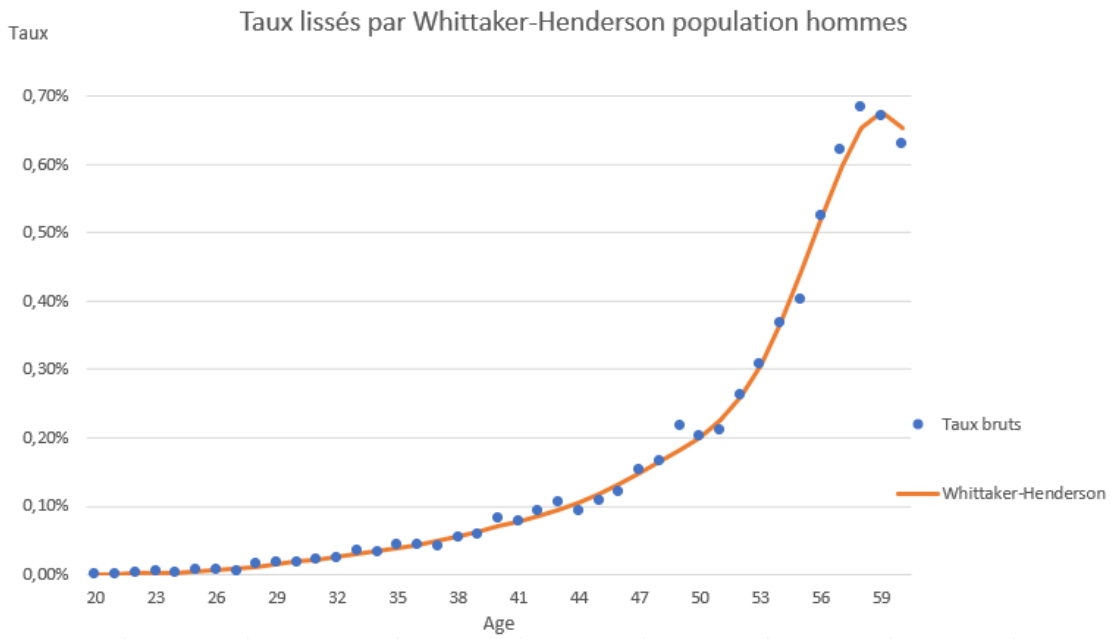


FIGURE 6 – Final smoothed rates for the male population

We then compared the resulting laws to those resulting from the previous study (which was conducted in 2015). We conclude that the NSNs have made a considerable contribution and that the overall incidence rates in disability were underestimated before.

Finally, thanks to these laws, we determine an "individual" rate per age resulting from classic life insurance formulas that we compare to a "group" rate taking into account the actuarial age, i.e. the average age of a group with respect to the insured risks. This last step concludes our study and at the same time responds to the need to update the laws.

Nevertheless, as mentioned above, a segmentation between executives and non-executives, or even 4 distinct laws (women, men, executives, non-executives) seems to us to be more judicious, so we have redone the entire process described. Once again, the gross rates were estimated by the Hoem method of moments, to which we applied a Whittaker-Henderson smoothing, giving us the following results :

For the managerial population :

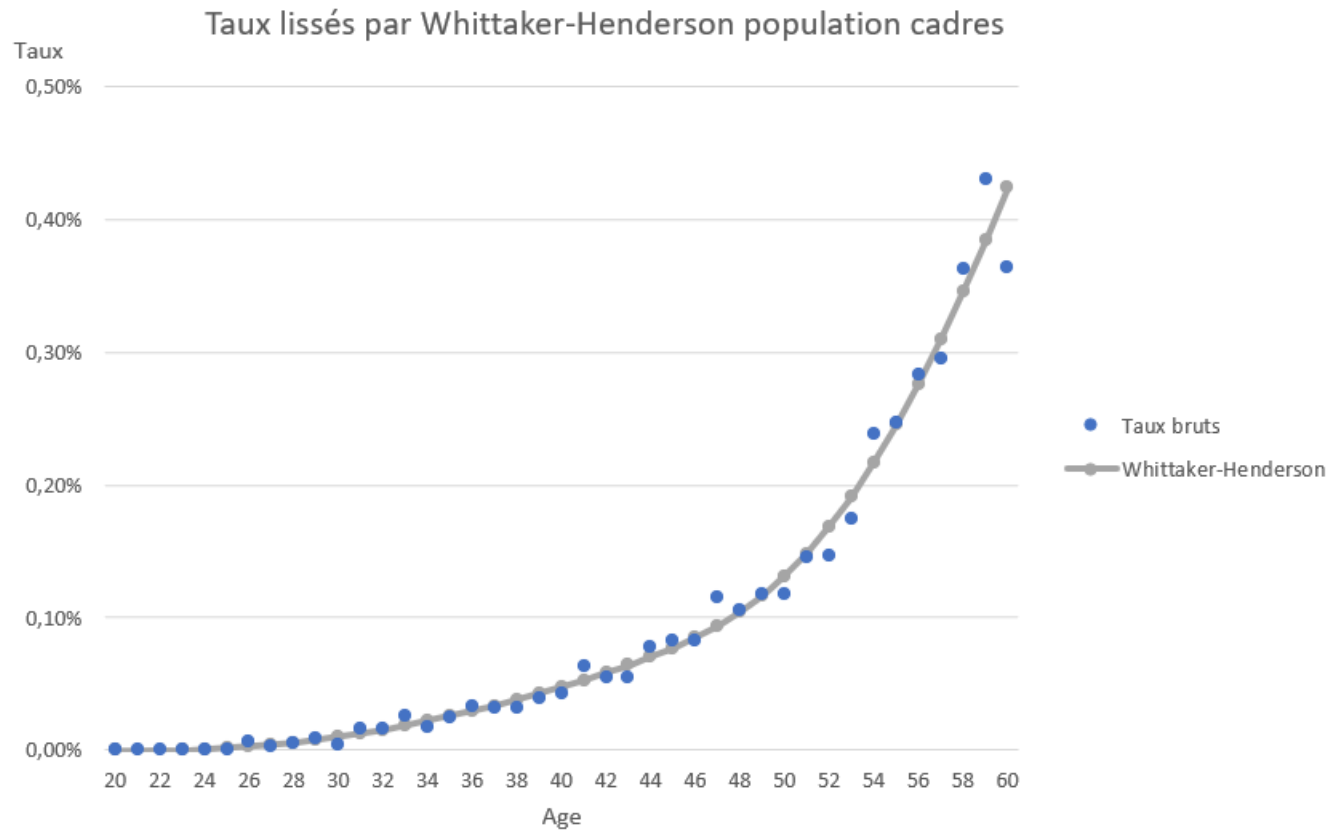


FIGURE 7 – Final smoothed rates for the managerial population

and for the non-managerial population :

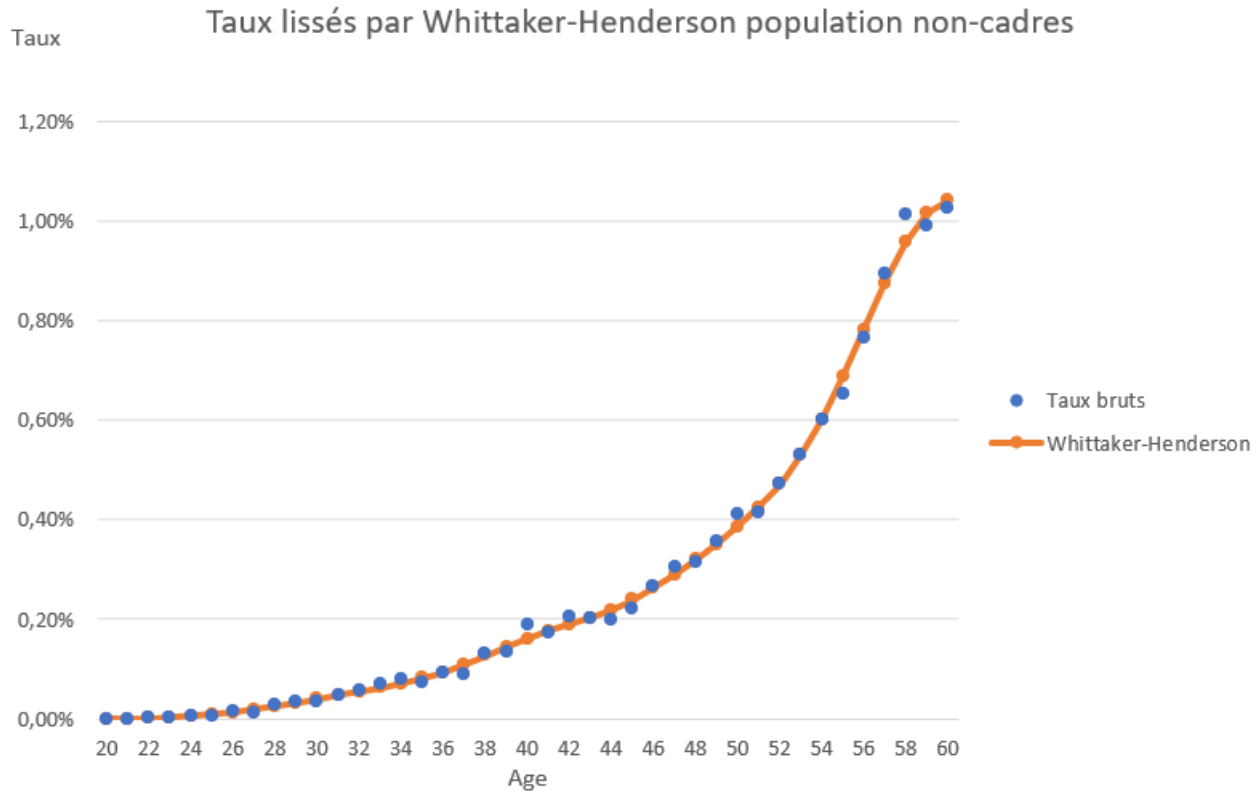


FIGURE 8 – Final smoothed rates for the non-managerial population

In view of the significant difference in rates between the two populations, an implementation of this segmentation in addition to the women/men segmentation seems appropriate to us because it would allow us to adjust the rate to the sub-populations concerned. We have then indicated the tariffs generated by such a segmentation according to the same process as the one mentioned for women and men.

In conclusion, this work has allowed us to see the construction of an incidence law from data collection to its final use, i.e. pricing. We were also able to understand the issues and characteristics specific to the world of group insurance. Finally, we were able to use our statistical knowledge to respond to operational constraints in the first instance, and then to go further by suggesting refining the segmentation of the pricing with a statistical justification in the second instance.

Remerciements

Je tenais en premier lieu à remercier Anne Fahy, la responsable de l'équipe Normes et études techniques France, qui m'a permis de faire mon alternance au sein de la Direction Technique Notoriété et Innovations France.

Je souhaitais également remercier l'ensemble des collaborateurs de l'équipe, qui ont su faciliter mon intégration d'une part et m'apprendre beaucoup sur le monde des collectives d'autre part.

Enfin, je remercie Antoine Nguyen, mon tuteur de stage qui m'a formé et accompagné tout au long de cette expérience en cherchant à me faire progresser en permanence.

1 Introduction

1.1 Présentation de l'entité

Créée en 1985 par Claude Bébéar, possédant 107 millions de clients, implanté désormais dans 64 pays, fort d'un chiffre d'affaires d'environ 100 milliards d'euros, AXA est aujourd'hui un leader mondial du monde de l'assurance.

AXA France, élément phare du groupe AXA représente quasiment un tiers de ce chiffre d'affaires et rassemble environ 33 000 collaborateurs pour un total de 6,3 millions de clients.

L'entité organisationnelle AXA Santé et Collectives d'AXA France est quant à elle chargée des assurances collectives et de la santé.

Parmi les différentes directions qui constituent cette entité figure la Direction Technique Notoriété et Innovations (DTNI), direction transverse, qui vise d'une part à garantir la rentabilité de nouvelles gammes et définir des règles de souscription mais également à développer le rayonnement d'AXA Santé et Collectives.

J'ai réalisé mon mémoire au sein de l'équipe Normes et Etudes Techniques France portant donc sur la création d'une table d'incidence en invalidité et visant ainsi à améliorer la tarification de produits liés à cette garantie prévoyance.

Avant de rentrer en détails dans la notion d'invalidité, il convient donc de définir plus globalement ce qu'est la prévoyance.

1.2 La prévoyance

La prévoyance est à proprement parler une partie de l'assurance de personnes et se positionne ainsi à cheval sur les deux domaines traditionnels que sont "la vie" et "la non-vie". Plus concrètement, ce domaine touche à tous les risques liés directement à une personne, par opposition à l'assurance dite "de dommages" qui couvre des biens.

Les garanties prévoyance sont souscrites pour faire face majoritairement à des décès, des incapacités de travail ou des invalidités via le versement d'un capital ou d'une rente. Des garanties de base sont souvent assurées par le régime de la sécurité sociale, mais les organismes d'assurance, les mutuelles ou les instituts de prévoyance peuvent proposer des garanties complémentaires pour augmenter la couverture de risque. Le contrat d'assurance lie alors l'assureur et le souscripteur du contrat : l'assureur s'engage à verser un capital ou une rente à une personne en échange d'une prime. La personne bénéficiant de cette rente ou de ce capital n'est pas nécessairement la personne qui souscrit le contrat.

Nous allons à présent décrire brièvement les principales garanties prévoyance :

L'assurance en cas de décès :

Cette assurance garantit le versement d'un capital ou d'une rente à un ou plusieurs bénéficiaires déterminés en cas de décès de l'assuré intervenant avant le terme du contrat. Ici, l'assuré diffère logiquement des bénéficiaires. Les garanties associées les plus connues sont la rente de conjoint et la rente éducation.

La garantie incapacité temporaire/invalidité permanente (IT/IP) :

En cas de maladie ou d'accident qui conduisent à un arrêt de travail, l'assuré ayant une garantie IT/IP reçoit des indemnités journalières. L'incapacité temporaire concerne les cas d'un arrêt de travail temporaire, l'invalidité permanente concerne les cas de diminution durable de la capacité de travail, c'est-à-dire les états physiques et mentaux plaçant l'assuré dans l'incapacité totale, permanente et présumée définitive d'exercer une activité rémunératrice. Notre étude va donc porter sur la garantie IP.

Notons que bien que peu fréquent, il peut arriver qu'un individu soit directement en invalidité sans passer par l'incapacité. L'objet de cette étude est de construire une loi d'incidence en invalidité quel que soit l'état dans lequel était l'individu avant d'entrer en invalidité (incapacité ou actif valide).

1.3 L'invalidité

Une personne est considérée invalide au sens de la Sécurité sociale si, après un accident ou une maladie survenue dans sa vie, sa capacité de travail ou de gain est réduite d'au moins $\frac{2}{3}$. Pour parler plus en profondeur de l'invalidité, il faut savoir qu'il existe 3 types d'invalidité : IP1, IP2 et IP3 qui correspondent aux éléments suivants :

1. L'IP1 : le salarié est encore capable d'exercer une activité rémunérée
2. L'IP2 : la victime a perdu $\frac{2}{3}$ de ses activités de travail et ne peut plus exercer de travail quel qu'il soit
3. L'IP3 : en plus d'être totalement incapable d'exercer quelconque activité, la personne a recours à une assistance pour effectuer les tâches ordinaires de la vie

La cause de l'invalidité peut être d'origine professionnelle ou non-professionnelle.

La catégorie d'invalidité étant liée à un taux d'incapacité, elle détermine directement l'indemnité payée par la Sécurité Sociale à la victime. En tant que complémentaire, AXA agit selon d'autres critères pour attribuer une pension d'invalidité, à savoir :

1. percevoir les rentes ou pensions versées par la Sécurité Sociale
2. dans le cas d'une invalidité d'origine non professionnelle, répondre aux critères d'attribution d'une pension d'invalidité fixés à l'article L 341-1 du Code de la Sécurité Sociale :
 - (a) Ne pas avoir atteint l'âge légal de départ à la retraite.
 - (b) Avoir une capacité de travail réduite d'au moins deux tiers.
 - (c) Justifier de douze mois d'immatriculation en tant qu'assuré social.
 - (d) Avoir travaillé au moins 600 heures ou cotisé au moins 2030 fois le SMIC horaire au cours des douze derniers mois.
3. dans le cas d'une invalidité d'origine professionnelle, se voir attribuer un taux d'incapacité permanente supérieur ou égal à 33%

Notons enfin que bien qu'étant censée avoir un caractère définitif par définition, la catégorie d'invalidité d'un individu peut être amenée à changer. Ces éléments peuvent faire l'objet de construction de différentes tables de passage de catégorie d'invalidité à une autre et par la même occasion de tables de maintien en invalidité par exemple.

Pour expliquer comment nous avons construit notre table d'incidence en invalidité, nous allons commencer par présenter le contexte ainsi que les données que nous avons utilisées et la manière dont nous les avons traitées.

1.4 Contexte et problématique

Cette création de table d'incidence en invalidité répond avant tout à un besoin de mettre à jour l'étude qui servait jusqu'à maintenant de référence dans l'outil de tarification. Cette étude, réalisée en 2015 et portant sur la période de 2011 à 2014, avait été réalisée sur la base d'un rapprochement entre les sinistres répertoriés dans les tables internes et les DADS-U (Déclarations Annuelles des Données Sociales Unifiées). Elle avait permis l'élaboration de deux lois d'incidence : une loi femmes et une loi hommes qui sont à ce jour implémentées dans l'outil.

Les DADS-U sont en quelque sorte l'ancêtre des DSN à la différence près qu'elles n'étaient pas autant utilisées. En ce sens, les DSN sont censées apporter un caractère beaucoup plus fiable à notre étude et ajuster par la même occasion la tarification de la garantie invalidité.

C'est ce processus de construction de lois femmes et hommes à partir des DSN que nous décrivons dans la suite de ce mémoire. Ce dernier répondra donc à la problématique suivante : en quoi les DSN constituent-elles un apport conséquent dans la création d'une table d'incidence en invalidité ?

2 Extraction et retraitements des données

Concernant le périmètre de notre étude, nous avons décidé de prendre tous les contrats collectifs en portefeuille, qu'ils soient en gestion directe ou en gestion déléguée, sur la plage temporelle s'étendant du 01/01/2017 au 01/01/2022, soit une durée de 5 ans.

Nous avons fait ce choix car l'intervalle de temps nous semble assez large pour établir une loi robuste s'appuyant sur un nombre important de données et en même temps assez fin pour qu'on puisse considérer cette loi comme étant d'actualité. Enfin, hormis les lissages et l'aspect tarification réalisés sous Excel, l'ensemble des opérations effectuées (extraction des données, retraitements, estimations des taux bruts) a été fait sous SAS.

Pour obtenir une table d'incidence en invalidité, il faut avoir une table présentant les éléments suivants :

1. un identifiant sinistre unique pour chaque assuré
2. une date de naissance
3. une date de survenance de l'invalidité
4. des dates de début et de fin de contrat de travail
5. des dates de début et de fin de contrat d'assurance
6. le sexe de la personne
7. la catégorie socio-professionnelle de la personne
8. le salaire de la personne

Le sexe et la catégorie socio-professionnelle de la personne permettent de créer des sous-catégories dans notre population et par la même occasion d'avoir les tables d'incidence en invalidité correspondantes. Ceci semble logique puisque nous pouvons nous douter que la loi d'incidence en invalidité des hommes diffère de celles des femmes et idem pour les cadres vis-à-vis des non-cadres. Par exemple, à AXA, la tarification de la garantie invalidité temporaire s'appuie sur une loi homme d'une part et une loi femme d'autre part, d'où la nécessité d'avoir la variable sexe. Pour la catégorie socio-professionnelle, un correctif est appliqué dans la tarification selon que la personne soit cadre ou non-cadre. En effet, lors de l'étude réalisée en

2015 qui avait permis de construire les lois d'incidence en invalidité, peu de données étaient disponibles pour réaliser une segmentation propre aux cadres/non-cadres, ce qui explique la présence d'un correctif.

Les variables comportant des dates (naissance, début et fin de contrat ou encore survenance) vont servir à calculer l'exposition de notre portefeuille. Quant aux salaires, ils peuvent être utiles pour déterminer la catégorie socio-professionnelle d'une personne si jamais elle prête à confusion mais nous y reviendrons ultérieurement.

Pour obtenir toutes les informations mentionnées ci-dessus, nous devons recueillir d'une part les sinistres invalidité dans les tables internes d'AXA et d'autre part toutes les autres informations que l'on trouve dans les DSN. En effet, il faut savoir qu'il n'est pas mentionné dans les DSN si la personne est en invalidité, d'où le fait que nous utilisons les tables internes d'AXA. Nous pouvons finalement lier ces deux tables avec l'identifiant sinistre.

2.1 Récupération et retraitements des sinistres invalidité

Depuis les tables internes d'AXA, nous pouvons recueillir l'ensemble des sinistres invalidité survenus entre le 01/01/2017 et le 01/01/2022 avec un identifiant. Cet identifiant est une variable qui résulte de la concaténation de plusieurs éléments :

1. le numéro de contrat de l'entreprise
2. la date de naissance de l'assuré
3. la date de survenance du sinistre
4. la prise en charge ou non de l'arrêt de travail
5. le prénom de l'assuré

Bien que cette maille soit déjà assez fine, il demeure encore des doublons. En effet cette table contient l'historique des provisions mathématiques tête par tête pour les exercices comptables des équipes inventaires, ce qui génère automatiquement des doublons à la maille que l'on a établie. Pour pallier ces doublons et obtenir qu'une ligne corresponde à un seul et unique identifiant et donc à un seul et unique individu, nous avons conservé la ligne associée à l'observation la plus récente.

Une fois cette opération effectuée, nous obtenons une table des sinistres invalidités intégralement dédoublonnée à la maille évoquée.

2.2 Récupération et retraitements des éléments de la DSN

Les déclarations sociales nominatives (DSN) sont largement utilisées depuis 2017. Obligatoires envers les organismes sociaux, tous les employeurs du secteur privé qui paient des salariés remplissent désormais une DSN. C'est une déclaration en ligne produite tous les mois à partir de la fiche de paie. Dans les DSN figurent des informations concernant chacun des salariés. Ces données sont ensuite transmises aux organismes sociaux.

L'idée est de profiter de l'ensemble de ces observations pour compléter les informations au sujet des individus présents en invalidité dans la table des sinistres et ainsi déterminer notre exposition totale. Par informations, on entend notamment les dates de début et de fin des contrats de travail et des contrats d'assurance, mais aussi le salaire, le genre et la catégorie socio-professionnelle : c'est-à-dire toutes les variables que nous avons listées plus haut.

Toutefois, comme déjà évoqué, les DSN ne fournissent pas d'information précise quant à la nature de l'arrêt de travail pour les personnes concernées. C'est pour cette raison que nous avons sélectionné au préalable dans les tables internes les sinistres "invalidité".

Après importation de la base DSN, nous effectuons un premier filtrage en conservant uniquement les contrats prévoyance.

Nous allons maintenant aborder un point très important : le traitement des DSN se fait à la maille identifiant DSN + numéro de RPP (Régime Professionnel de Prévoyance) où la variable "identifiant DSN" résulte de l'anonymisation des individus présents dans les DSN (nom+prénom). Notons que comme nous traitons une table d'incidence, il apparaît peu judicieux d'introduire la variable numéro de RPP dans notre maille. En effet, bien qu'un individu puisse posséder plusieurs numéros de RPP différents, il ne tombe qu'une seule fois en invalidité. Cependant, cette variable est indispensable dans notre maille dans le sens où l'utilité de la table d'incidence à terme est son implémentation dans le processus de tarification. Or, comme nous travaillons sur le périmètre des collectives, la tarification repose justement sur cette notion de numéro de RPP d'où le fait que l'on doit en tenir compte dans notre maille.

Notons toutefois que bien qu'un même individu puisse être compté deux fois comme étant invalide, il sera également compté en double dans notre exposition ce qui ne pose donc pas de souci dans notre construction des taux d'incidence en invalidité (il n'y a pas de phénomène de sur sinistralité). Une fois encore, plusieurs doublons sont apparus lors de la construction de cette table et qui, par la même occasion, ont suggéré des retraitements. Tout d'abord, il a fallu retraiter les dates de début et de fin de contrat de travail ou d'assurance. En effet, de nombreux doublons étaient dus à des plages de temps qui ne coïncidaient pas sur ces

variables. Par prudence, il a été décidé de conserver les dates de début les plus anciennes et les dates de fin les plus larges de façon à avoir une exposition la plus longue possible. Après ces retraitements, il demeure des doublons dus aux salaires, nous avons conservé, là encore par prudence, le salaire le plus élevé.

On obtient ainsi une table finale intégralement dédoublonnée à la maille identifiant DSN + numéro de RPP.

2.3 Jointure des tables et derniers retraitements

Au stade actuel, nous avons à disposition d'une part la table des sinistres invalidités dédoublonnée à la maille d'un identifiant sinistre unique et d'autre part la table dite "de la DSN" dédoublonnée à la maille identifiant DSN + numéro de RPP. L'idée étant de joindre ces deux tables, nous avons utilisé une table "passerelle" qui à un identifiant DSN faisait correspondre un identifiant sinistre. De cette manière nous avons pu obtenir une table à la maille identifiant sinistre + identifiant DSN + numéro de RPP.

Le but de cette jointure est d'avoir d'une part l'ensemble de nos sinistrés que l'on a pu rapprocher des DSN auquel on ajoute tous les individus des DSN mais n'ayant pas été identifiés comme invalides et qui constituent donc notre exposition totale.

En faisant cette jointure, des doublons se sont encore présentés correspondant à deux cas de figure distincts :

1. plusieurs identifiants sinistres pour un même identifiant DSN et un même numéro de RPP : nous avons retenu l'identifiant sinistre associé à la date d'extraction du sinistre la plus récente
2. plusieurs identifiants DSN pour un même identifiant sinistre : nous avons retenu l'identifiant DSN associé à la dernière observation (i.e la date de fin de contrat de travail la plus récente)

De ces retraitements résultent donc une table, issue de la jointure des DSN et de la table des sinistres, entièrement dédoublonnée à la maille identifiant sinistre + identifiant DSN + RPP.

Derniers filtrages de la table finale obtenue

Dans la table obtenue à la dernière étape, nous avons donc tous les individus présents dans les DSN avec un contrat prévoyance peu importe qu'ils aient subi une invalidité ou non. Ce périmètre est un peu trop large dans le sens où avoir un contrat prévoyance ne signifie pas nécessairement avoir un contrat qui couvre la garantie invalidité. Nous devons donc restreindre notre périmètre pour n'avoir que les individus dont le contrat couvre cette garantie.

Une fois cette action effectuée, nous avons enfin notre table complète soit un total d'environ 3 000 000 individus.

Bien que les ordres de grandeurs soient respectés, les chiffres exacts ne sont pas communiqués par souci de confidentialité.

Retraitements des cadres et des non-cadres

Comme évoqué précédemment, il existe dans les DSN une variable CSP qui prend les modalités suivantes associées au libellés suivants :

	CSP	first(LIBELLE_CSP)
1	1	AGRICULTEURS
2	2	ARTISANS, COMMERÇANTS ET CHEFS D'ENTREPRISE
3	3	CADRES, PROFESSIONS INTELLECTUELLES SUPERIEURES
4	4	PROFESSIONS INTERMEDIAIRES
5	5	EMPLOYES
6	6	OUVRIERS
7	9	AUTRES

FIGURE 9 – Table de correspondance

Globalement, les catégories évoquées sont assez explicites quant à la répartition des cadres et des non-cadres. Toutefois, il apparaît que la catégorie 4 dite "professions intermédiaires" est constituée à la fois de cadres et de non-cadres. Pour décider à laquelle de ces deux catégories appartenaient les individus de la CSP 4, nous avons choisi le critère du salaire (variable que nous avons également retenue). Autrement dit, nous avons effectué un choix complètement arbitraire selon lequel pour un salaire supérieur à un Plafond Annuel de la Sécurité Sociale (PASS), une personne est considérée comme étant cadre et non-cadre dans le cas contraire.

Toutefois, si le salaire n'était pas renseigné, nous avons choisi pour chaque numéro de RPP de mettre le salaire moyen à partir des salaires renseignés du même numéro de RPP. Dans le cas où aucun salaire n'était renseigné pour un même numéro de RPP, nous avons mis le salaire moyen de la CSP en question obtenu à partir de tous les salaires renseignés.

Une fois cette étape réalisée, l'intégralité des individus sont considérés comme cadres ou non-cadres selon la maille CSP + salaire.

3 Statistiques descriptives

3.1 Statistiques descriptives de la population en portefeuille

On entend par population en portefeuille tous les individus présents dans les DSN et dont le contrat couvre l'invalidité (et non tous les individus présents dans les DSN) : c'est-à-dire nos 3 000 000 d'individus environ.

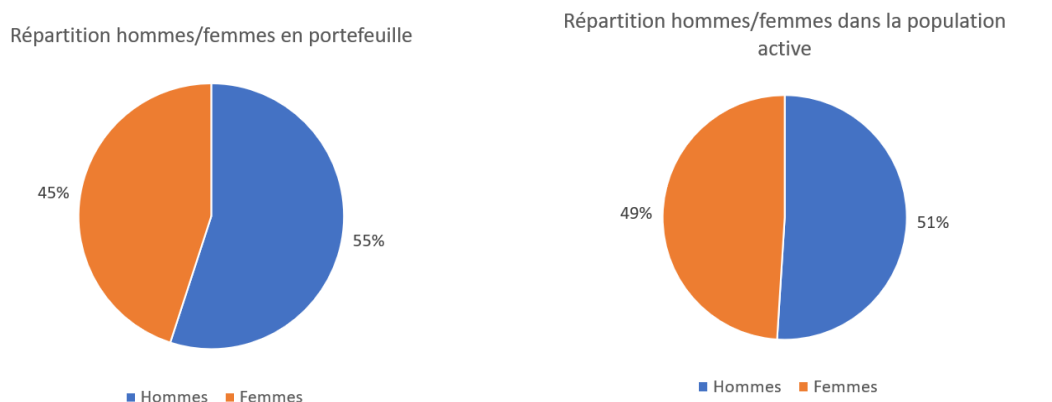
Parmi ceux-là on retrouve environ 55% d'hommes et 45% de femmes, ainsi que 33% de cadres et 67% de non-cadres.

Ces chiffres sont éloignés de ceux fournis par l'INSEE relatifs à l'ensemble de la population française active :

51% d'hommes pour 49% de femmes et 19% de cadres pour 81% de non-cadres. Cet écart assez conséquent reflète le fait qu'AXA possède en portefeuille une part relativement importante de personnes cadres.

Ces différences sont représentées ci-dessous pour mieux les visualiser :

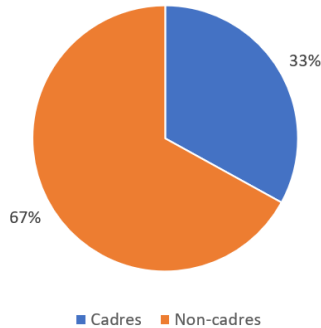
Dans un premier temps, la répartition selon le sexe :



(a) Répartition des hommes et des femmes en portefeuille (b) Répartition des hommes et des femmes dans la population active

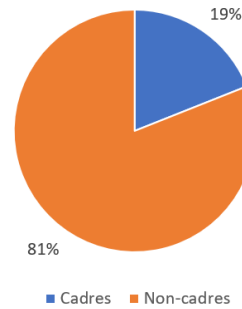
Dans un second temps, la répartition selon la CSP :

Répartition cadres/non-cadres en portefeuille



(a) Répartition des cadres et des non-cadres en portefeuille

Répartition cadres/non-cadres dans la population active



(b) Répartition des cadres et des non-cadres dans la population active

3.2 Statistiques descriptives de la population sinistrée

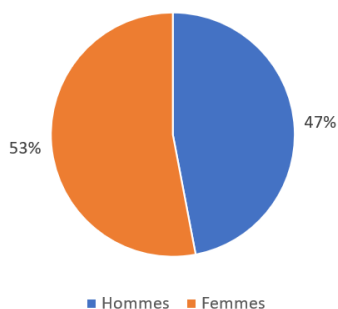
Nous faisons maintenant un focus sur notre portefeuille en considérant seulement les personnes en invalidité. Parmi celles-ci, nous retrouvons 47% d'hommes et 53% de femmes soit une répartition différente de celle pour la totalité du portefeuille. Au niveau de la segmentation socio-professionnelle, nous retrouvons dans la population sinistrée 20% de cadres et 80% de non-cadres. Notons qu'intuitivement, il apparaît cohérent que la proportion de non-cadres en invalidité soit plus importante que la proportion de non-cadres en portefeuille car les non-cadres semblent davantage exposés au risque d'invalidité que les cadres.

Quant aux chiffres fournis par l'INSEE sur les personnes en invalidité, ce sont 51% d'hommes et 49% de femmes ainsi que 20% de cadres et 80% de non-cadres.

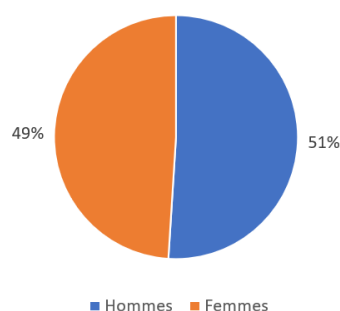
Nous consignons aussi les résultats obtenus dans les diagrammes circulaires ci-dessous avant de livrer notre interprétation :

Dans un premier temps, la répartition selon le sexe :

Répartition invalides hommes/femmes en portefeuille



Répartition invalides hommes/femmes dans la population active

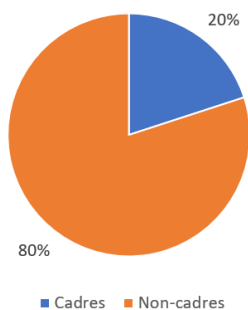


(a) Répartition des hommes et des femmes en portefeuille

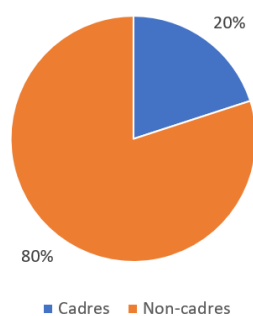
(b) Répartition des hommes et des femmes dans la population active

Dans un second temps, la répartition selon la CSP :

Répartition invalides cadres/non-cadres en portefeuille



Répartition invalides cadres/non-cadres dans la population active



(a) Répartition des cadres et des non-cadres en portefeuille

(b) Répartition des cadres et des non-cadres dans la population active

Plusieurs remarques par rapport à ces résultats :

1. Dans la population active, il n'y a pas de différences de répartition entre les personnes sinistrées et celles qui ne le sont pas alors qu'on observe à AXA des différences de répartition (plus de femmes sinistrées et plus de non-cadres sinistrés). Ceci peut éventuellement s'expliquer par un "effet volume", en effet les statistiques de l'INSEE reposent sur l'étude de la population active soit environ 28 000 000 d'individus tandis que, comme évoqué auparavant, le portefeuille d'AXA relatif à la garantie invalidité est d'environ 3 000 000 d'individus.
2. Le fait que les non-cadres soient davantage représentés en invalidité qu'en portefeuille semble logique étant donné les risques liés à leurs activités de manière générale.

Ayant étudiés successivement les répartitions de notre portefeuille et de la population active selon le sexe et la CSP et ce également pour les personnes en invalidité, nous allons maintenant nous intéresser aux âges des personnes en invalidité.

En portefeuille, la moyenne d'âge des personnes en invalidité au moment de leur entrée en invalidité est de 51,5 ans environ.

Si l'on s'intéresse à la répartition des âges :

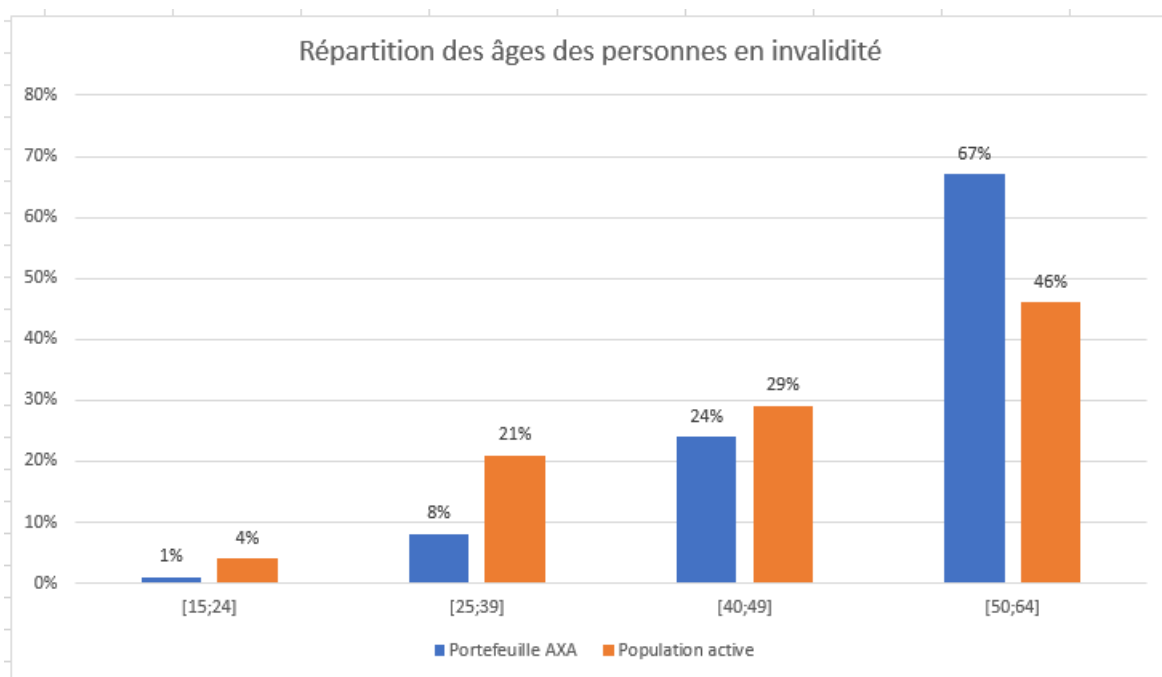


FIGURE 14 – Comparaison des répartitions des invalidités en portefeuille et dans la population active par tranches d'âge

Bien que la proportion de sinistrés augmente avec l'âge naturellement dans les deux cas, nous pouvons voir que la population des sinistrés est bien plus âgée dans le portefeuille AXA. Ceci peut en partie s'expliquer par la présence plus importante de femmes.

Enfin, à titre informatif, en termes de volume total ce sont un peu plus de 16 000 invalidités qui constituent le portefeuille d'invalides AXA sur la plage étudiée.

Maintenant que nous avons décortiqué la population constituant nos données, nous allons établir les taux bruts d'entrée en invalidité par différentes méthodes puis les analyser.

4 Estimation des taux bruts d'entrée en invalidité

4.1 Analyse de survie : Censure et troncature

Nous possédons désormais une table intégralement retraitée et prête à l'utilisation sur laquelle nous allons appliquer différents modèles pour déterminer nos taux bruts d'incidence en invalidité. Avant d'entrer dans le détail des modèles, il convient de détailler dans un premier temps le cadre de ces derniers associé à l'analyse de survie et notamment aux notions de censure et de troncature.

L'analyse de survie est une science utile en assurance notamment mais aussi en biostatistiques (efficacité d'un traitement) ou encore dans n'importe quel processus de fiabilité.

Tout d'abord commençons par évoquer la durée de survie : il s'agit de la période de temps écoulé avant qu'un évènement se produise (dans notre cas la survenance de l'invalidité). L'objectif est d'étudier la distribution des temps de survie qu'on appelle fonction de survie.

Mathématiquement, la fonction de survie est définie comme suit :

$$S(t) = P(X > t), \text{ où } t \text{ est positif ou nul et } X \text{ désigne la durée de survie}$$

Nous reviendrons ultérieurement et de manière plus approfondie sur l'aspect théorique de la construction de la fonction de survie.

Comme notre étude est bornée dans le temps (entre 2017 et 2022), nous avons forcément des données incomplètes à cause des effets de censure et de troncature que nous allons expliquer.

Commençons par la censure qui est le plus intuitif des deux phénomènes rencontrés en analyse de survie :

Cas de la censure à droite

Un individu est dit "censuré à droite" s'il n'a pas subi l'évènement lors de la dernière observation, auquel cas sa véritable durée de survie n'est pas observée. Soit X_i la survenance en invalidité d'un individu i et C_i l'observation censurée du même individu i , nous étudions en réalité $T_i = \min(X_i, C_i)$. L'observation peut être censurée pour diverses raisons comme la rupture du contrat d'assurance pendant la période d'observation, la survenance de l'invalidité après la date de fin de l'étude ou encore la défaillance du système informatique. Ce phénomène de censure à droite est représenté de manière importante dans notre cas du fait de l'arrêt de l'étude au 01/01/2022.

Nous joignons ci-contre le schéma permettant de bien saisir l'enjeu lié à cette notion :

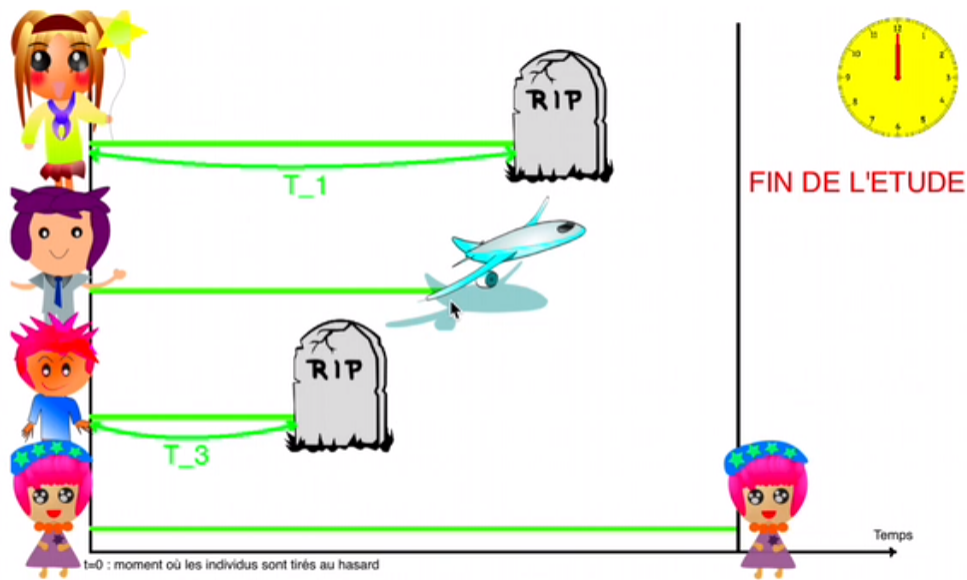


FIGURE 15 – Schéma représentant le phénomène de censure

[11]

Avec ce schéma, nous pouvons voir que les individus 1 et 3 correspondent à des informations complètes. Quant aux deux autres, on pourrait très bien imaginer par exemple que l'individu 2 ait rompu son contrat d'assurance mais qu'il ait subi une invalidité avant la fin de l'étude sans que nous n'ayons eu cette information. L'individu 4 représente un cas probablement très fréquent d'individus qui n'ont pas subi d'invalidité sur la plage d'étude et

dont on ne connaît donc pas la survenance de l'invalidité ou non : c'est le cas de censure à droite.

Notons un point très important, nous considérons que X_i et C_i sont deux éléments complètement indépendants ce qui semble une hypothèse assez naturelle. En effet, cette hypothèse d'identification va être très importante par la suite d'un point de vue théorique pour pouvoir réaliser les calculs.

Cas de la censure à gauche

Un individu est dit "censuré à gauche" s'il a subi l'évènement avant d'être observé. Dans notre cas cela signifierait être tombé en invalidité avant le 01/01/2017 mais tout de même être observé en portefeuille par la suite ce qui semble improbable. Ces cas de censure à gauche sont en général assez marginaux et restent beaucoup moins fréquents que ceux de censure à droite.

Enfin comme nous disposons pour chacun de nos sinistrés la date de survenance du sinistre, nous excluons de fait la censure par intervalles.

Cas de la troncature à gauche

On assiste également à un phénomène de troncature lorsqu'on conditionne notre échantillon. Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui-même. Ainsi, une variable X est tronquée par un sous ensemble éventuellement aléatoire A de \mathbb{R}_+ si au lieu de X , on observe X uniquement si X appartient à A . La troncature diffère de la censure dans le sens où elle résulte d'un biais de sélection et également dans le sens où nous sommes incapables de quantifier le nombre d'observation tronquées (alors que nous connaissons le nombre d'observations censurées).

En l'occurrence dans notre cas précis, les individus étant entrés en invalidité avant le 1er janvier 2017 et que l'on ne parvient pas à retrouver dans notre portefeuille ont été exclus de fait de notre étude et constituent donc en cela un certain biais.

Nous introduisons une fois de plus un schéma facilitant la compréhension du phénomène de troncature à gauche :

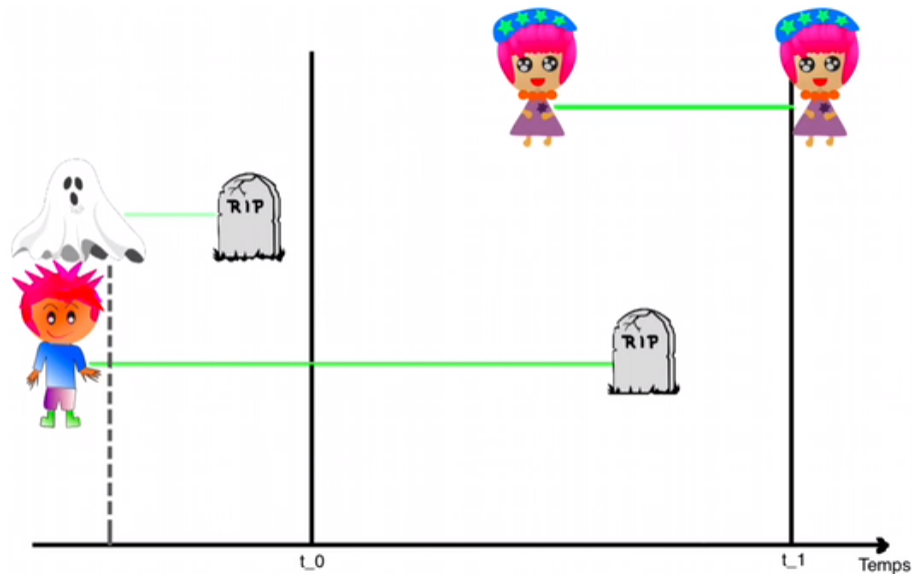


FIGURE 16 – Schéma représentant le phénomène de troncature

[11]

Ici, le premier individu correspond à un phénomène de censure à droite, le second à un phénomène de troncature à gauche (du fait de la survenance de l'évènement avant le début de l'étude) et le troisième correspond à une observation complète.

Notons que ces deux phénomènes doivent être absolument pris en compte dans le calcul des taux. En effet, nous ne pouvons pas les ignorer et construire notre table d'incidence uniquement sur les informations complètes sans quoi, les durées de survie seraient faussées et c'est précisément ce que nous allons voir en introduisant les modèles de Kaplan-Meier et de Hoem.

4.2 L'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur non-paramétrique dans le sens où il ne prend quasiment pas d'hypothèses sur la distribution. En réalité, s'il ne tendait pas à corriger les phénomènes de censure et de troncature évoqués précédemment, l'estimateur de Kaplan-Meier se résume tout simplement à une fonction de répartition. C'est pour cela qu'il très utilisé en analyse de survie car il permet notamment de tenir compte des phénomènes de censures à droite voire même de troncatures à gauche qui nous concernent justement.

En effet, bien que nous allons revenir dessus par la suite, rappelons que nous étudions non pas X_i directement à savoir l'observation d'entrée en invalidité d'un individu mais bien $T_i = \min(X_i, C_i)$ où C_i désigne une observation censurée. La quantité T_i est associée directement à l'indicatrice de $(X_i \leq C_i)$ qui vaut 1 si l'observation est complète ou 0 si elle est censurée.

Nous avons donc à notre disposition de manière assez directe la fonction de répartition de T_i mais également les fonctions de "sous-répartition" de T_i conditionnellement au fait que l'indicatrice vaille 1 ou 0. Ces fonctions de "sous-répartition", bien que croissantes, n'atteignent pas la valeur 1 du fait de leur conditionnement. Notons que leur somme vaut justement la fonction de répartition de T_i

Ces fonctions mentionnées ne nous intéressent pas directement mais l'enjeu de l'estimateur de Kaplan-Meier est d'établir une fonction (discrète) de ces trois éléments qui permettrait de tenir compte du phénomène de censure contenu dans la variable T_i .

C'est pour cela justement que nous déterminons une fonction de survie et de manière assez logique on en déduit le taux brut d'incidence en invalidité puisque ce taux vaut $1-S(t)$ où $S(t)$ est la fonction de survie. Ici, le terme "fonction de survie" désigne en réalité ("la fonction de ne pas tomber en invalidité").

Nous faisons alors deux hypothèses :

1. les entrées en invalidité sont indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre q_k où q_k désigne la probabilité de tomber en invalidité pour l'individu k . Ceci implique immédiatement que i_k à savoir le nombre d'entrées en invalidité suit une loi binomiale de paramètre n_k et q_k assez logiquement.
2. la variable aléatoire X_i est complètement indépendantes de la variable aléatoire C_i . En effet, par exemple, les entrées en invalidité sont indépendantes de la date de sortie de l'étude (qui est une cause de censure courante). Cette hypothèse est appelée hypothèse d'identification. Pour détailler son utilisation, elle permet notamment lors du calcul de $P(T_i > t)$ de scinder l'intersection des probabilités en produit des probabilités :
$$P(T_i > t) = P(\min(X_i, C_i) > t) = P(X_i > t \text{ et } C_i > t) = P(X_i > t)P(C_i > t)$$

Avant d'introduire quelques notations, nous allons expliquer brièvement le principe de la construction de la fonction de survie :

L'idée est la suivante : soient t_1, t_2 et t_3 tels que $t_1 < t_2 < t_3$. Ne pas être tombé en invalidité au temps t_3 , c'est ne pas être tombé en invalidité aux temps t_2 et t_1 .

Autrement dit, mathématiquement parlant, cela signifie que :

$$\begin{aligned} P(X > t_3) &= P(X > t_2, X > t_3) \\ &= P(X > t_3|X > t_2) * P(X > t_2) \text{ par la formule de Bayes} \\ &= P(X > t_3|X > t_2) * P(X > t_2|X > t_1) * P(X > t_1) \text{ et ainsi de suite...} \end{aligned}$$

Remarquons que le raisonnement évoqué s'appuie indirectement sur la notion de taux instantané qui est la grandeur phare en analyse de survie, grandeur à partir de laquelle il est possible de retrouver la fonction de répartition voire la densité dans le cas de distributions continues. En s'inspirant de cette idée, nous posons les notations suivantes :

1. a_k les âges auxquels les assurés entrent en invalidité ou bien auxquels ils sortent de l'étude (ceux auxquels ils sont censurés), c'est à dire l'âge correspondant à $Ti = \min(Xi, Ci)$ déjà évoqué plus haut et on considère ces âges rangés par ordre croissant pour pouvoir ensuite appliquer le théorème de Bayes dans l'esprit des temps t_1, t_2 et t_3 mentionnés juste avant.

Il est important que les " a_k " soient rangés dans l'ordre croissant, d'un point de vue théorique on parle alors de statistique d'ordre : la statistique d'ordre de rang k d'un échantillon statistique est égal à la k -ième plus petite valeur. C'est un outil fondamental dans le domaine de la statistique non-paramétrique.

2. i_k le nombre d'entrée en invalidité à l'âge a_{k+1}
3. c_k le nombre de censures à droite entre les âges a_k et a_{k+1}
4. t_k le nombre de censures à gauche à l'âge a_{k+1}
5. n_k le nombre d'assurés présents en portefeuille à l'âge a_k : notons au passage la relation de récurrence suivante : $n_{k+1} = n_k + t_k - c_k - i_k$.

Une fois ces quelques éléments présentés, il vient assez naturellement que la fonction de survie $S(a_k) = P(X > a_k)$ vaut par récurrence : $(1 - q_k)S(a_k - 1) = \prod(1 - q_i)$ pour i qui va de 1 à a_k .

En remplaçant justement q_k par son estimateur du maximum de vraisemblance à savoir $\frac{i_k}{n_k}$, on obtient que $(1 - q_k)S(a_k - 1) = \prod(1 - \frac{i_j}{n_j})$ pour j qui va de 1 à a_k .

De là découle alors directement notre estimateur de taux final qui vaut $1-S(a_k)$ soit $q_k = 1 - \prod(1 - \frac{i_j}{n_j})$

Dans notre cas, nous avons appliqué le modèle de Kaplan-Meier uniquement pour le cas de la censure à droite (qui est le modèle "classique" de cet estimateur, ce pour quoi il a été conçu). Notons tout de même qu'il peut être étendu au cas de la troncature à gauche grâce aux travaux de Wang M.-C., Jewell N.P. and Tsai W.-Y. en 1986.

Cette extension du modèle demande également des hypothèses supplémentaires parmi lesquelles l'hypothèse phare d'identification selon laquelle la distribution des troncatures est indépendante de celle des censures et des observations complètes.

Par ailleurs, nous noterons sans rentrer dans les détails que l'estimateur de Kaplan-Meier donne une place prépondérante aux observations complètes et compense plutôt bien les censures en milieu de distribution (dues souvent à une rupture du contrat d'assurance) mais est plus limité pour décrire la queue de distribution à droite en raison du nombre très élevé de censures. En effet, de nombreuses observations sont censurées du fait de l'arrêt de l'étude à une date fixe et du caractère relativement rare de la survenance d'une invalidité. Ces notions de propriétés de cet estimateur reposent entre autres sur les concepts de martingales avec des processus de comptage que nous n'aborderons pas ici.

4.3 L'estimateur des moments de Hoem

L'estimateur de Hoem est tout comme celui de Kaplan-Meier un estimateur non-paramétrique, peut-être moins connu que celui de Kaplan-Meier, car appliqué quasiment exclusivement dans le domaine de l'assurance sur des problématiques telles que la construction de tables de mortalité ou d'invalidité par exemple. Il prend également en compte les censures à droite et corrige ainsi une partie du biais engendré par les observations incomplètes. Cet estimateur est tout simplement l'estimateur du maximum de vraisemblance et sa construction applique donc la méthode classique qui suit. En considérant les hypothèses selon lesquelles les entrées en invalidité sont indépendantes et identiquement distribuées et réparties linéairement entre deux âges, et en posant les notations suivantes :

1. i_k le nombre d'entrée en invalidité entre les âges a_k et a_{k+1}
2. d_k et f_k les bornes de l'exposition de l'assuré k entre les âges a_k et a_{k+1}
3. n_k le nombre d'individus entre les âges a_k et a_{k+1}

Nous avons l'entrée en invalidité d'un individu k qui suit une loi de Bernoulli de paramètre $(f_k-d_k)q_k$ et les i_k qui suivent donc une loi binomiale de paramètres n_x et $(f_k-d_k)q_k$.

Nous tâchons ensuite de maximiser la vraisemblance de cette loi. Pour cela, nous passons par la log-vraisemblance dans un premier temps avant de dériver puis déterminer le maximum de notre fonction en vérifiant au passage son caractère concave.

On obtient alors une expression de l'estimateur de $q_k = \frac{1}{\sum_{j=1}^{n_k} (f_j - b_j)}$.

Avant de passer à l'exposition des résultats bruts, nous soulignerons que l'hypothèse selon laquelle les invalidités sont réparties linéairement entre deux âges peut être discutable.

4.4 Résultats

Nous allons dans cette partie exposer les résultats obtenus avec les deux estimateurs évoqués, les comparer et sélectionner l'un des deux en justifiant notre choix. Il faut avant tout savoir que les lois d'incidence en invalidité implantées dans l'outil de tarification sont les lois femmes et hommes. Nous allons donc successivement présenter pour chacun des deux estimateurs les lois de la population globale, de la population femmes et de la population hommes.

Les taux affichés ne sont pas les taux réellement obtenus par souci de confidentialité mais respectent les ordres de grandeurs et les positions des courbes les unes par rapport aux autres ou des nuages de points les uns par rapport aux autres.

Avant de comparer nos deux estimateurs, tâchons de voir déjà comment notre exposition et nos invalidités sont réparties selon les âges allant de 20 à 60 ans :

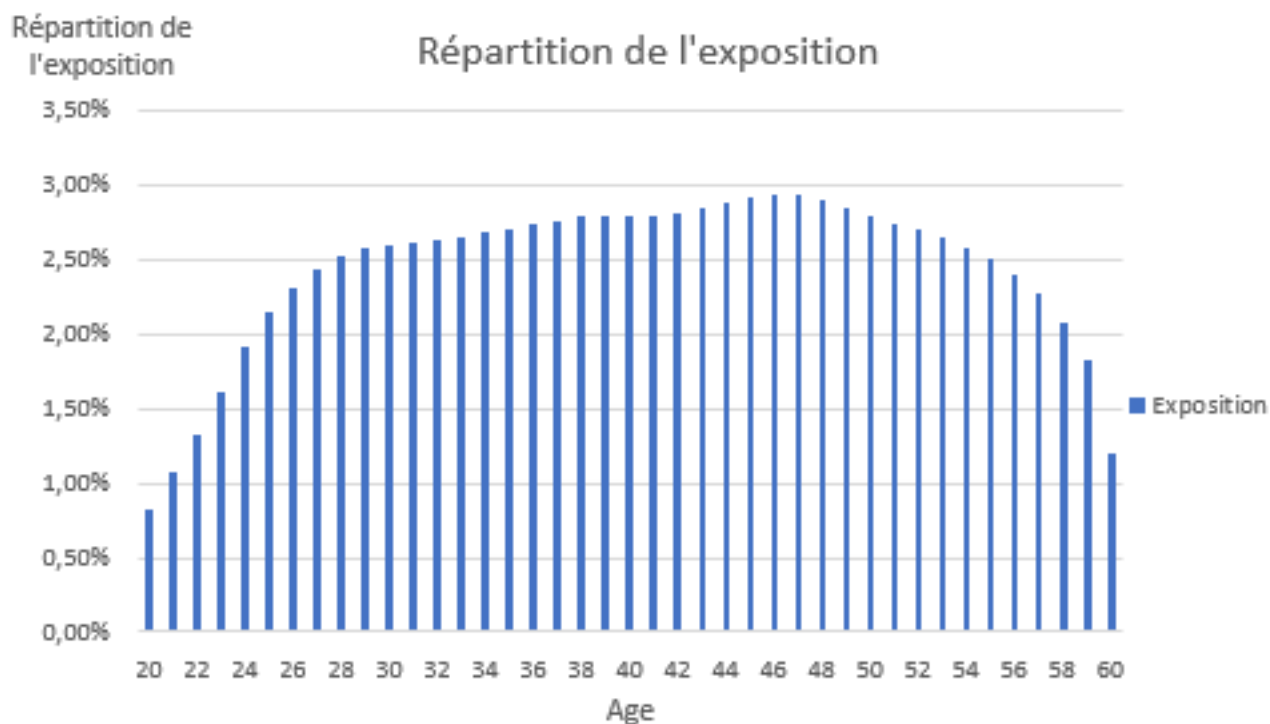


FIGURE 17 – Répartition de l'exposition en fonction de l'âge

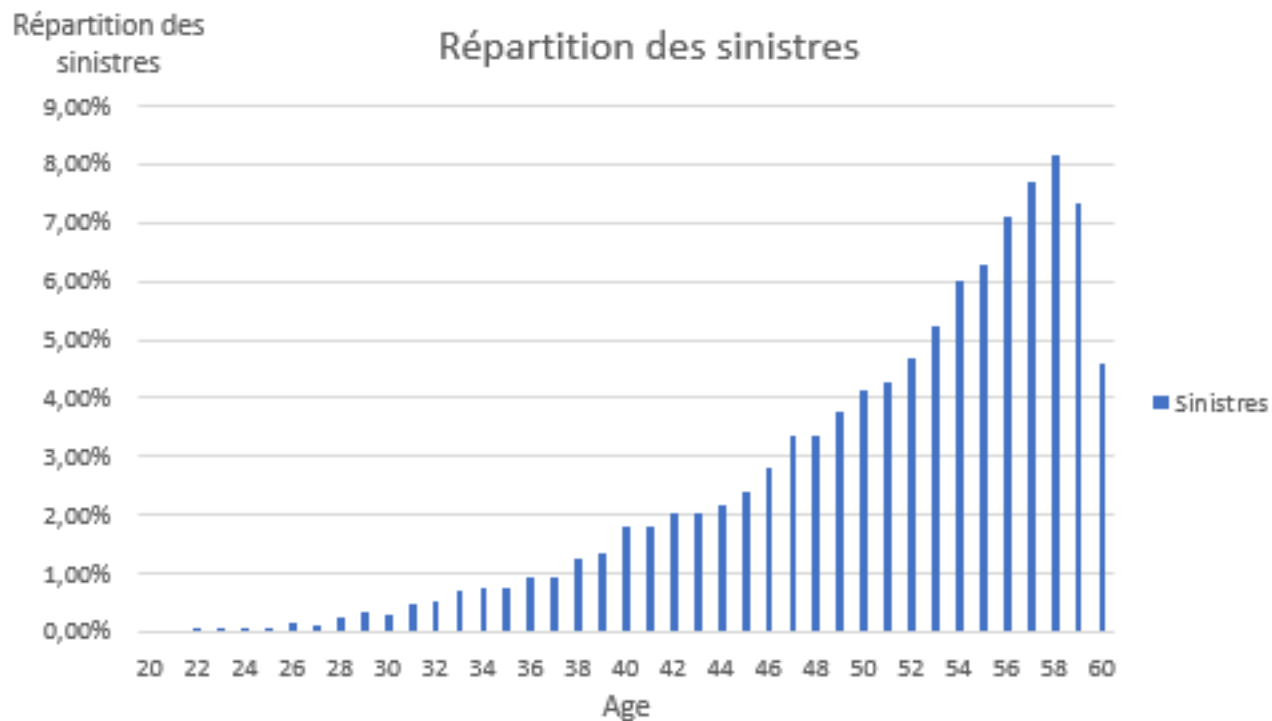


FIGURE 18 – Répartition des sinistres en fonction de l'âge

Nous pouvons voir que l'exposition est majoritairement répartie sur un plateau s'étendant entre 28 et 54 ans soit là où l'on trouve logiquement le plus d'actifs et concernant nos sinistrés ils augmentent avec l'âge avant de chuter très brusquement à l'approche de la retraite.

Encore une fois cette baisse de sinistrés sur les âges élevés pourrait évoluer avec un âge de départ à la retraite repoussé.

Comparons désormais les taux obtenus avec nos deux estimateurs pour la population totale :

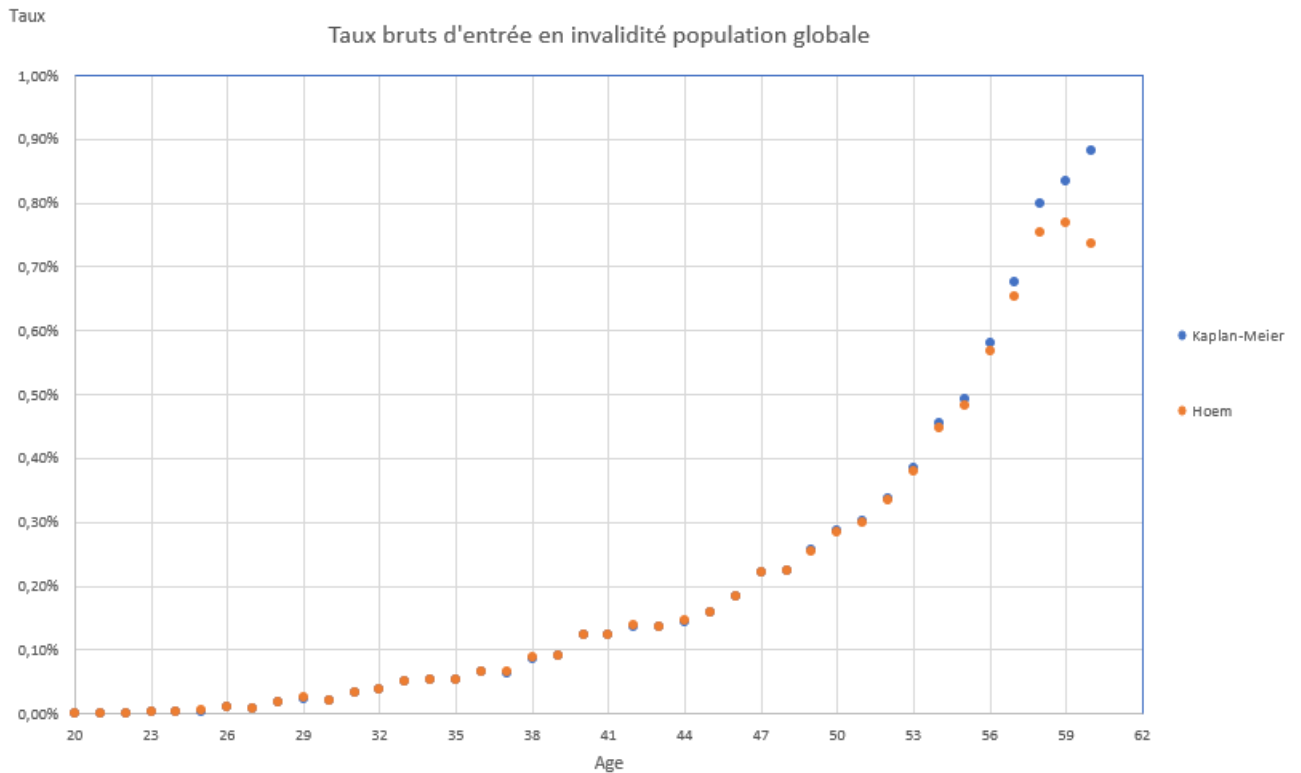


FIGURE 19 – Taux bruts d'entrée en invalidité pour l'ensemble de la population avec les estimateurs de Kaplan-Meier et de Hoem

Nous pouvons d'ores et déjà remarquer que les nuages de points des deux estimateurs sont très proches l'un de l'autre et atteignent un pic vers des âges élevés (plus de 57 ans) ce qui semble cohérent. Quant aux positions relatives des nuages, celui représentant l'estimateur d'Hoem est légèrement au-dessus jusqu'à l'âge de 57 ans puis c'est ensuite l'inverse entre 58 et 60 ans. Ceci s'explique principalement par le fait que l'estimateur de Kaplan-Meier est déterminé à l'aide d'une fonction de survie comme nous l'avons vu lors de sa construction qui se fait par récurrence. Par conséquent, comme la probabilité de rester en vie à un âge dépend de tous les âges précédents, il y a un phénomène de "temps de retard" dans cet estimateur qui explique qu'aux alentours de l'âge de 60 ans la chute des taux est perçue avec un léger décalage. Ceci est normal puisque le départ anticipé à la retraite est possible entre 56 ans et 58 ans, cela coïncide donc avec la baisse d'exposition également.

Nous avons regardé les mêmes données restreintes aux populations femmes puis hommes

et mis les graphiques associés dans la section annexe.

Nous pouvons sensiblement faire les mêmes remarques que pour la population globale concernant le plateau entre 28 et 54 ans pour l'exposition ainsi que la croissance des sinistres avec l'âge bien que cette dernière soit légèrement moins régulière.

Quant aux nuages de points, ils sont sensiblement identiques sauf sur la plage s'étendant de 57 ans à 60 ans et ce, pour les mêmes raisons que celles relatives à la population globale.

Après avoir vu ces trois populations sous la lumière de l'estimateur de Kaplan-Meier et de l'estimateur de Hoem avec à chaque fois les mêmes interprétations, il nous est apparu plus judicieux de préférer l'estimateur de Hoem car, du fait de taux plus élevés entre 20 et 57 ans, il semble plus prudent. De plus, l'importante baisse de l'exposition vers 60 ans due vraisemblablement à l'approche de la retraite semble biaiser quelque peu l'estimateur de Kaplan-Meier comme expliqué plus haut. Pour encore mieux comprendre ce phénomène, nous allons superposer la courbe de nos taux bruts obtenus par l'estimateur de Hoem avec celle de l'exposition pour les trois populations concernées.

Nous montrons ci-dessous le graphique correspondant à la population globale et mettons en annexe les graphiques pour les populations femmes et hommes. Notons tout de même que nous observons les mêmes phénomènes pour ces deux dernières populations.

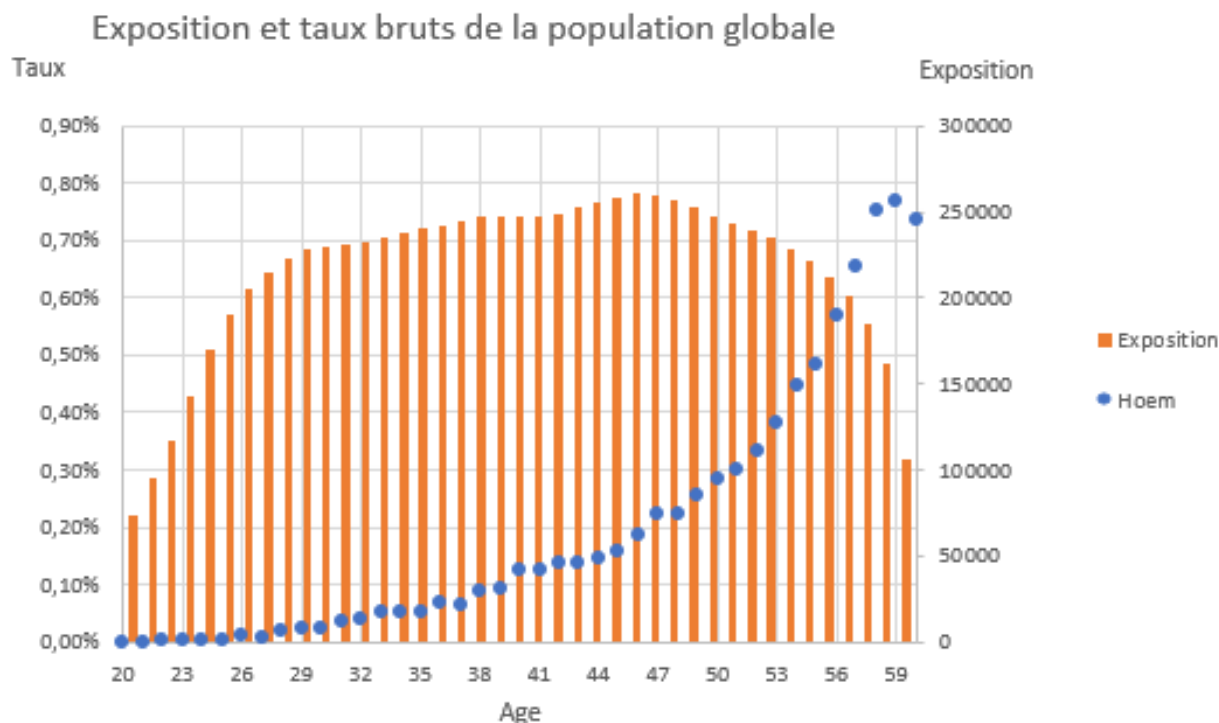


FIGURE 20 – Taux bruts estimés par Hoem et exposition en fonction de l'âge pour l'ensemble de la population

En somme, nous pouvons voir que la chute légère des taux s'accompagne de la fin du "plateau" que l'on observait sur la variable exposition.

Maintenant que nous avons notre estimateur, nous allons lisser les taux bruts par différentes méthodes de lissage avant d'en sélectionner une en justifiant notre choix par des arguments statistiques.

5 Lissage des taux

Pour chaque type de lissage, nous l'appliquons aux taux bruts femmes et hommes obtenus par l'estimateur de Hoem. Nous ne l'appliquons plus à la population globale car la loi associée n'est pas implantée dans l'outil de tarification.

Dans un premier temps nous exposerons les résultats des lissages puis dans un second temps nous présenterons les tests statistiques permettant de valider ou d'invalidier les lissages et par la même occasion de choisir le lissage qui nous donnera les taux "définitifs".

Notons aussi que l'ensemble des lissages ont été réalisés sous Excel avec des constructions de matrices adaptées pour chaque lissage lorsque cela était nécessaire.

Enfin, nous rappellerons que les taux présentés ne sont pas les taux réels par souci de confidentialité mais conservent bien les ordres de grandeurs.

5.1 Lissage par splines cubiques

Le principe du lissage par splines cubiques est d'approximer la courbe des taux bruts par un polynôme de degré 3. Plus précisément, le lissage par splines cubiques associe un polynôme de degré 3 à un nombre de segments souhaités.

Soit n le nombre de segment choisi. Un polynôme de degré 3 étant défini par 4 paramètres, il faut donc déterminer $4n$ paramètres...

De plus, il faut que l'on assure la continuité de la courbe aux points où commence un autre intervalle. Cela implique que chaque fonction polynomiale doit admettre une dérivée première ainsi qu'une dérivée seconde. Usuellement, nous utilisons des coefficients c_0 , c_1 , c_2 qui résument les conditions mentionnées :

1. c_0 : chacun des n polynômes est jointif au suivant ce qui représente $n-1$ conditions
2. c_1 : les polynômes ont la même dérivée première aux $n-2$ points de jonctions soit $n-1$ conditions
3. c_2 : les polynômes ont la même dérivée seconde aux $n-2$ points de jonctions soit $n-1$ conditions

Ceci fait un total de $3n-3$ conditions.

Pour faciliter la compréhension, appliquons cette théorie à notre cas. En admettant que l'on ait constitué ces n intervalles de la façon suivante $[a_0 ; a_1] ; [a_1 ; a_2] \dots [a_{n-2} ; a_{n-1}]$ où les " a_k " désignent les âges entiers entre 20 et 60 ans, on veut que $\hat{q}_x = p_k(x)$ si x appartient à $[a_{k-1} ; a_k]$, où $p_k(a)$ est le polynôme de degré 3 correspondant.

Ainsi on a les $3n-3$ conditions suivantes pour tout k :

1. $c_0 : p_k(a_k) = p_{k+1}(a_k)$
2. $c_1 : p'_k(a_k) = p'_{k+1}(a_k)$
3. $c_2 : p''_k(a_k) = p''_{k+1}(a_k)$

Les coefficients sont ensuite déterminés à partir de la minimisation de la fonction suivante :

$$M = \sum_{i=20}^{60} w_x (q_x - \hat{q}_x)^2$$

Ceci se résout de manière matricielle en faisant intervenir la matrice des poids w_x . C'est une matrice diagonale où nous avons décidé que chaque point de la diagonale valait l'exposition de l'âge correspondant sur l'exposition totale. Cette matrice des poids est construite arbitrairement : on peut aussi choisir par exemple d'affecter le même poids à chacun des âges (soit $\frac{1}{41}$ dans notre cas).

La difficulté réside dans la segmentation de notre intervalle s'étendant de 20 ans à 60 ans. En effet, il faut que les segments soient suffisamment grands (plus de 5 observations sinon les taux lisses correspondent exactement aux taux bruts) mais suffisamment petits pour distinguer des comportements distincts sur des segments distincts. En l'occurrence pour le cas qui nous préoccupe, la pente ne change jamais brusquement hormis à partir de 45 ans environ où le nombre de sinistres s'élève davantage. Il faudrait donc arriver à discerner une première tendance sur les âges plus faibles et une seconde tendance sur les âges plus élevés. La tendance qu'ont les taux à baisser à partir de 57 ans ne pourra pas être détectée par le lissage des splines cubiques et la création d'un troisième segment du fait que ce segment est justement trop court (3 points) et que donc la création d'un polynôme de degré 3 à partir de 3 observations n'a pas vraiment de sens (il faut au moins 5 points). C'est pour cette raison que nous avons choisi de limiter notre lissage à 2 segments :

1. l'un s'étendant de 20 ans à 45 ans
2. l'autre s'étendant de 46 ans à 60 ans

Nous obtenons alors les lissages suivants pour les populations femmes et hommes :

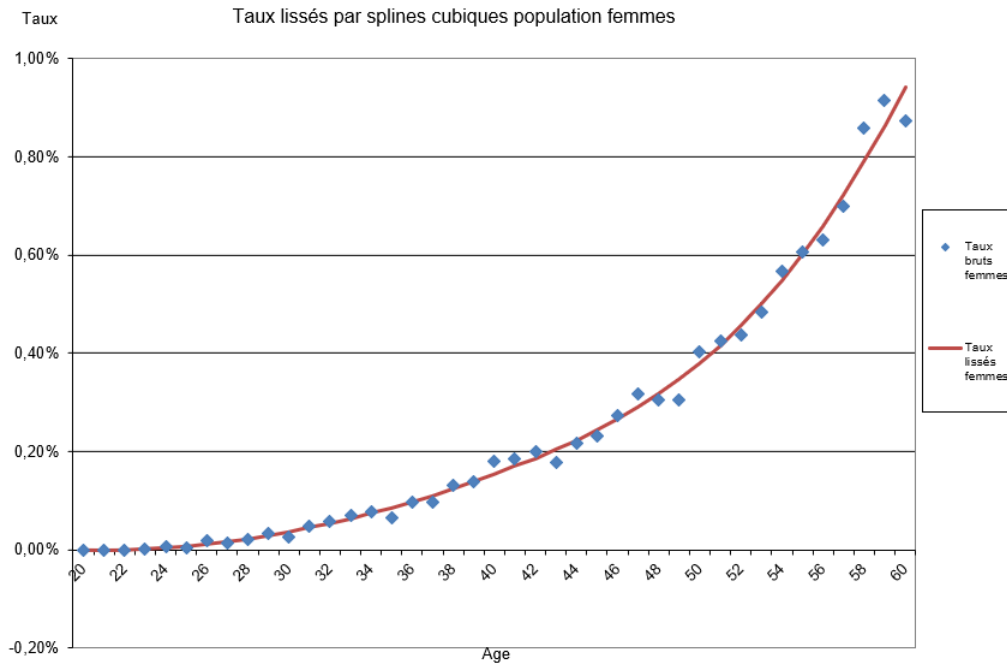


FIGURE 21 – Lissage par splines cubiques pour la population femmes

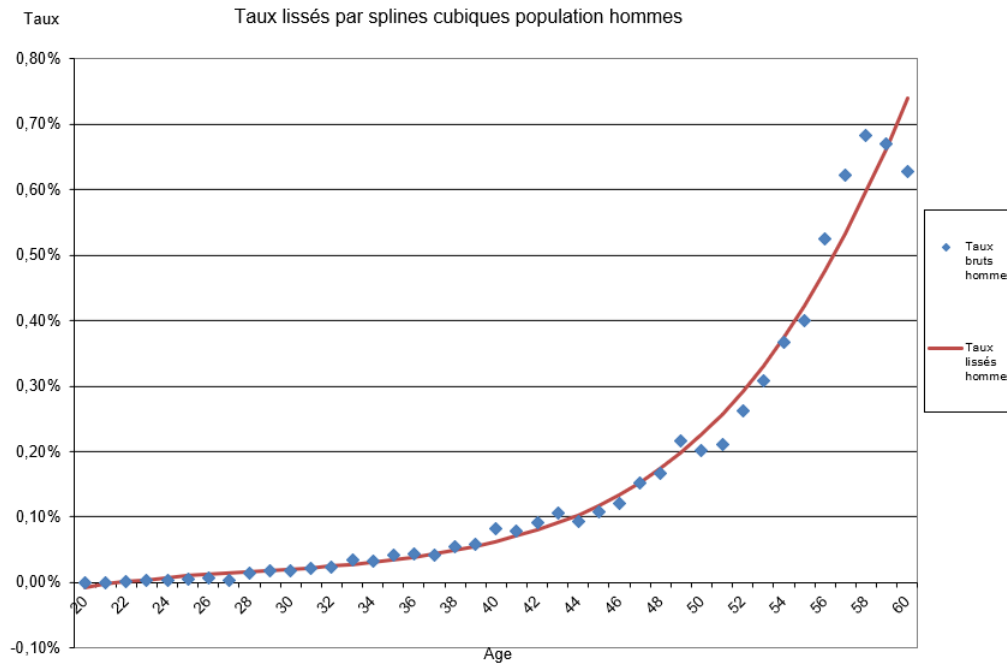


FIGURE 22 – Lissage par splines cubiques pour la population hommes

Comme déjà dit plus haut, nous évoquerons ultérieurement la qualité du lissage des splines cubiques mais nous noterons tout de même qu'un inconvénient majeur de ce lissage est que, dans le cas de figure où l'on a un changement de pente que l'on peut qualifier d'important entre deux intervalles successifs, l'estimation des paramètres du polynôme devient compliquée pour satisfaire les hypothèses de continuité et de dérivabilité et le lissage s'en retrouve impacté.

Nous allons maintenant présenter le lissage de Whittaker-Henderson.

5.2 Lissage par Whittaker-Henderson

Le lissage de Whittaker-Henderson repose sur deux critères :

1. un critère de fidélité : $F = \sum_{i=1}^p w_i (q_i - \hat{q}_i)$ (où p est le nombre de points) qui représente le fait que nos taux lissés sont en adéquation avec nos taux bruts. Là encore, les " w_i " représentent les poids que l'on affecte à nos observations.
2. un critère de régularité : $F = \sum_{i=1}^{p-z} (\Delta^z \hat{q}_i)^2$ où Δ^z est l'opérateur de différenciation d'ordre z, avec z un paramètre du modèle.

L'enjeu du lissage de Whittaker-Henderson est de minimiser une combinaison linéaire de ces deux critères, à savoir la quantité suivante : $M = F + h * S$ où h est le second paramètre du modèle. En termes d'interprétation de ce second paramètre, nous pouvons voir qu'il influe sur l'importance des deux critères l'un sur l'autre : plus h est grand plus il donne de l'importance au critère de régularité et plus h est petit plus il donne de l'importance au critère de fidélité.

La résolution de la minimisation de la quantité M peut encore une fois se faire matriciellement.

Détaillons sa résolution :

On pose :

1. Q : le vecteur des taux observés
2. \hat{Q} le vecteur des taux lissés
3. W la matrice diagonale des poids
4. K_z la matrice de taille (N-z,N) et dont les termes sont les coefficients binomiaux d'ordre z.

Avec ces notations là, on a : $F = (Q - \hat{Q})'W(Q - \hat{Q})$ et $S = (K_z\hat{Q})'(K_z\hat{Q})$.

$$\begin{aligned}\text{Or } M &= F + h * S = (Q - \hat{Q})'W(Q - \hat{Q}) + h * (K_z\hat{Q})'(K_z\hat{Q}) \\ &= \hat{Q}'W\hat{Q} - 2\hat{Q}'WQ + Q'WQ + h\hat{Q}'K'_zK_z\hat{Q}.\end{aligned}$$

Maintenant que nous avons cette expression de M, nous la dérivons par rapport à \hat{Q} :
 $M' = 2W\hat{Q} - 2WQ + 2hK'_zK_z\hat{Q}$

Nous en déduisons alors l'estimateur de \hat{Q} qui vaut : $(W+hK'_zK_z)^{-1}WQ$.

Bien que nous ayons présenté théoriquement les deux critères, nous avons choisi d'effectuer les lissages pour des valeurs de h valant 1, 100 et 1000 et pour les valeurs de z valant 2,3 et 4 de manière à bien percevoir l'influence de chacun des 2 paramètres.

Comme pour les splines cubiques, nous présentons les résultats d'abord pour les femmes puis pour les hommes :

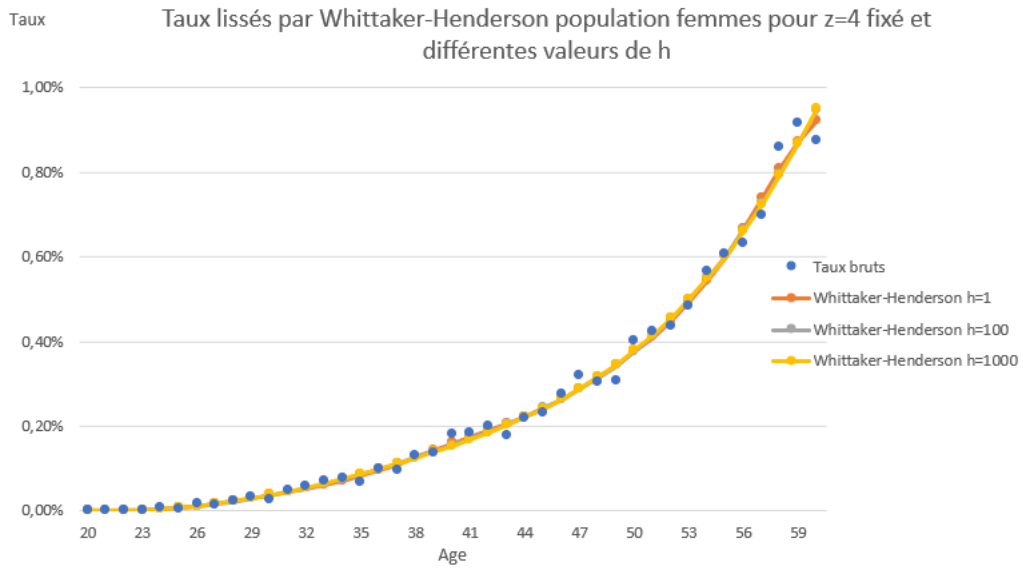


FIGURE 23 – Lissage par Whittaker-Henderson avec variation du paramètre h pour la population femmes

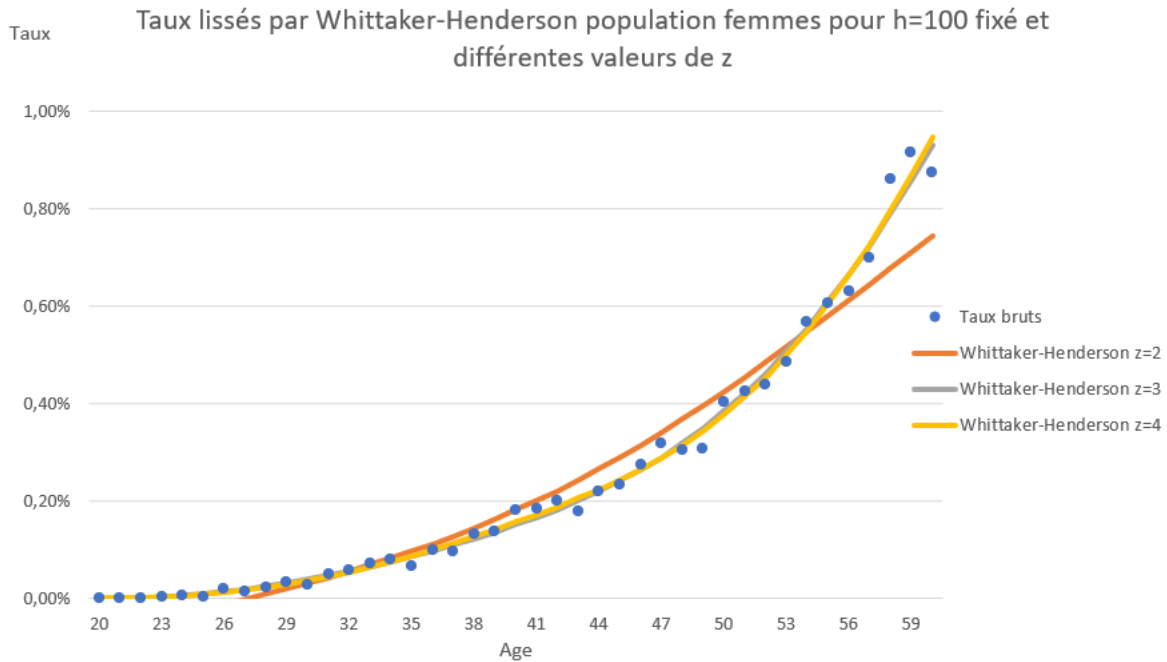


FIGURE 24 – Lissage par Whittaker-Henderson avec variation du paramètre z pour la population femmes

Du fait du volume de données important, ce dernier tend à minimiser l'impact des paramètres h et z : nous pouvons notamment le voir sur le paramètre h où les trois courbes sont quasiment superposées. Toutefois, l'impact du paramètre z est bien illustré sur le second graphique et nous suggère de considérer une valeur de z valant 3 ou 4 dans ce cas. Voyons les résultats analogues au sujet de la population hommes :

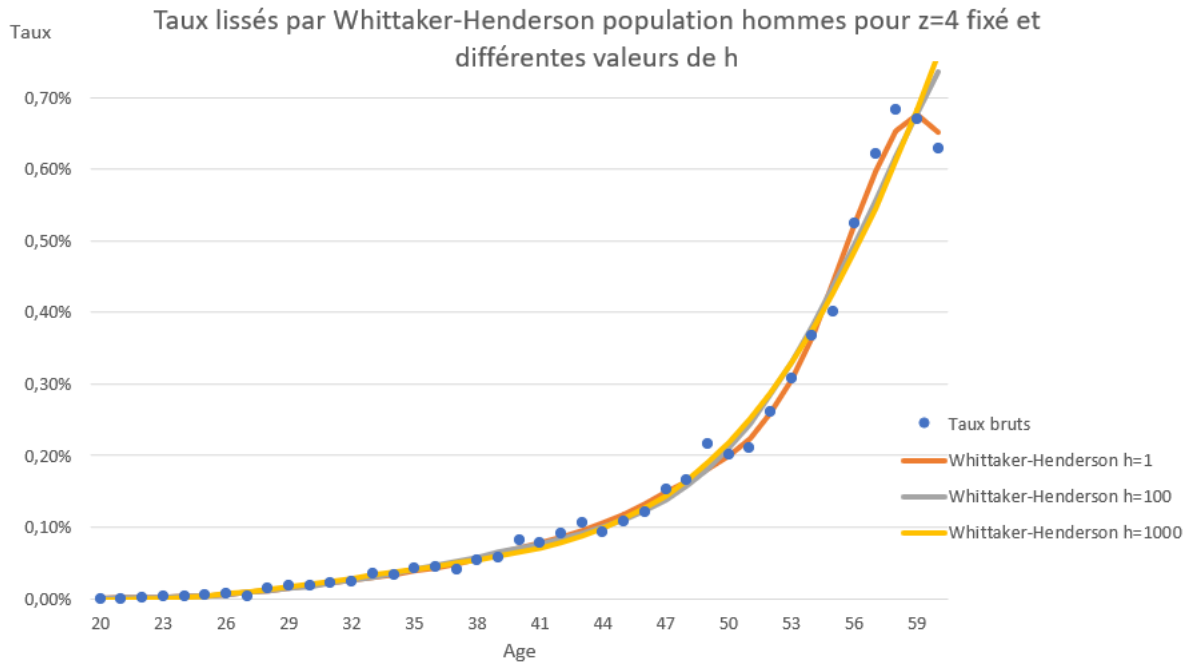


FIGURE 25 – Lissage par Whittaker-Henderson avec variation du paramètre h pour la population hommes

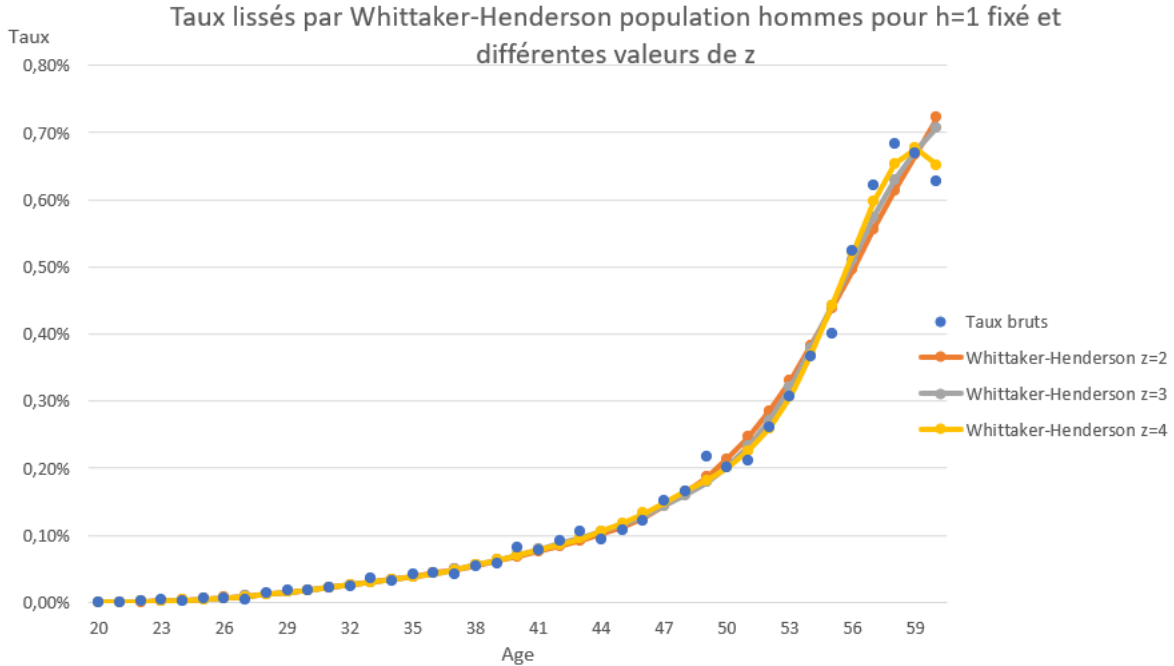


FIGURE 26 – Lissage par Whittaker-Henderson avec variation du paramètre z pour la population hommes

Là encore les extrémités hautes des courbes permettent de distinguer les différences dues aux différentes valeurs de z et h .

Nous présentons maintenant le lissage par moyennes mobiles qui est davantage accessible en termes de compréhension.

5.3 Lissage par moyennes mobiles

Le lissage par moyennes mobiles est une technique non paramétrique dont l'expression la plus simple est la suivante :

$$\hat{q} = \frac{\sum_{i=-v}^v q_{x+i}}{2v+1}$$

Comme l'indique son nom, l'objectif est de déterminer des valeurs lissées à partir des v valeurs à droite et des v valeurs à gauche de l'observation que l'on veut lisser soit un total de $2v+1$ valeurs.

Un des problèmes majeurs de ce lissage est que si parmi ces $2v+1$ valeurs figurent une ou plusieurs valeurs extrêmes, le lissage va être directement impacté et ce d'autant plus si v est petit.

En effet, plus v est grand, moins on accorde d'importance à l'observation que l'on lisse.

Dans notre cas, étant donnée la régularité de la pente qui croît de manière relativement similaire, nous avons choisi de prendre $v=3$ ce qui nous donne les résultats suivants :

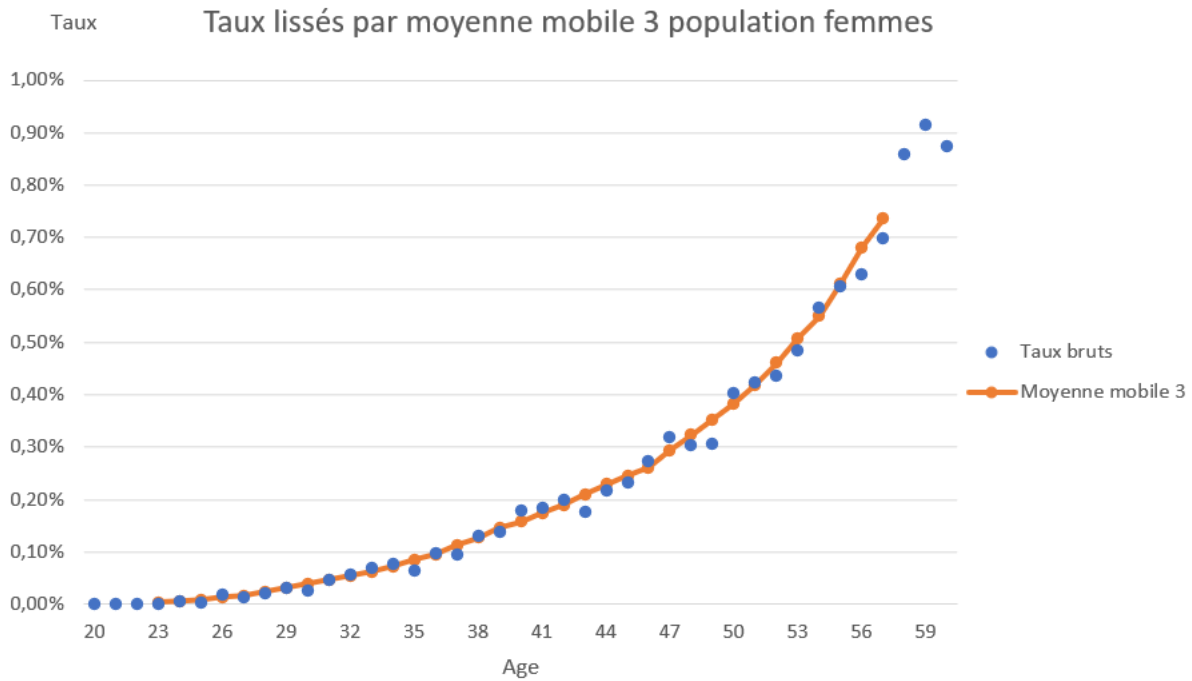


FIGURE 27 – Lissage par moyenne mobile pour la population femmes

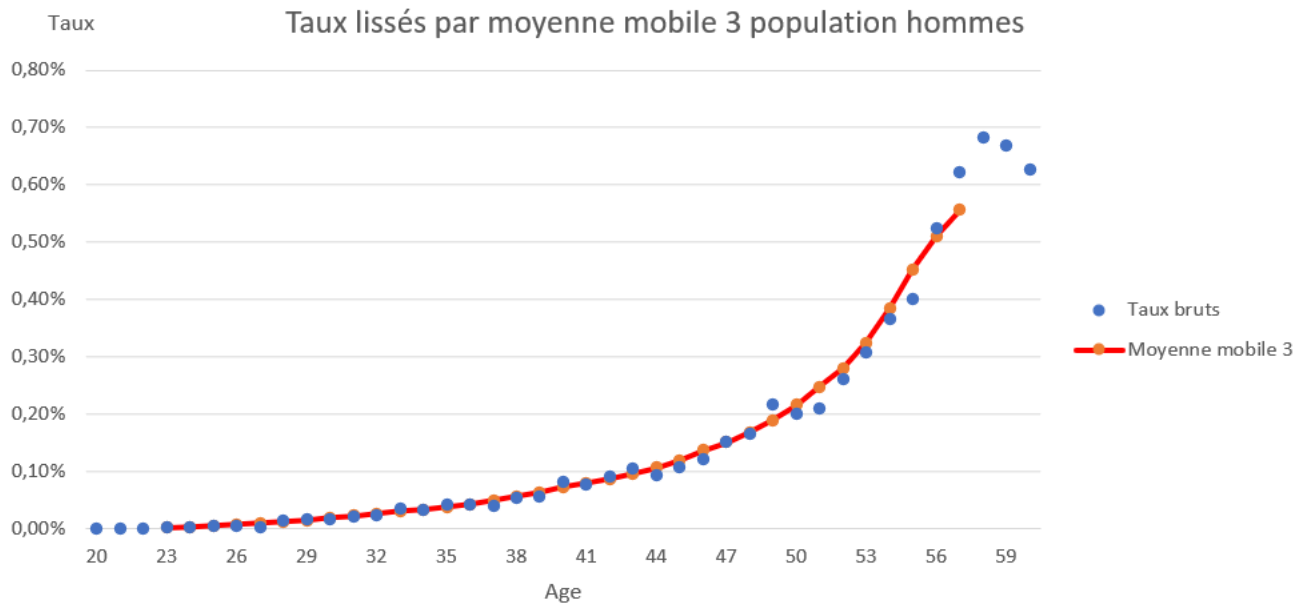


FIGURE 28 – Lissage par moyenne mobile pour la population hommes

Visuellement, les lissages ont l'air satisfaisants mais bien évidemment présentent l'inconvénient de ne pas pouvoir donner de taux lissés pour des âges strictement supérieurs à 60-v et strictement inférieurs à 20+v soit 6 âges car nous avons fixé v=3 ce qui représente un manque de valeurs lissées pour 15% des âges...

Pour pallier à cette sensibilité des valeurs extrêmes, nous pouvons considérer les valeurs lissées suivantes :

$$\hat{q} = \frac{\sum_{i=-v}^v q_{x+i} - \max(q_{x+i}) - \min(q_{x+i})}{2v-1}$$

Cette formule correspond précisément à celle des moyennes mobiles écrêtées.

De manière assez intuitive on retire les 2 valeurs les plus extrêmes ; la plus grande et la plus petite. Comme nous avons choisi v=3, le lissage par moyenne mobile écrêtée repose sur 5 valeurs.

Nous exposons de suite les résultats liés aux populations femmes et hommes :

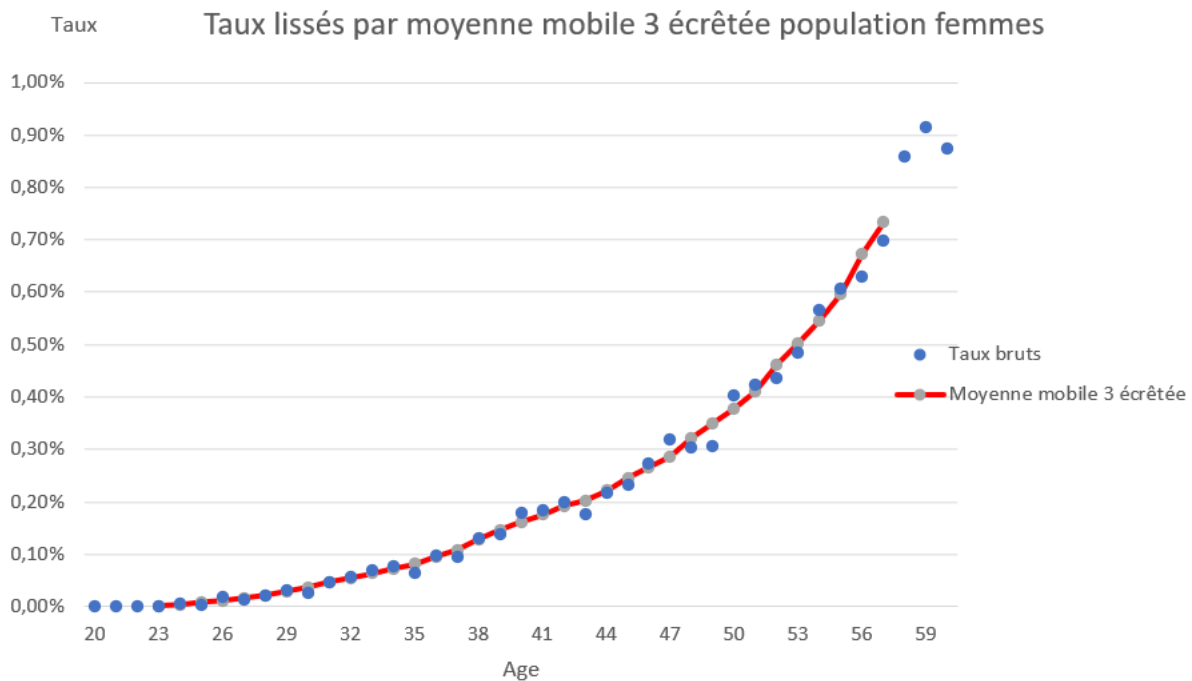


FIGURE 29 – Lissage par moyenne mobile écrêtée pour la population femmes

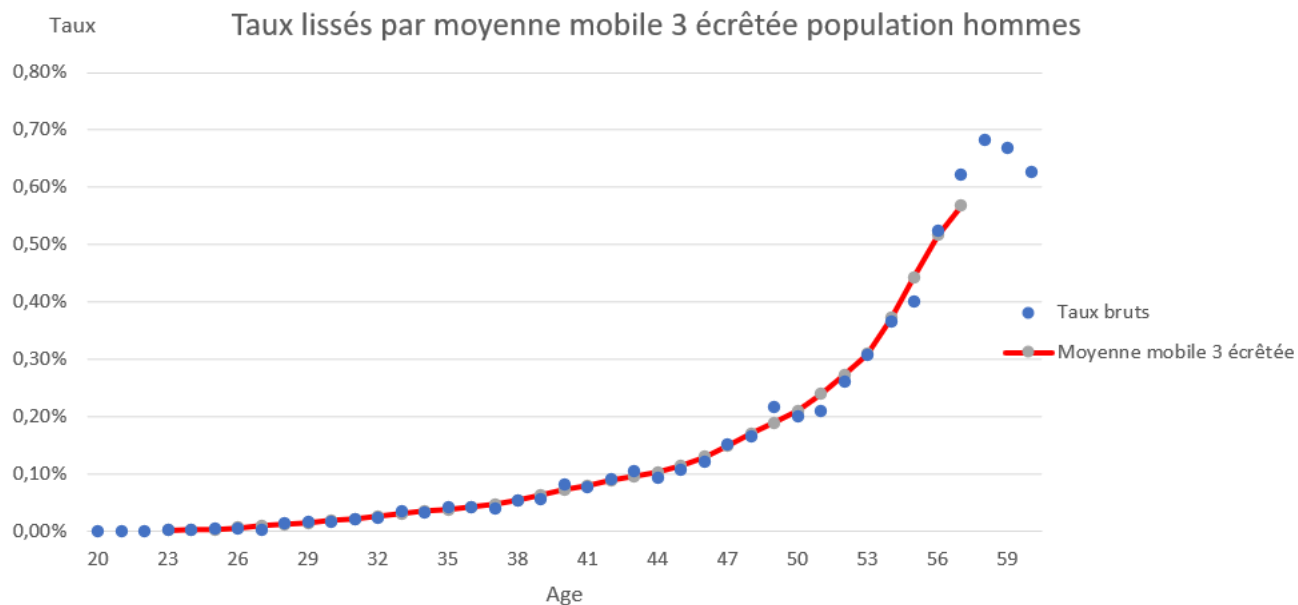


FIGURE 30 – Lissage par moyenne mobile écrêtée pour la population hommes

Comme notre courbe des taux bruts est relativement régulière, hormis sur les âges très élevés, mais que les moyennes mobiles ne fournissent pas de valeurs lissées pour les âges de 58 ans, 59 ans et 60 ans, nous voyons que les moyennes mobiles "classiques" et écrêtées sont quasiment identiques. Nous joignons tout de même en annexe les graphiques comparatifs des moyennes mobiles pour les deux populations étudiées.

Présentons enfin les lissages par la méthode des noyaux avant de tester nos lissages.

5.4 Lissage par la méthode des noyaux

Cette méthode non-paramétrique de lissage, aussi appelée méthode de Nadaraya-Watson, repose sur le modèle de régression suivant :

$Y_i = m(x_i) + e_i$ avec m une fonction totalement inconnue et les e_i qui satisfont classiquement les conditions suivantes :

$$E[e_i] = 0, V[e_i] = \sigma_{e_i}^2 \text{ et } Cov(e_i, e_j) = 0 \text{ pour tout } i \text{ différent de } j.$$

La régression par noyaux vise précisément à estimer l'espérance conditionnelle :

$m(X) = E[Y|X]$. Pour cela on veut déterminer les quantités suivantes :

$$m(x) = E[Y|X = x].$$

Pour ce faire, on utilise des noyaux représentés par une fonction K qui a les propriétés suivantes :

1. régulière
2. positive
3. $K(u) = K(-u)$

Si l'on revient à notre problématique, on a $Y = Q$ les taux bruts et $m = \hat{Q}$ les taux lissés donc on a $Q_{x_i} = \hat{Q}_{x_i} + e_{x_i}$.

Si l'on pose :

1. $\hat{f}_h(x)$ un estimateur de la densité de X .
2. N le nombre d'observations qu'on utilise pour déterminer $\hat{f}_h(x)$
3. h le paramètre du lissage par noyaux, qu'on appelle fenêtre, qui permet de définir le degré de lissage

Avec ces éléments, nous avons :

1. $\hat{f}_h(x) = \frac{\sum_{n=1}^N K\left(\frac{X-X_N}{h}\right)}{Nh}$
2. $\sqrt{Nh}(\hat{f}_h(x) - f(x))$ qui tend en loi vers $\mathcal{N}(0; \int K(u)^2 du)$

Revenons maintenant à l'estimation de notre espérance conditionnelle :

$$\hat{(q)}_x = E[q_x | X = x] = \int q_x \frac{f(q_x; x)}{f(x)} dq_x$$

Nous estimons le numérateur et le dénominateur séparément.

D'une part on a directement une estimation du dénominateur avec

$$\frac{\sum_{n=1}^N K\left(\frac{X-X_N}{h}\right)}{Nh}$$

Quant au numérateur, à l'aide des propriétés de parité de K et du fait qu'il s'agit d'une densité donc d'intégrale valant 1 sur l'espace entier, on l'estime avec la quantité suivante :

$$\frac{\sum_{n=1}^N K\left(\frac{X-X_N}{h}\right) q_{X_N}}{Nh}$$

Enfin, et pour achever la partie théorique, l'estimateur de Nadaraya-Watson s'obtient en prenant $N=2v+1$ et x_N les $2v+1$ âges voisins de l'observation en question.

En réalité, cet estimateur peut se résumer à une simple moyenne mobile à la différence près où l'on affecte à chaque observation des poids fournis par le noyau K et non le même poids à chaque observation. Quant aux noyaux que nous allons tester, nous les présentons ci-dessous :

1. le noyau gaussien qui est le plus répandu qui est de densité bien connue $K(u) = \frac{\exp \frac{-u^2}{2}}{\sqrt{2\pi}}$
2. le noyau d'Epanechnikov de densité $k(u) = \frac{3(1-u^2)}{4}$ si $|U|$ est plus petit que 1 et 0 sinon
3. le noyau cubique de densité $k(u) = \frac{34(1-u^2)^3}{35}$ si $|U|$ est plus petit que 1 et 0 sinon

Ces noyaux ont la forme suivante si nous les consignons dans un même graphique :

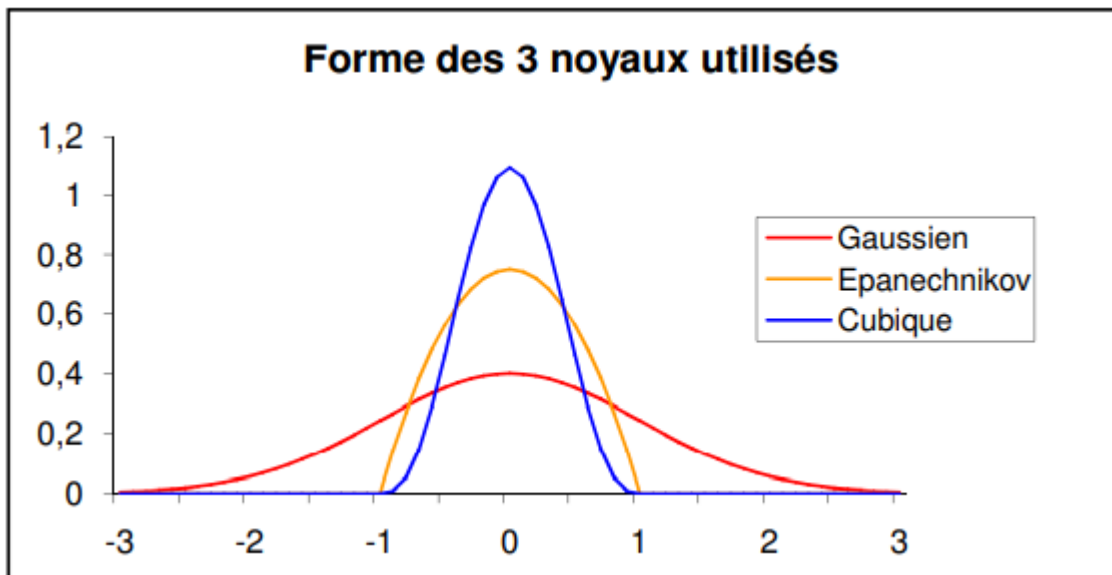


FIGURE 31 – Forme des 3 noyaux utilisés

Comme pour les moyennes mobiles et pour les mêmes raisons, nous avons choisi de considérer les 3 plus proches voisins pour effectuer le lissage ($v=3$) et nous testons h pour les valeurs 2,3 et 4.

Nous joignons ci-dessous les lissages associés aux trois noyaux pour une valeur de h fixé :

1. $h=4$ pour les femmes
2. $h=2$ pour les hommes

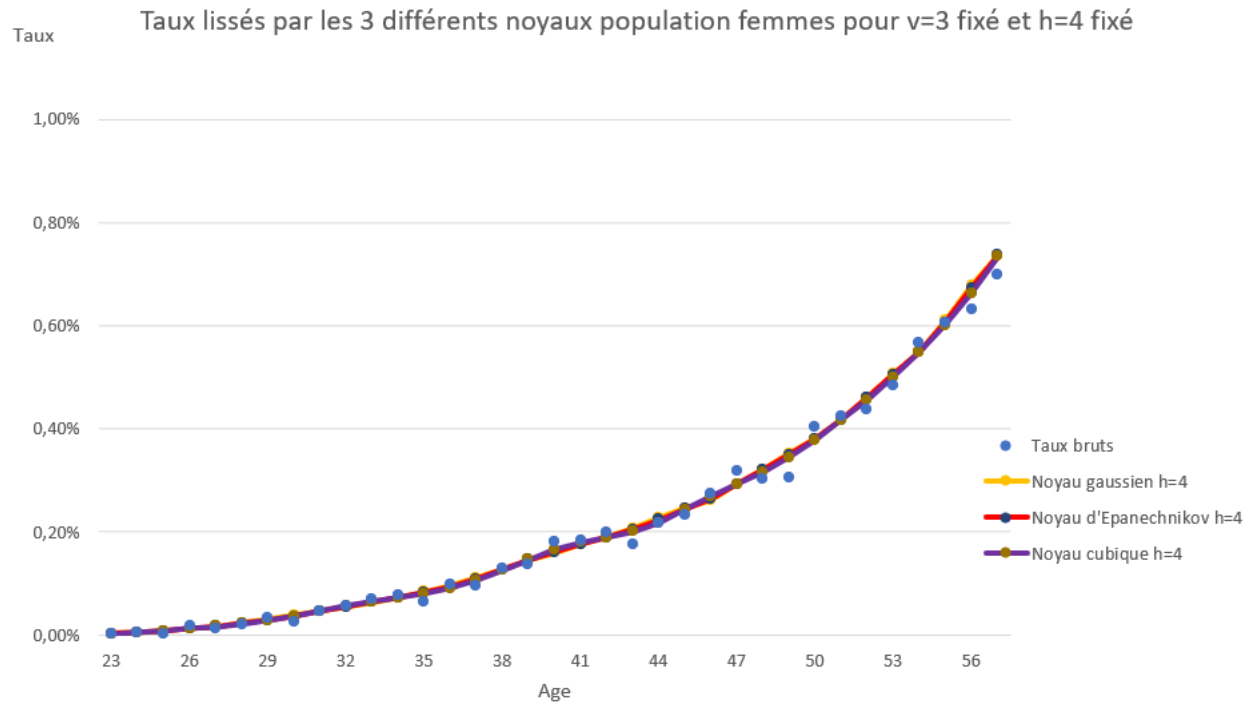


FIGURE 32 – Comparaison des lissages par noyaux pour la population femmes

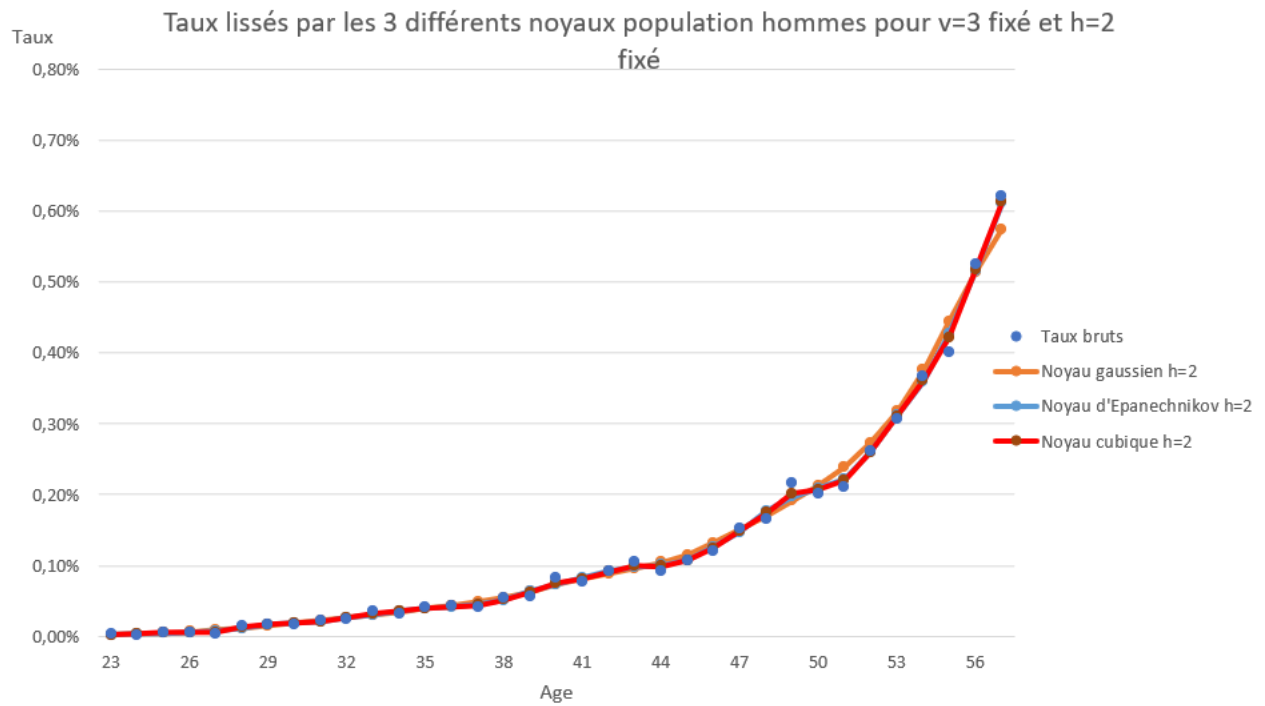


FIGURE 33 – Comparaison des lissages par noyaux pour la population hommes

Ces trois noyaux affectant donc des poids légèrement différents à chacune des observations présentent tout de même un certain nombre de points communs que sont les propriétés mentionnées antérieurement.

Pour voir plus précisément l'impact du paramètre h sur le lissage, nous avons joint en annexe les graphiques correspondants aux lissages effectués pour les 3 noyaux traités séparément en faisant varier h entre 2 et 4 comme expliqué plus haut. En résumé, il apparaît de suite que visiblement la fenêtre h n'a pas énormément d'impacts étant donnée la certaine régularité des taux bruts de base.

De plus, notons que les taux lissés ne sont disponibles qu'entre 23 ans et 57 ans inclus du fait du point suivant (déjà relevé pour les moyennes mobiles) ; nous ne disposons pas de taux lissés pour des âges strictement supérieurs à $60-v$ et strictement inférieurs à $20+v$ soit 6 âges car nous avons fixé $v=3$ ce qui représente un manque de valeurs lissées pour 15% des âges...

Notons à titre informatif que nous avons réalisé l'ensemble des lissages (et des résolutions matricielles pour les lissages concernés) sur Excel.

Après avoir étudié 4 méthodes de lissage différentes : par splines cubiques, par Whittaker-Henderson, par moyennes mobiles et par la méthode des noyaux, nous allons faire différents tests sur les valeurs lissées que nous vous présentons et résumons dans la partie suivante. A l'issue de cette dernière, nous choisirons donc des taux lissés à caractère définitifs qui seront alors utilisés pour l'élaboration d'un tarif.

6 Tests de lissage et validation de la table

Bien que les lissages puissent déjà fournir visuellement une idée quant à leur qualité, nous devons appliquer des méthodes statistiques de lissage permettant de les valider ou non.

Nous allons commencer par présenter le test du changement de signe :

6.1 Test du changement de signe

Le principe de ce test non-paramétrique est de regarder la fréquence de changement de signe entre les taux bruts et les taux lissés : nous allons donc observer la différence de ces deux quantités.

Pour cela, on teste l'hypothèse nulle H_0 :

$$P(q_x - \hat{q}_x) > 0 = P(q_x - \hat{q}_x) < 0 = \frac{1}{2} \text{ contre l'hypothèse inverse } H_1 :$$

$$P(q_x - \hat{q}_x) > 0 \neq \frac{1}{2} (\Rightarrow P(q_x - \hat{q}_x) < 0 \neq \frac{1}{2}).$$

Si H_0 est validée cela revient à dire que nous avons autant de fois les faits suivants :

1. les taux lissés sont supérieurs aux taux bruts
2. les taux bruts sont supérieurs aux taux lissés

Autrement dit, sans perdre de généralités et en admettant les âges indépendants les uns des autres, on peut considérer, que pour un âge sur deux, le signe de la différence change : $P((q_x - \hat{q}_x) \text{ et } (q_{x+1} - \hat{q}_{x+1}) \text{ sont de signes opposés}) = \frac{1}{2}$

En effet cela vient du fait que :

$$P((q_x - \hat{q}_x) \text{ et } (q_{x+1} - \hat{q}_{x+1}) \text{ sont de signes opposés}) \\ = P((q_x - \hat{q}_x) > 0)P((q_{x+1} - \hat{q}_{x+1}) < 0) + P((q_x - \hat{q}_x) < 0)P((q_{x+1} - \hat{q}_{x+1}) > 0)$$

Alors :

$\mathbb{1}_{(q_x - \hat{q}_x)(q_{x+1} - \hat{q}_{x+1}) < 0}$ suit une loi de Bernoulli de paramètre $\frac{1}{2}$ sous H_0 . En passant tout naturellement à la somme :

$$S_+^- = \sum_{i=1}^{N-1} \mathbb{1}_{(q_x - \hat{q}_x)(q_{x+1} - \hat{q}_{x+1}) < 0}$$

On en déduit très rapidement que cette statistique suit une loi binomiale de paramètres $(N-1, \frac{1}{2})$. En fonction du nombre de données à lisser, on utilise le test à distance finie pour N petit ou bien une approximation gaussienne pour N jugé suffisamment grand. Nous détaillons de suite ces deux méthodes un peu plus en détail :

1. le test à distance finie consiste à choisir, pour un seuil α , le plus grand entier k_α tel que, pour X une variable aléatoire qui suit une loi binomiale de paramètres $(N-1, \frac{1}{2})$, on ait :

$$P_{H_0}(X < k_\alpha) < \frac{\alpha}{2}.$$

A partir de là, nous regardons la quantité S_+^- :

- (a) si elle appartient à $]k_\alpha ; N-1-k_\alpha[$ alors on accepte l'hypothèse nulle et on valide le lissage
 - (b) dans le cas contraire, nous rejetons H_0 et invalidons le lissage par la même occasion
2. l'approximation gaussienne consiste à dire que S_+^- suit une loi normale de paramètres $(\frac{N-1}{2} ; \frac{N-1}{4})$. En retranchant la moyenne et en divisant par l'écart-type (c'est-à-dire la racine carrée de la variance), nous étudions la quantité suivante :

$$Z = \frac{S_+^- - \frac{N-1}{2}}{\sqrt{\frac{N-1}{4}}}$$

Cette quantité suit une loi normale centrée réduite sous H_0 qui nous permet alors de rejeter ou non l'hypothèse nulle.

Dans notre cas, nous considérons que N est relativement petit (41 observations) : c'est pourquoi nous effectuons le test à distance finie.

On a donc S_+^- qui suit une loi binomiale de paramètres $(40, \frac{1}{2})$.

Si l'on reprend la description du test à distance finie juste au-dessus, nous devons fixer un seuil α qui fait office de seuil de confiance en quelque sorte : nous le fixons à 10%.

A ce stade, on cherche alors le plus grand entier k_α tel que, pour X une variable aléatoire qui suit une loi binomiale de paramètres $(40, \frac{1}{2})$, on ait :

$P_{H_0}(X < k_\alpha) < \frac{\alpha}{2}$. Autrement dit, on cherche le plus grand entier k_α tel que :

$P(S_+^- < k_\alpha) < 5\%$ et par symétrie :

$P(S_+^- > 40 - k_\alpha) < 5\%$

Nous affichons en annexe les quantités $P(X=k)$ et $F(X \leq k)$ pour toutes les valeurs entières de k et pour X qui suit une loi binomiale de paramètres $(40; \frac{1}{2})$:

En parcourant le tableau, on peut voir que le premier entier k qui vérifie la condition est 15 donc on ne rejette pas H_0 si le nombre de changements de signes est compris dans $]k_\alpha; N-1-k_\alpha[$ soit entre 15 et 40-15 soit entre 15 et 25.

Nous joignons également en annexe les résultats de ce test pour les populations femmes et hommes.

Nous pouvons voir d'ores et déjà que le lissage par splines est invalidé ainsi que le lissage de Whittaker-Henderson et par noyaux pour certains paramètres.

Malgré tout, le test du changement de signe n'est pas suffisant pour décrire la qualité d'un lissage car il décrit uniquement les positions relatives des valeurs lissées et des valeurs brutes mais potentiellement les valeurs lissées ne reflètent pas assez bien la réalité tout en étant réparties de part et d'autre des taux bruts.

Pour résumer, valider le test du changement de signe est une condition nécessaire mais pas suffisante, nous excluons donc à son issue le lissage par splines et les lissages de Whittaker-Henderson et par noyaux pour certaines valeurs de paramètres ainsi que le lissage par moyenne mobile "classique" pour la population hommes.

6.2 Test du SMR

Nous passons au test du Standardized Mortality Ratio (Indice standardisé de mortalité) qui s'applique aussi aux invalidités. C'est tout simplement un ratio entre les invalidités réellement observées et les invalidités estimées (c'est-à-dire le produit de nos taux lissés par notre exposition) :

$$SMR = \frac{i_x}{\sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i}}$$

Logiquement nous testons donc $H_0 : "SMR=1"$ contre $H_1 : "SMR \neq 1"$. Mathématiquement, sous H_0 , on considère que le nombre d'entrée en invalidité suit une loi de Poisson de paramètre $\lambda = \sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i}$ ce qui est assez courant pour la modélisation de phénomènes

jugés "rares". Nous regardons alors l'intervalle de confiance à 95% du nombre d'entrée en invalidité ce qui implique de regarder les valeurs de la fonction de répartition de notre statistique pour lesquelles on atteint les valeurs 0,025 et 0,975.

A noter que ces 2 valeurs varient bien évidemment selon les valeurs lissées mais valent globalement 0,978 pour la borne inférieure et 1,021 pour la borne supérieure à titre indicatif.

Notons que si le nombre d'invalidités est important, il n'est pas aberrant de considérer que une loi normale de paramètres $(\sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i}; \sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i})$ approxime notre loi de Poisson de paramètre : $(\sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i})$.

Dans ce cas précis, notre SMR suivrait alors une loi normale de paramètre $(1; \frac{1}{\sum_{x_i=1}^{41} n_{x_i} \hat{q}_{x_i}})$ sous l'hypothèse nulle.

Comme d'habitude, nous présentons dans les tableaux en annexe les résultats du test du SMR pour les populations femme et homme.

Nous pouvons voir que les résultats sont rigoureusement identiques. Plusieurs remarques par rapport à ces derniers :

1. les lissages par moyennes mobiles et par noyaux sont invalidés car, du fait du nombre restreint de valeurs lissées (35 valeurs lissées pour 41 valeurs brutes), on observe conséquemment un écart lors du calcul du SMR. Nous pourrions choisir de restreindre le calcul du SMR sur la plage s'étendant de 23 ans à 57 ans mais ce serait engendrer un biais supplémentaire et surtout ne pas avoir de valeurs lissées proprement pour les âges extrêmes.
2. les lissages par splines et par Whittaker-Henderson sont en revanche validés et pour celui de Whittaker-Henderson, nous avons même un SMR égal à 1 et ce, quelles que soient les valeurs de h et z, ce qui veut dire que nous obtenons exactement le bon nombre d'invalidités en multipliant nos taux lissés par notre exposition. Néanmoins, il semble donc que les paramètres z et h qui pour rappel, permettent de contrôler les critères de régularité et de fidélité, n'ont pas d'influence particulière sur la valeur du SMR.

Rappelons que le lissage par splines avait été exclu lors du test du changement de signes, nous allons donc opter pour un lissage par Whittaker-Henderson pour nos deux catégories de population.

Au vu de l'allure des graphiques d'une part, mais aussi des résultats du test du changement de signe qui invalidaient le lissage par Whittaker-Henderson pour certaines valeurs de z et h , on retient définitivement :

1. pour la population femmes : $h=100$ et $z=4$
2. pour la population hommes : $h=1$ et $z=4$

Nous mettons donc ci-après les deux graphiques des taux lissés définitifs correspondants aux paramètres énoncés.

Pour les femmes :

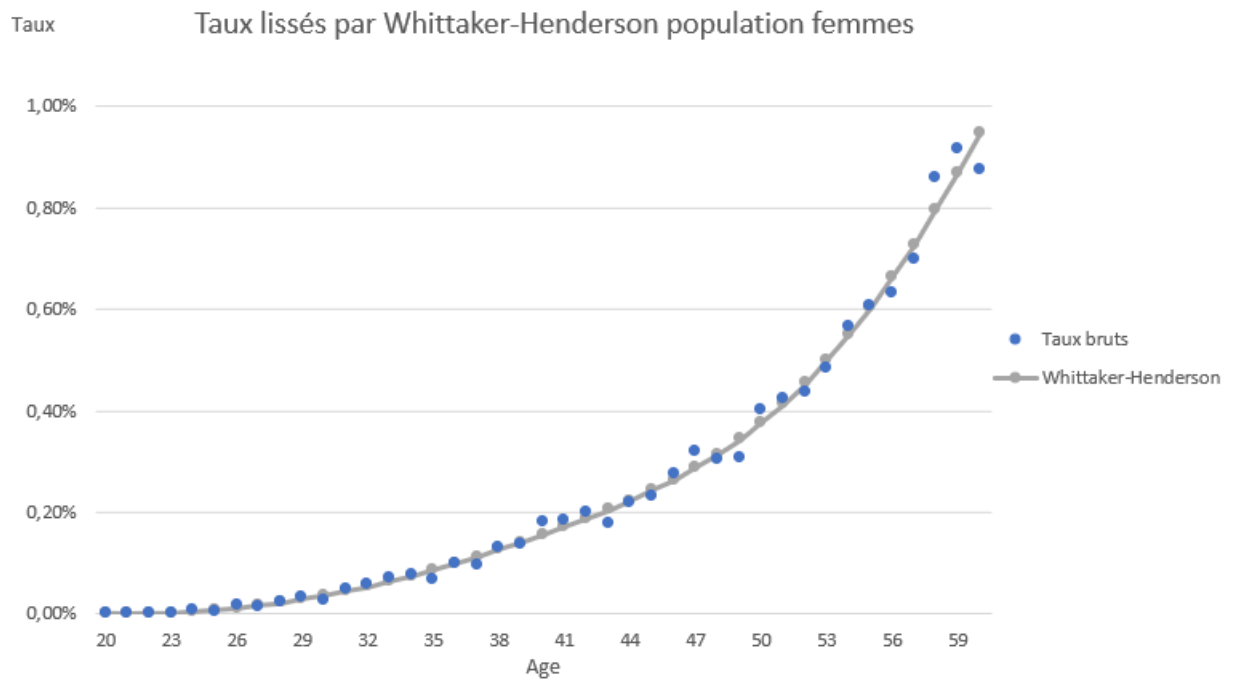


FIGURE 34 – Taux lissés définitifs pour la population femmes

et pour les hommes :

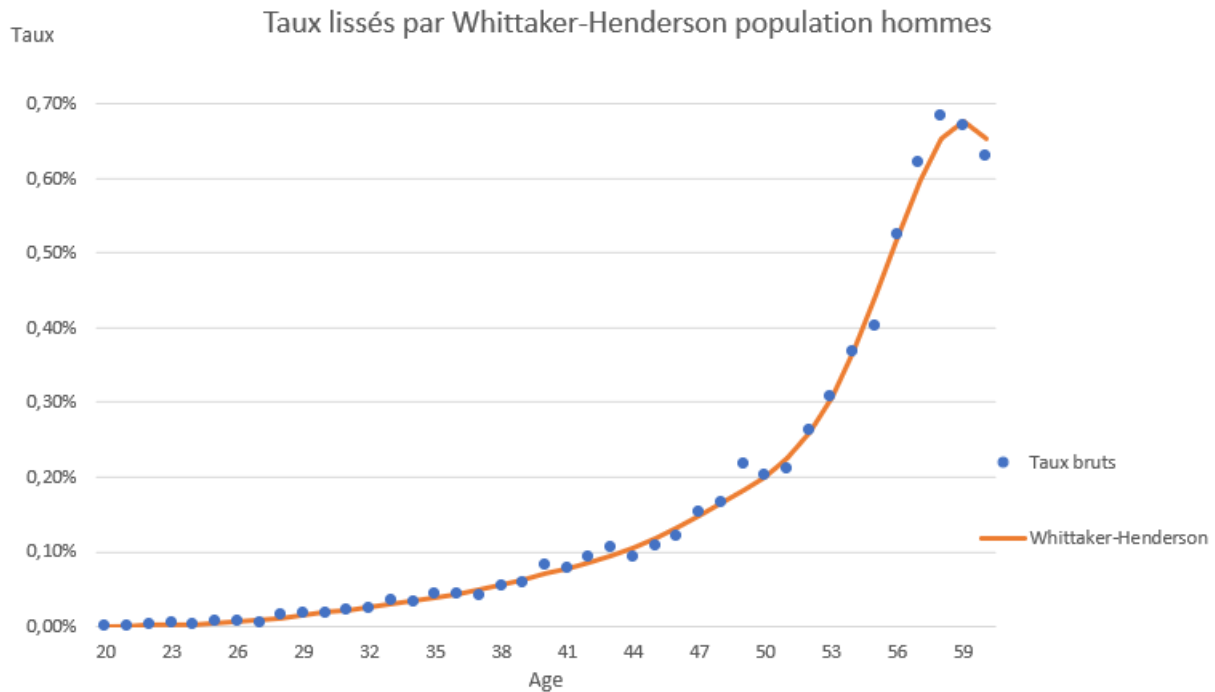


FIGURE 35 – Taux lissés définitifs pour la population hommes

Le choix des paramètres diffère selon les deux populations car comme nous pouvons le voir, sur les âges élevés, il y a un léger phénomène de redescente des taux chez la population masculine que nous n’observons pas sur la population féminine. En donnant des valeurs plus élevées à h et z , nous arrivons à percevoir ce phénomène dans une certaine proportion ce qui évite à notre courbe des taux lissés de s’envoler et donc de surestimer l’incidence en invalidité dans cette tranche d’âge.

Maintenant que nous avons validé ces deux tables dont le but est l’implémentation dans l’outil de tarification, nous allons d’abord comparer nos lois aux lois précédemment implémentées, donner une explications aux différences observées et expliquer l’apport de notre étude justement.

7 Comparaison avec les lois d'incidence en invalidité existantes

La véritable "nouveauité actuarielle" est l'utilisation des DSN comme base de données. Les tables d'incidence en invalidité jusqu'alors implémentées dans l'outil ont été construites en 2015 à partir d'une étude réalisée sur la période 2011-2014 et à partir des DADS-U (Déclarations Annuelles des Données Sociales Unifiées) et qui n'avaient pas la même importance qu'ont les DSN aujourd'hui.

Lors de l'étude précédente, très peu de sinistres avaient été retracés dans les DADSU et selon des critères moins fiables que ceux utilisés lors de notre étude. Les sinistres dont a pu retracer l'exposition représentent moins de 10% des sinistres totaux lors de l'étude de 2015 contre plus de 90% en ce qui concerne notre étude en 2022 soit un déficit important de sinistres dont nous connaissons l'exposition lors de la précédente étude...

Certes, la plage d'étude est deux fois plus large mais l'écart est colossal et s'explique justement par la source d'informations (à savoir les DSN versus les DADSU).

Nous montrons ci-après les courbes définitives, celles des taux lissés.

Pour les femmes :

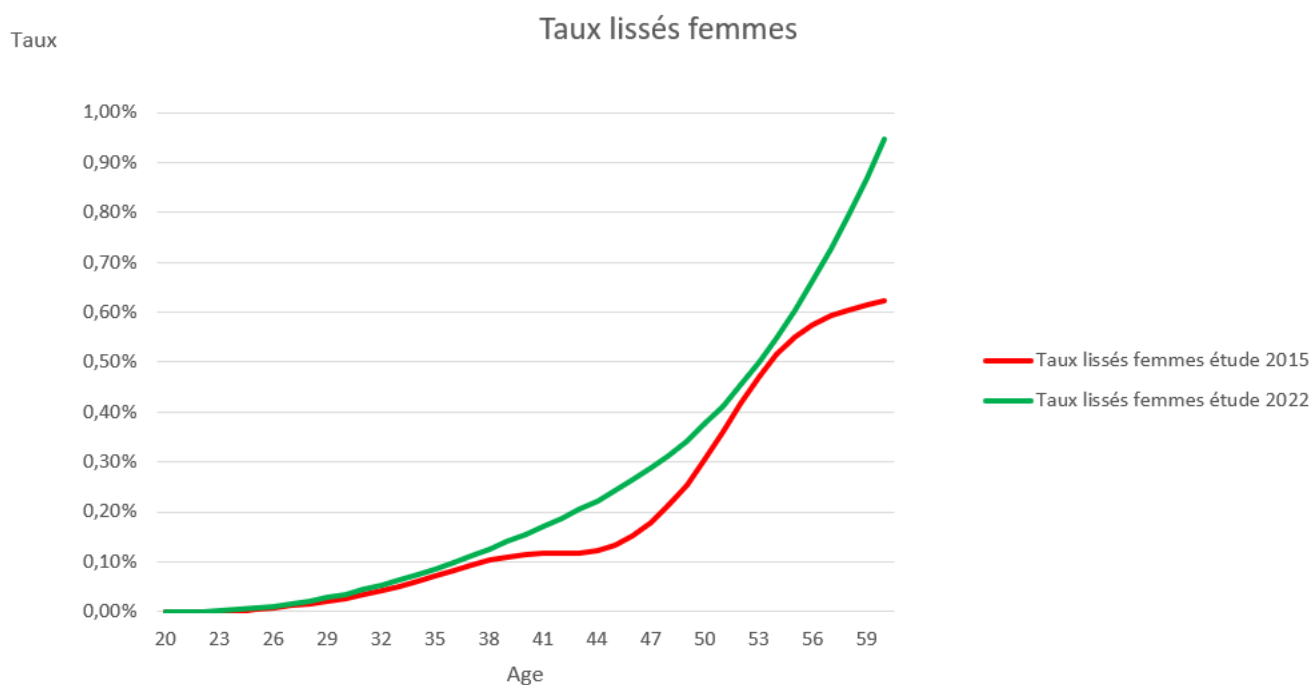


FIGURE 36 – Comparaison des taux lissés des deux études pour la population femmes

et pour les hommes :

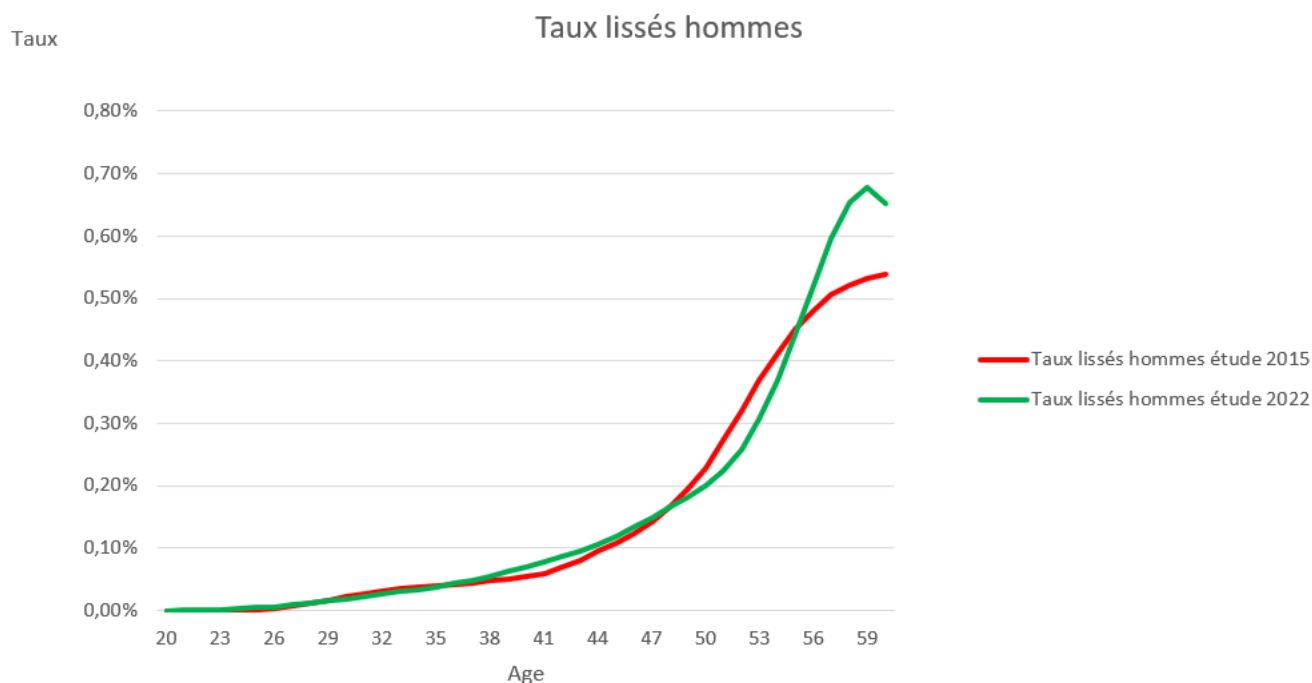


FIGURE 37 – Comparaison des taux lissés des deux études pour la population hommes

La comparaison des lois définitives fait état d’une sous-estimation globale de la sinistralité auparavant notamment sur la population femmes mais aussi légèrement sur la population hommes, et cette sous-estimation est d’autant plus importante que l’âge augmente.

Pour bien la mesurer nous allons appliquer les anciennes lois hommes et femmes à nos données actuelles afin de comparer le nombre d’invalidités prédit par rapport aux nombres d’invalidités réel :

	Femmes	Hommes
Nombre d’invalidités prédit	6395	6947
Nombre d’invalidités réel	8249	7226

FIGURE 38 – Comparaison du nombre d’invalidités total prédit et du nombre d’invalidités total réel

Comme le suggéraient les graphiques précédents, les anciennes lois sous-estiment la sinistralité de manière plus importante chez la population femmes que chez la population hommes. En l’occurrence, nous retrouvons au niveau du nombre d’invalidités un déficit de 4% contre un déficit de plus de 20% chez les femmes ce qui est conséquent. Comme évoqué

maintes fois, le but de ces lois est de mieux tarifier la garantie invalidité, nous allons donc décrire désormais le processus de tarification.

8 Tarification

En assurances collectives, la tarification se fait avant tout à la maille du numéro de RPP, nous en avons déjà parlé beaucoup plus haut lors du traitement de nos données.

Comme nous avons deux lois : une loi femme et une loi homme, nous appliquons le raisonnement ci-dessous aux deux sous-populations décrites.

En l'absence de chargements, le tarif individuel d'une rente invalidité est alors donné par la formule suivante :

Tarif individuel = (Taux d'incidence_{age_x})(annuité) = (Taux d'incidence_{age_x})(coût segmenté_{age_x})(salaire moyen_{age_x}).

Nous avons déjà les taux d'incidence issus des lois que l'on a construites. Nous avons également calculé le salaire moyen par âge car nous avons jugé que cette variable avait un impact sur la tarification (nous aurons l'occasion d'y revenir ultérieurement). Enfin, ce que l'on appelle "coût segmenté" est la quantité suivante :

$$\text{coût segmenté}_{age_x} = \sum_{age > x} \frac{P(\text{l'individu est toujours invalide sachant qu'il est devenu invalide à l'âge } x)(Flux_{age_x})}{(1-i)^{60-age}} - \frac{m-1}{2m}(1 - {}_n E_x)$$

Plusieurs remarques par rapport à cette formule du coût segmenté :

1. *i* désigne le taux technique (que l'on considérera à 1 dans notre cas)
2. dans notre tarification, nous avons pris le cas d'une rente de 1€ donc Flux=1
3. la probabilité au numérateur s'obtient grâce à une table de maintien que nous avons à disposition et qui est unique (on utilise la même table de maintien pour les femmes et les hommes)
4. le terme $-\frac{m-1}{2m}(1 - {}_n E_x)$ est issu des formules classiques d'assurance vie pour calculer une rente viagère payable *m* fois dans l'année, à termes échus pendant *n* années. Dans notre cas, nous prendrons *m*=12 car les rentes invalidités sont souvent mensuelles.

De plus, comme cette formule dépend logiquement de l'âge, nous avons fait nos calculs sur une seule et unique année : en l'occurrence l'année 2020.

Expliquons maintenant pourquoi le tarif individuel est différent du tarif groupe. Par sa construction-même, le tarif individuel est construit par âge et comme en assurance collectives nous tarifons à la maille du numéro de RPP, le tarif est construit sur l'âge moyen de l'entreprise.

Mais en toute logique, pour bien comprendre le biais engendré, une entreprise de deux individus : 1 de 30 ans et 1 de 50 ans ne porte pas le même risque qu'une entreprise de deux individus de 40 ans. Pourtant, l'âge moyen est le même : 40 ans. Et encore nous n'avons pas introduit dans cet exemple la notion de sexe...

Nous allons donc calculer le tarif groupe pour la population femmes d'une part et pour la population hommes d'autre part, c'est-à-dire le tarif qui correspond à la formule suivante :

$$\text{Tarif groupe} = \sum_{x=20}^{60} (\text{Proportion des salariés d'âge } x)(\text{Taux d'incidence}_{age_x})(\text{annuité}) = \sum_{x=20}^{60} (\text{Proportion des salariés d'âge } x)(\text{Taux d'incidence}_{age_x})(\text{coût fragmenté}_{age_x})(\text{salaire moyen}_{age_x}).$$

En fait le tarif groupe est la moyenne pondérée des tarifs individuels.

Nous avons pu réaliser cette segmentation en calculant dans un premier temps tous les âges moyens de nos numéros de RPP puis ensuite, en agrégeant tous les numéros de RPP possédant le même âge moyen, nous obtenons la démographie intégrale des âges des individus.

Une fois le tarif groupe obtenu, nous le comparons à la table des tarifs individuels pour voir l'écart engendré par la démographie de notre portefeuille.

Quoiqu'il en soit, nous présentons ci-après les tarifs individuel et groupe femmes : (là encore les résultats ne sont pas fidèles à la réalité mais respectent les ordres de grandeur)

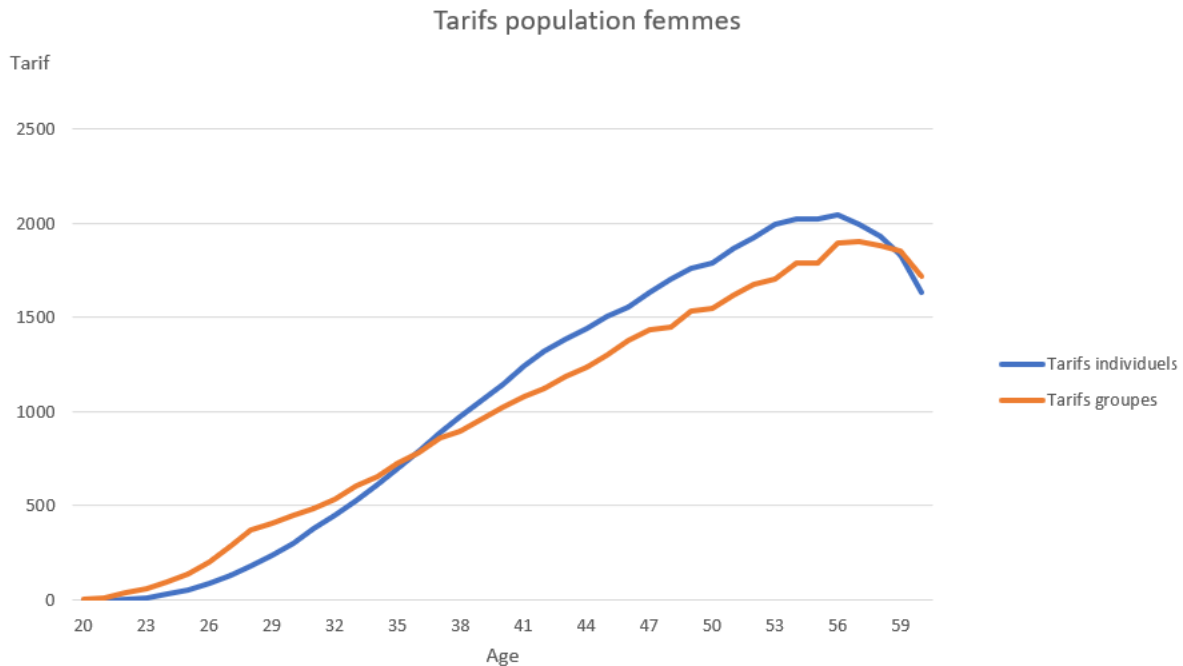


FIGURE 39 – Comparaison du tarif individuel et du tarif groupe pour la population femmes

puis ceux des hommes :

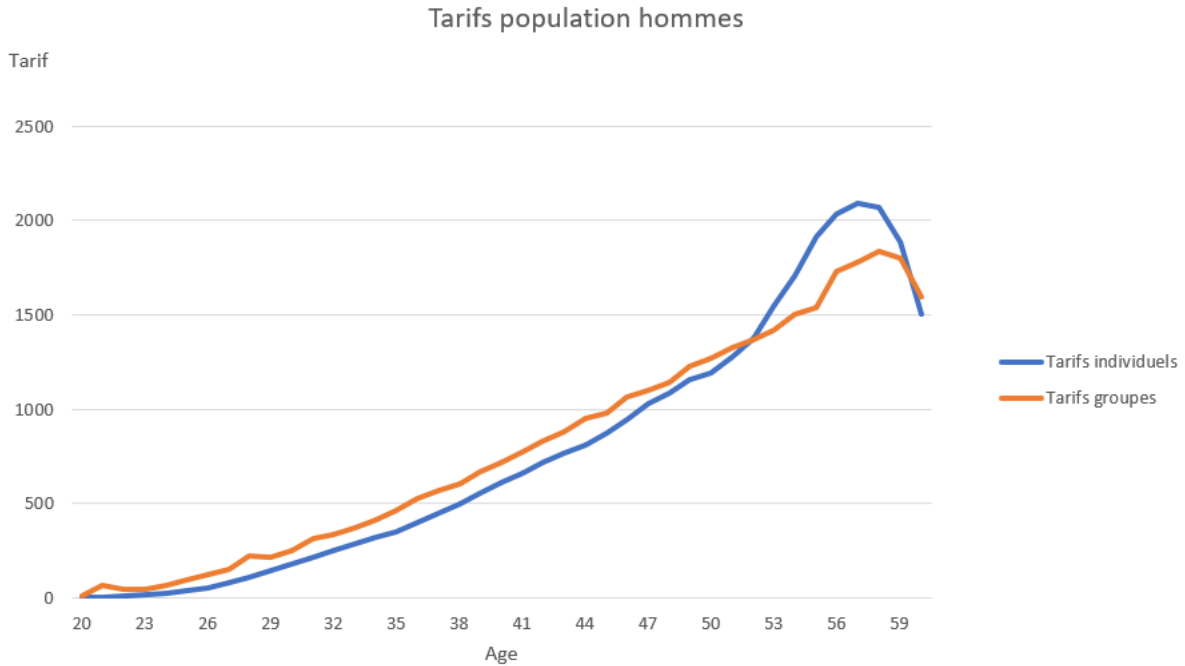


FIGURE 40 – Comparaison du tarif individuel et du tarif groupe pour la population hommes

Nous pouvons voir chez les deux populations les mêmes phénomènes :

1. au niveau des tarifs individuels, une courbe régulièrement croissante jusqu'à l'âge de 54 ans environ avant de former une cloche et de redescendre un peu. Ceci montre en fait que bien que les salaires augmentent avec l'âge en théorie, il y a un effet compensateur clairement non négligeable du coût qui baisse significativement car la somme $\sum_{age>x} \frac{P(\text{l'individu est toujours invalide sachant qu'il est devenu invalide à l'âge } x)(Flux_{age_x})}{(1-i)^{60-age}} - \frac{m-1}{2m}(1-nE_x)$ contient de moins en moins de termes.
2. les tarifs groupe qui prennent donc en considération la répartition de tous les âges pour un âge moyen fixé, semblent indiquer qu'il y a un phénomène de sous tarification sur les petits âges (avant 38 ou 42 ans selon les femmes ou les hommes) puis ensuite le contraire à savoir un phénomène de sur tarification sur les âges supérieurs (notamment au-delà de 45 ans). Nous observons surtout dans les 2 cas une forme de plateau ascendant entre 33 ans et 50 ans avant que la courbe marque des ruptures plus prononcées probablement du fait de l'existence de moins voire peu de données pour des âges moyens supérieurs à 50 ans.
3. enfin sur les âges très élevés (au-delà de 55 ans) les 2 courbes sont parfaitement alignées.

Si jamais l'on souhaite obtenir l'âge actuariel, à savoir non pas l'âge moyen d'un groupe au sens arithmétique du terme mais l'âge moyen d'un groupe au regard des risques assurés, la méthode est la suivante :

1. Pour un âge donné, nous regardons le tarif groupe, donc obtenu à partir du détail de la démographie, et nous recherchons quel tarif individuel s'en rapproche le plus
2. Bien souvent, le tarif individuel qui se rapproche le plus du tarif groupe n'est pas strictement égal au tarif groupe (d'où la notion de valeur la plus proche). Nous effectuons alors un système de barycentre entre l'âge correspondant à la valeur la plus proche et l'âge supérieur (respectivement inférieur) si le tarif groupe était supérieur (respectivement inférieur) à la valeur la plus proche du tarif individuel.

Pour effectuer ce calcul de barycentres, nous faisons l'approximation selon laquelle le tarif augmente linéairement entre deux âges successifs.

3. L'âge actuariel vaut alors à l'âge correspondant dans la table des tarifs individuels, plus ou moins augmenté de 0,50 au maximum.

Enfin, remarquons que notre méthode engendre tout de même un certain biais : nous considérons l'âge moyen d'une entreprise pour notre tarification. Or, pour calculer notre tarif groupe, nous séparons les femmes et les hommes du fait de nos deux lois construites. Nous pouvons très bien imaginer que pour une entreprise dont l'âge moyen est 40 ans, l'âge moyen des femmes soit supérieur et celui des hommes inférieur par exemple.

Pour être tout à fait complet, il aurait fallu étudier la dispersion des âges à la maille du sexe pour chaque âge moyen et effectuer les pondérations nécessaires ensuite.

Ceci vient conclure l'ensemble de cette étude qui a permis de retracer l'élaboration d'un tarif en assurances collectives depuis la collecte des données jusqu'à la tarification finale en passant donc par les étapes d'estimation des taux bruts, de lissage et de tests des lissages entre autres.

Toutefois, cette étude vient "seulement" répondre à des contraintes opérationnelles puisque, comme expliqué en introduction, s'agissant de la garantie invalidité, l'outil de tarification actuel est implémenté avec les lois hommes et femmes uniquement.

Or, intuitivement, il apparaît peut-être plus judicieux d'opter pour une segmentation cadre/non-cadre pour ce type de garantie voire carrément d'implémenter à l'avenir dans l'outil de tarification quatre lois : une loi femmes, une loi hommes, une loi cadres et une loi non-cadres.

9 Lois d'incidence pour les cadres et les non-cadres

Dans le retraitement de nos données, nous avons déjà effectué des travaux sur le remplissage des salaires manquants permettant d'identifier l'appartenance à la catégorie cadre ou non-cadre pour la CSP 4 : professions intermédiaires qui était litigieuse.

Autrement dit, nous avons dans notre base initiale tous les ingrédients pour construire deux lois distinctes cadres et non-cadres. Nous allons donc présenter dans ce qui suit exactement la même démarche que celle que nous avons déjà réalisée mais avec cette fois, une segmentation cadre/non-cadre.

Nous gardons néanmoins à l'esprit que, pour l'instant, cette partie est une ébauche de ce qui pourrait être fait à l'avenir mais ne répond pas aux contraintes opérationnelles.

9.1 Quelques statistiques descriptives

Dans un premier temps, nous présentons la répartition de l'exposition et la répartition des invalidités de la population cadres :

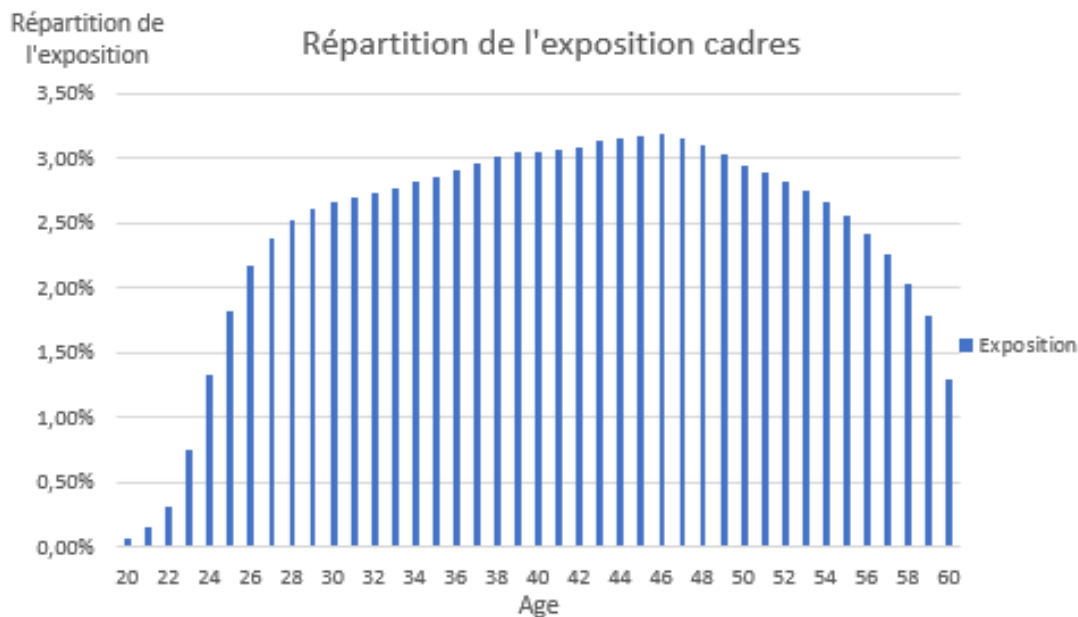


FIGURE 41 – Répartition de l'exposition des cadres en fonction de l'âge

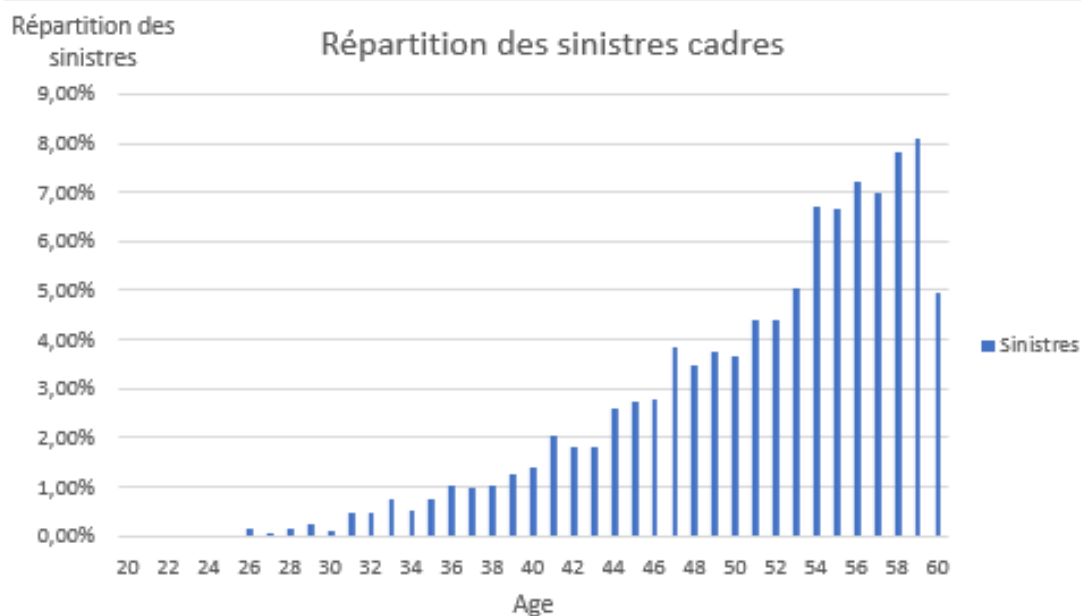


FIGURE 42 – Répartition des sinistres des cadres en fonction de l'âge

puis celles des non-cadres :

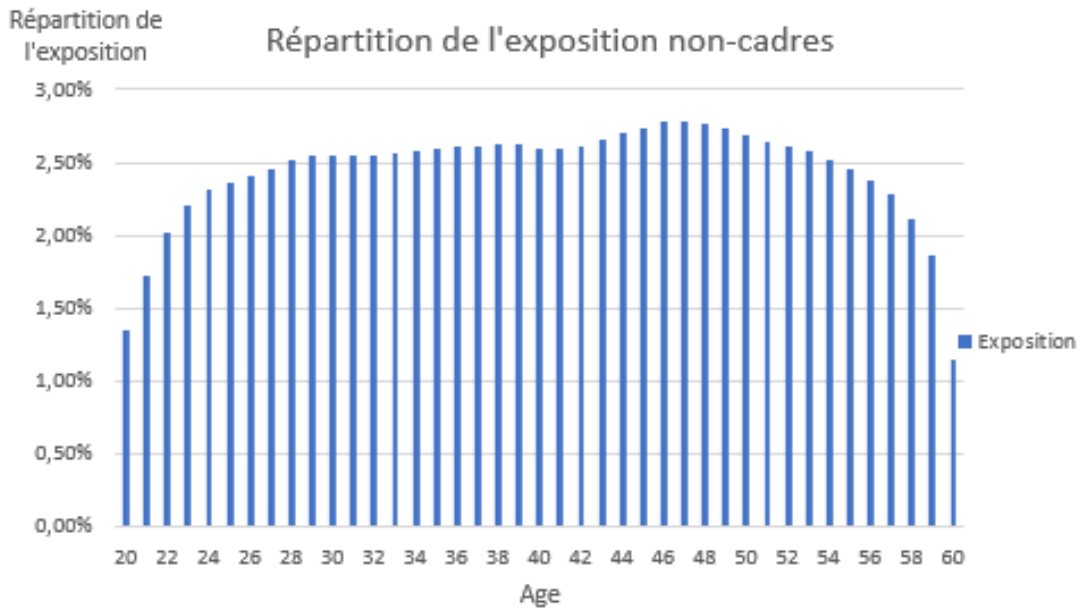


FIGURE 43 – Répartition de l'exposition des non-cadres en fonction de l'âge

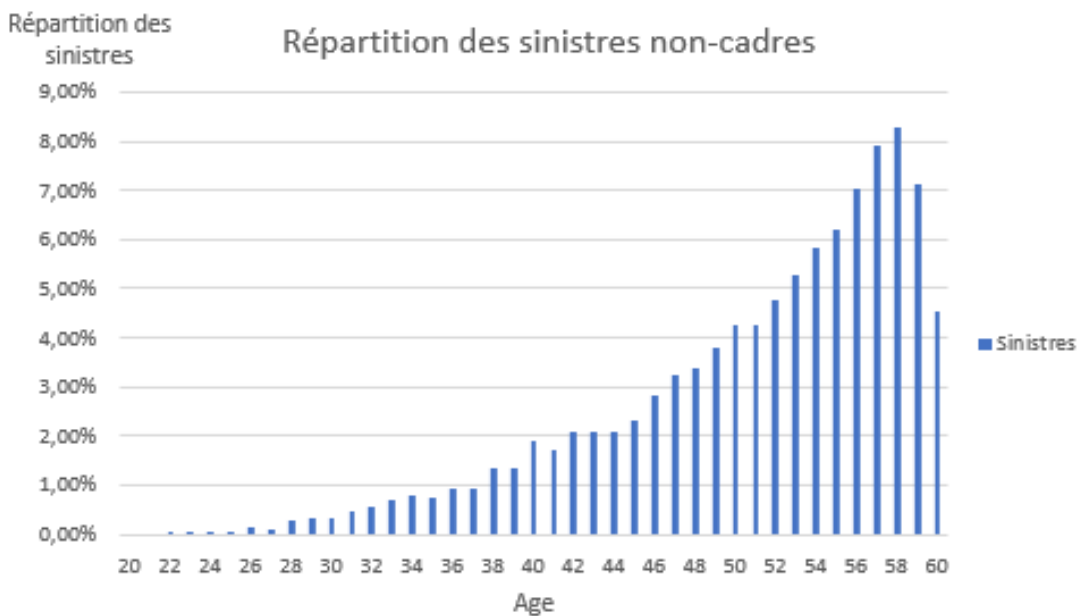


FIGURE 44 – Répartition des sinistres des non-cadres en fonction de l'âge

En termes de volume, nous savons déjà que le total s'élève à plus de 16 000 invalidités réparties comme suit : plus de 3 000 invalidités pour les cadres et un peu plus de 13 000

invalidités pour les non-cadres.

En dehors de ça, nous tirons les mêmes conclusions que pour les hommes et les femmes à savoir la présence d'un plateau pour la répartition des expositions, quelque peu plus large pour les non-cadres, et une sinistralité croissante avec l'âge.

9.2 Estimation des taux bruts d'entrée en invalidité

Pour l'estimation des taux bruts, nous avons encore opté pour l'estimateur de Hoem et exposons ci-après les résultats pour la population cadres :

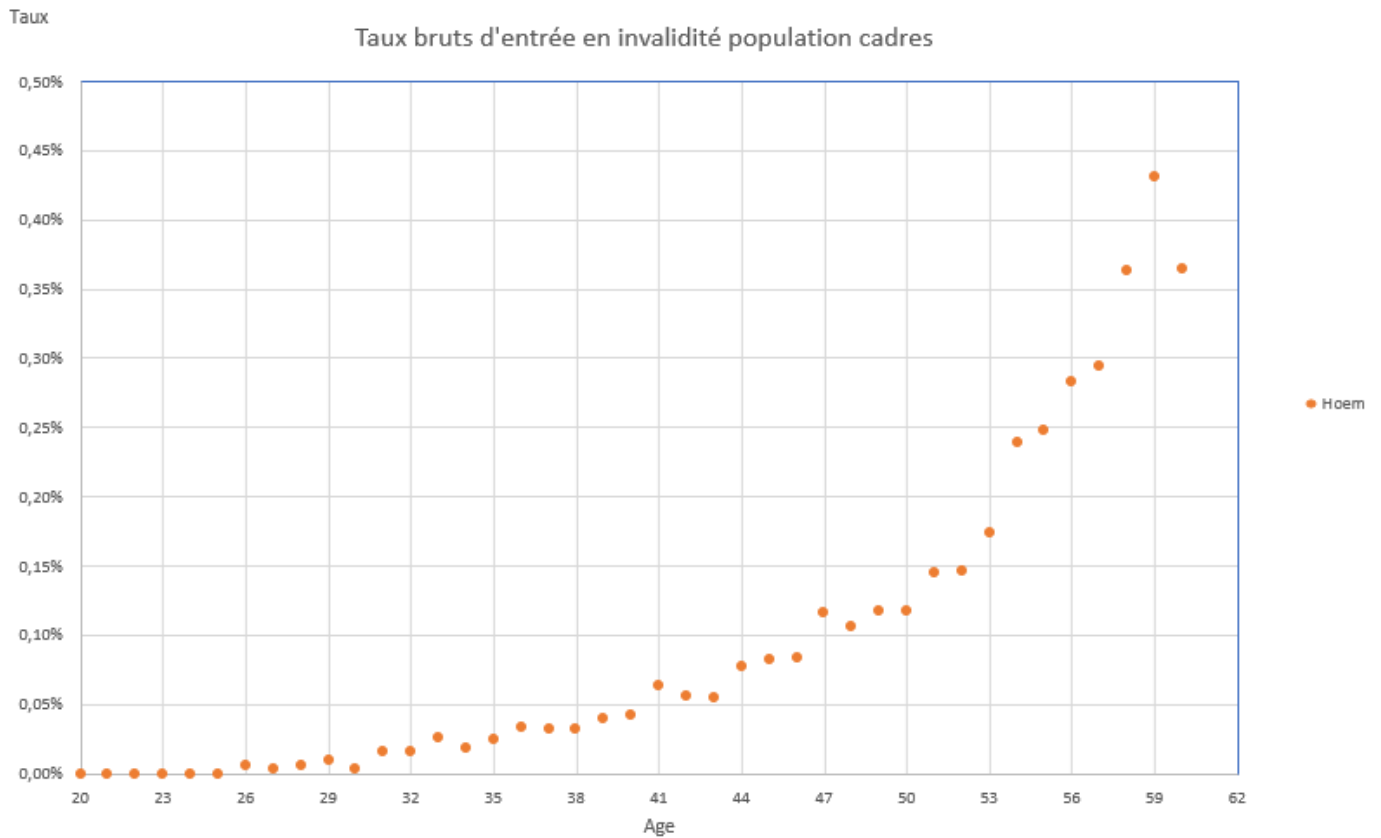


FIGURE 45 – Taux bruts d'entrée en invalidité pour la population cadres avec l'estimateur de Hoem

puis celles des non-cadres :

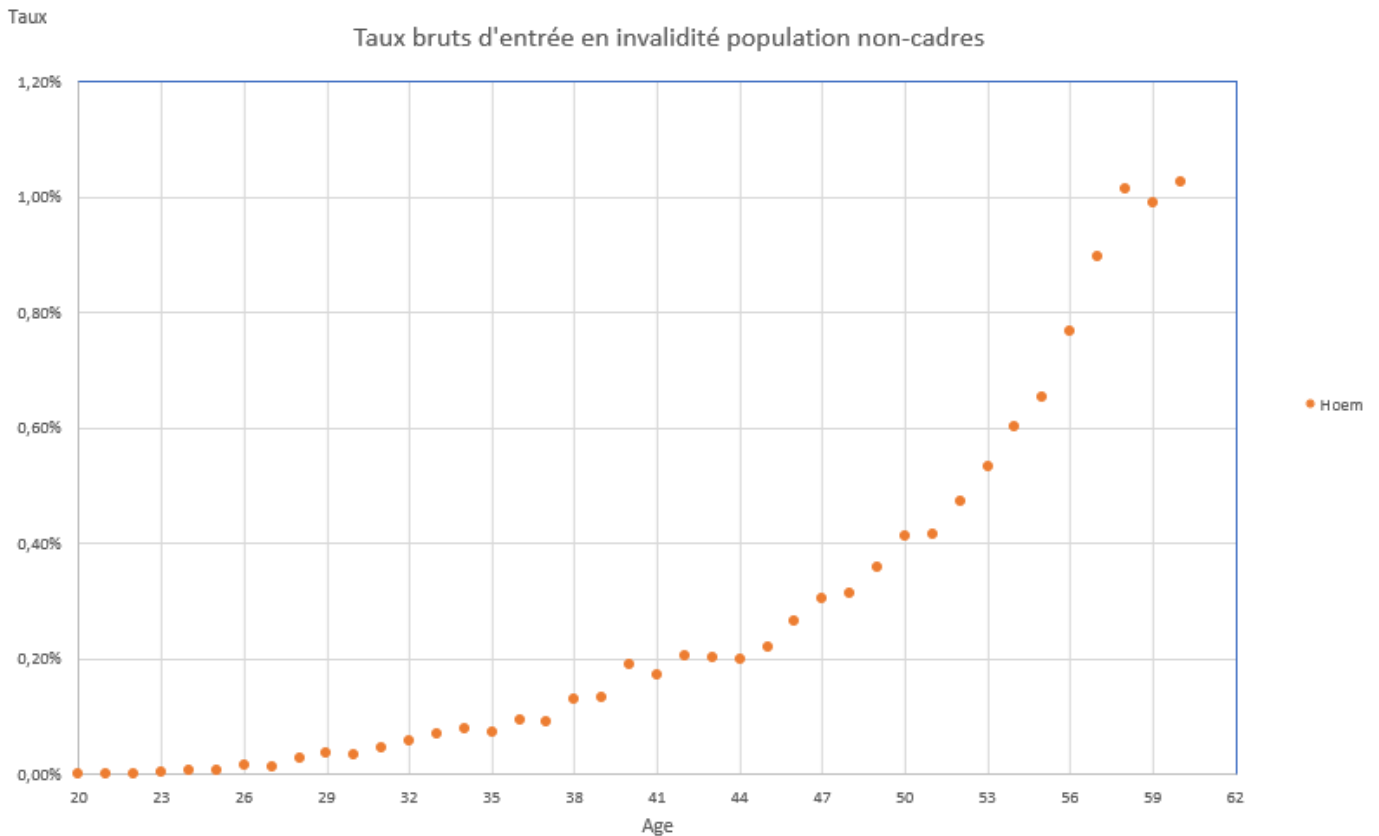


FIGURE 46 – Taux bruts d’entrée en invalidité pour la population non-cadres avec l’estimateur de Hoem

Comme nous pouvions nous y attendre, les taux sont bien plus élevés chez les non-cadres que chez les cadres, ils sont même parfois jusqu’à plus de deux fois plus élevés.

Rien que cette observation semble déjà justifier l’utilité de construire deux lois distinctes cadres et non-cadres et de les implémenter dans l’outil de tarification.

Les taux affichés ont là aussi été modifiés mais respectent les ordres de grandeur.

Notons qu’il est parfois possible d’utiliser le modèle de Cox qui est un modèle assez largement utilisé en analyse de survie et qui sert notamment à estimer les taux d’une partie de la population à partir des taux d’une autre partie de la population (typiquement les taux des cadres à partir de ceux des non-cadres). C’est de plus un modèle qui présente l’avantage de tenir compte des phénomènes de censure et de troncature.

Toutefois, il est souvent utilisé lorsque l’on manque de données sur une partie de la population. Or, comme expliqué plus haut, nous avons plus de 3 000 invalidités pour la

population cadres, c'est pourquoi nous avons considéré que ce volume était suffisamment important pour déterminer une loi cadre.

A titre indicatif, nous joignons en annexe les graphiques superposant les taux bruts avec la répartition de l'exposition pour les cadres.

Une fois de plus, la chute des taux à la toute fin coïncide avec une baisse importante de l'exposition. Passons maintenant aux lissages des taux.

9.3 Lissage des taux

Nous n'allons pas expérimenter de nouveau les lissages par splines cubiques, par moyennes mobiles ou encore par la méthode des noyaux car nous rencontrerions les mêmes limites que celles évoquées pour les femmes et les hommes. En effet, ce sont des limites intrinsèques aux lissages et qui ne dépendent donc pas des données étudiées (manque de valeurs lissées pour la méthode des noyaux et les moyennes mobiles et disparité trop importante pour le lissage par splines qui ne permet pas de valider le test du changement de signes notamment).

Pour ces raisons, nous allons donc nous limiter aux lissages de Whittaker-Henderson en prenant neuf configurations différentes avec le paramètre z valant 2, 3 ou 4 et le paramètre h valant 1, 100 ou 1000.

Nous exposons ci-après quelques courbes obtenues pour la population cadres :

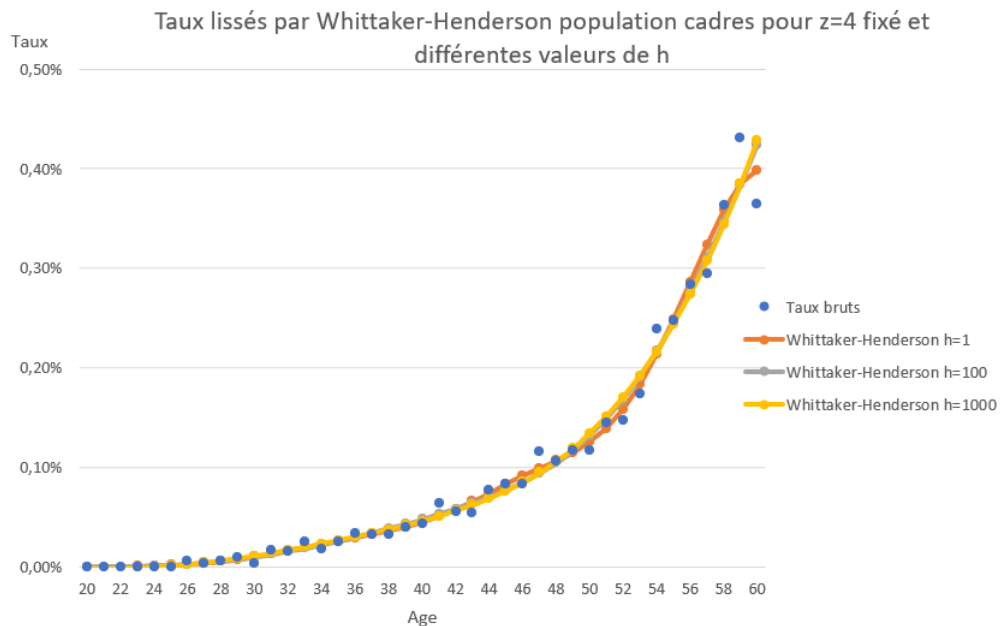


FIGURE 47 – Lissage par Whittaker-Henderson avec variations du paramètre h pour la population cadres

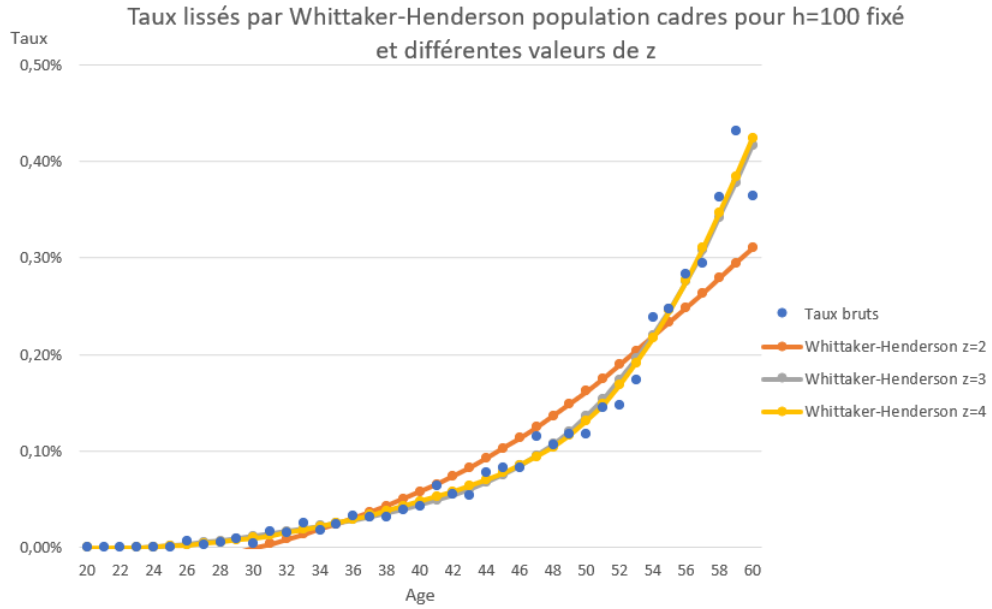


FIGURE 48 – Lissage par Whittaker-Henderson avec variations du paramètre z pour la population cadres

et pour les non-cadres :

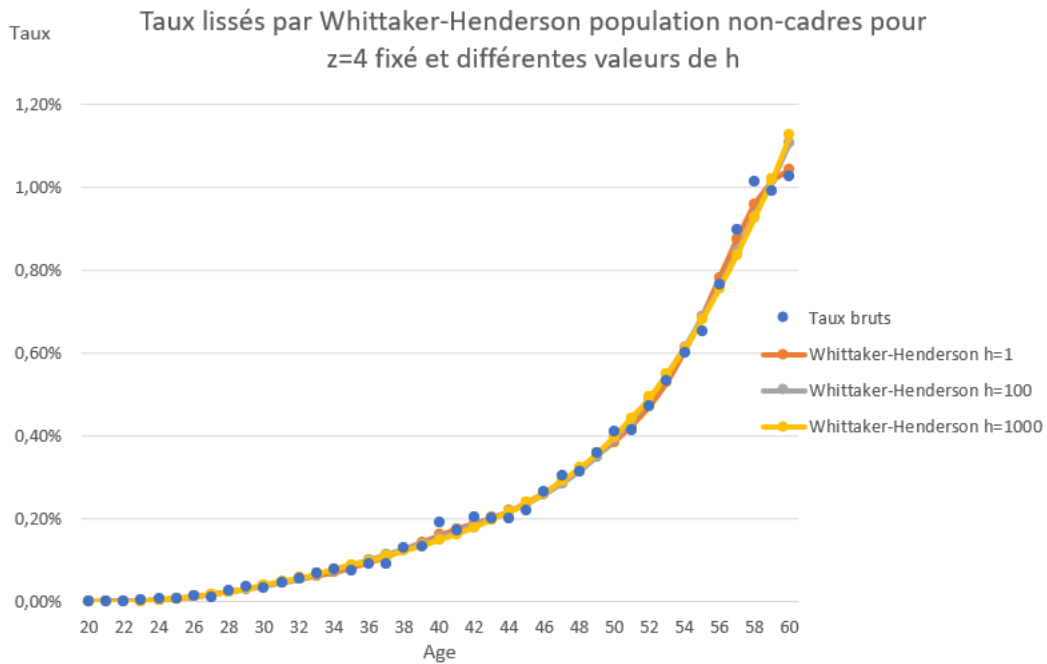


FIGURE 49 – Lissage par Whittaker-Henderson avec variations du paramètre h pour la population non-cadres

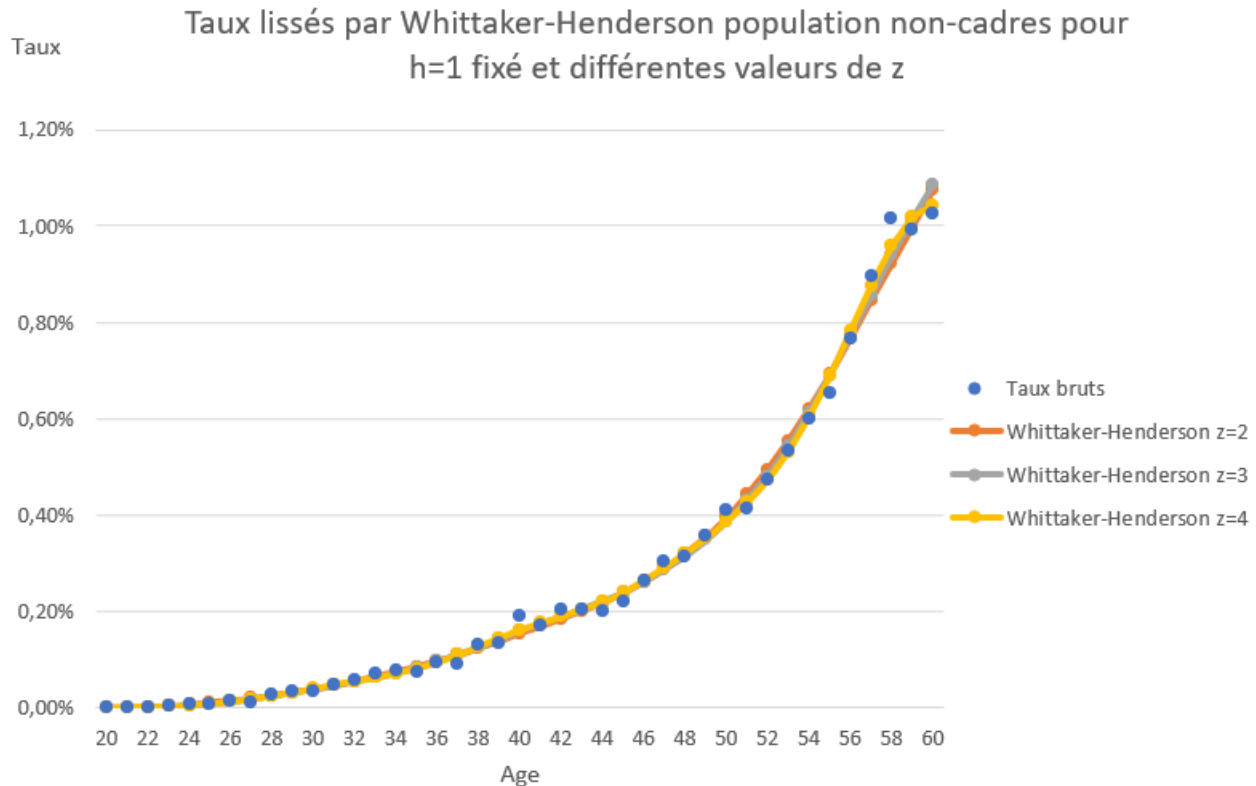


FIGURE 50 – Lissage par Whittaker-Henderson avec variations du paramètre z pour la population non-cadres

Encore une fois étant donné le volume de données, les taux bruts forment déjà une courbe relativement lisse réduisant donc l'impact des paramètres du lissage sauf pour des âges assez élevés notamment sur la population cadres.

Pour choisir les paramètres z et h définitivement, nous effectuons des tests de lissage.

9.4 Tests de lissage et validation de la table

Comme les tests de lissage sont les mêmes que ceux réalisés pour les femmes et les hommes, nous ne faisons pas apparaître les parties théoriques et passons directement aux résultats.

Pour le test du SMR, tous les lissages ont un SMR de 1 peu importe les paramètres encore une fois.

Cette qualité de précision provient une fois de plus des DSN très utilisées qui permettent de fournir la quasi-totalité des expositions pour les invalidités enregistrées. Il en résulte de fait que le produit de l'exposition par les taux lissés donne le nombre exact d'invalidités.

Quant au test du changement de signes, il est davantage discriminant : nous présentons l'ensemble des résultats de ce test en annexe pour les population cadres et non-cadres.

En tenant compte de cela, de l'allure des courbes et de la tendance qui se dessinait sur les âges plus importants, nous avons retenu les paramètres $h=100$ et $z=4$ pour les cadres et $h=1$ et $z=4$ pour les non-cadres, soient les courbes des taux lissés définitifs suivantes :

pour les cadres :

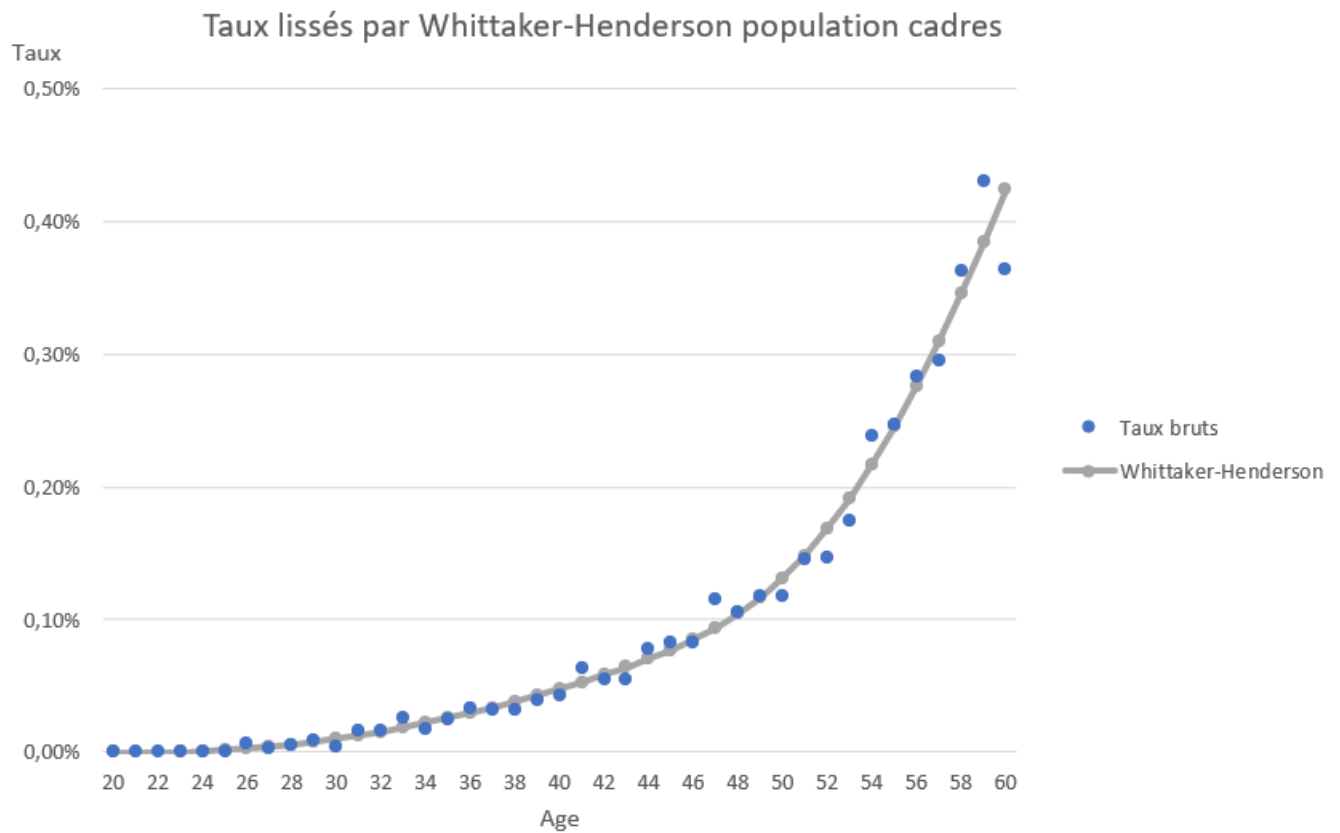


FIGURE 51 – Taux lissés définitifs pour la population cadres

et pour les non-cadres :

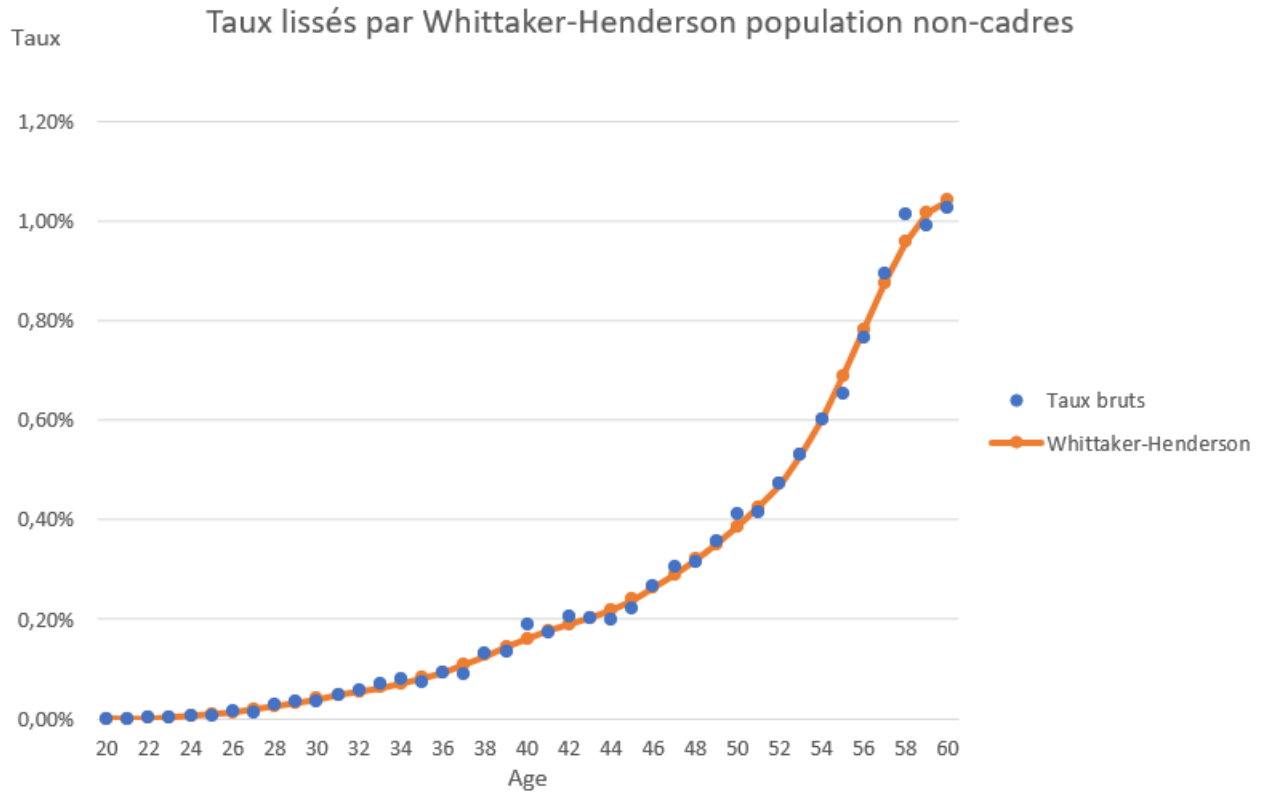


FIGURE 52 – Taux lissés définitifs pour la population non-cadres

9.5 Tarification

Pour la tarification, nous avons procédé de manière analogue en calculant d’une part le tarif individuel pour un âge fixé, puis ensuite le tarif groupe en tenant compte de la dispersion des âges en ayant restreint notre population aux cadres puis aux non-cadres.

En revanche, notons une fois de plus que nous engendrons le biais suivant lors du calcul de notre tarif groupe : comme nous séparons les cadres et les non-cadres mais que nous considérons l’âge moyen des entreprises tout individu confondu (cadres et non cadres), il peut y avoir dans la réalité un âge moyen des cadres supérieurs à celui des non-cadres ou inversement... Là encore, nous prenons en compte le salaire dans le calcul du tarif.

Voici les résultats pour les cadres :

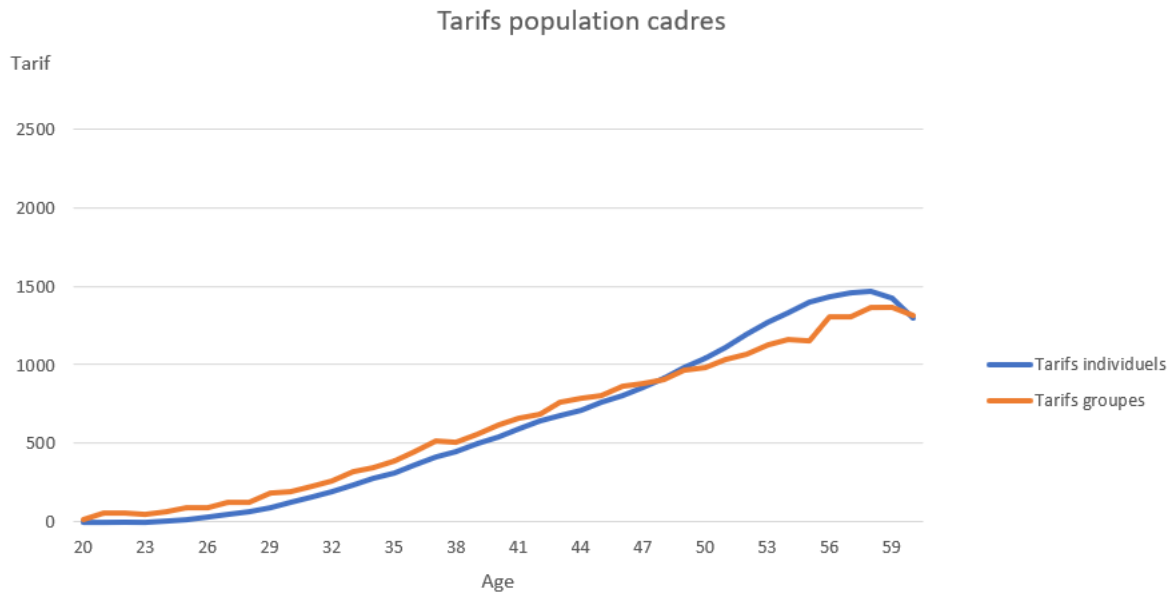


FIGURE 53 – Comparaison du tarif individuel et du tarif groupe pour la population cadres

et pour les non-cadres :

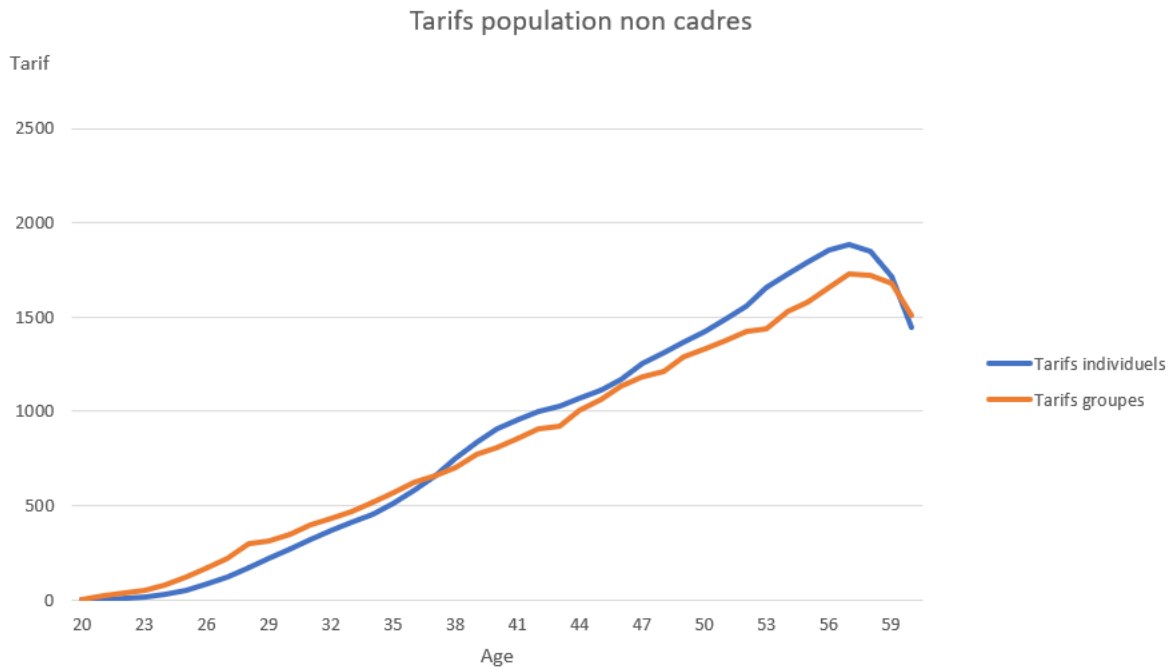


FIGURE 54 – Comparaison du tarif individuel et du tarif groupe pour la population non-cadres

Le comportement global du tarif groupe par rapport au tarif individuel reste sensiblement le même que pour les populations femmes et hommes : à savoir une sous tarification pour les âges faibles et inversement pour les âges plus élevés.

De plus, le tarif groupe de la population cadres semble souffrir du manque de données sur les âges élevés étant donnés les brusques variations de la courbe...

Quant à la fin de la distribution du tarif groupe, elle semble une fois de plus complètement en accord avec le tarif individuel. Ce dernier présente une forme de cloche du fait de la compensation du coût par rapport à l'incidence et au salaire.

9.6 Conclusion sur les lois cadres/non-cadres

En conclusion, la segmentation cadres/non-cadres apparait de toute évidence importante au vu de l'écart très significatif entre les taux d'incidence de chacune de ces deux lois (parfois plus de 2,5 fois plus d'incidence chez les non-cadres).

Comme nous avons vu qu'il existait aussi des différences entre les hommes et les femmes, l'idéal serait donc de mettre en place quatre lois pour chacune des catégories citées.

Quoi qu'il en soit, le travail sur cette segmentation de la population semble avoir du sens et répondre à l'idée que l'on pouvait s'en faire intuitivement.

10 Conclusion

La construction d'une table d'incidence en invalidité en assurances collectives nous a permis de comprendre d'une part les étapes successives aboutissant à la construction d'une loi et d'autre part les difficultés propres au milieu des assurances collectives : nécessité de penser à une loi à la maille individu + numéro de RPP car l'utilité finale est la tarification et s'appuie sur ce principe intrinsèque.

Nous avons eu l'occasion de retraiter en conséquence une base de données afin de satisfaire à cette maille puis ensuite d'utiliser des modèles d'analyse de survie que nous avons pu comparer d'un point de vue théorique afin de justifier nos décisions.

La multitude de lissages effectués et de tests sur ces derniers nous a permis de saisir les avantages et inconvénients liés à chacun de ces lissages : ceux qui accordent trop d'importance à la régularité de la courbe lissée ou ceux qui accordent trop d'importance aux observations brutes (notion de fidélité). Sans surprise, nous avons choisi le lissage de Whittaker-Henderson qui, avec ses deux paramètres, permet de jouer justement sur l'importance accordée à tel ou tel critère.

Enfin, nous avons pu voir l'application concrète de l'élaboration de ces lois en allant jusqu'à la notion de tarification, mettant en jeu celle d'âge actuariel, ce qui présente l'avantage de voir les fruits et donc l'impact direct de notre travail statistique. En somme, nous avons mis en jeu nos capacités à répondre à une problématique de l'entreprise en construisant des lois conformes à ce qui existe dans l'outil de tarification et dans un second temps nos capacités dites statistiques pour proposer une autre alternative plus fine qui devrait permettre d'affiner le tarif propre à la garantie invalidité.

Références

- [1] Alan. <https://alan.com/fr-fr/assurance-sante/prevoyance-entreprise/a/differents-contrats-prevoyance>.
- [2] Allianz. <https://espaceclient.allianz.fr/pmt/guide/Prevoyance>.
- [3] Améli. <https://www.ameli.fr/assure/droits-demarches/invalidite-handicap/invalidite>.
- [4] AtouSante. <https://www.atousante.com/apptitude-inaptitude/amenagement-poste-reclassement/amenagement-invalidite/categories-invalidite/>.
- [5] AXA. <https://www.axa.fr/qui-sommes-nous.html>.
- [6] Wikipédia AXA. <https://fr.wikipedia.org/wiki/Axa>.
- [7] Marie Line Chabanol. <https://www.math.u-bordeaux.fr/~mchabano/Agreg/ProbaAgreg1213-COURS1-Statordre.pdf>.
- [8] Romain Decamps. Création et étude d'une table d'incidence en invalidité. *Mémoire*, 2015.
- [9] INSEE. <https://www.insee.fr/fr/accueil>.
- [10] Inserm. <https://www.cepidc.inserm.fr/documentation/tests-statistiques-relatifs-aux-indicateurs-de-mortalite-en-population>.
- [11] O.Lopez. <https://moodle-sciences.upmc.fr/moodle-2021/course/view.php?id=3257>.
- [12] Service Public. <https://entreprendre.service-public.fr/vosdroits/F34059>.
- [13] Philippe Saint-Pierre. Introduction à l'analyse des durées de survie. *Cours*, 2021.
- [14] Stringfixer. <https://stringfixer.com/fr/Kernelregression>.
- [15] Thomas Guillon Verne. Construction de tables de mortalité d'expérience sur de petits échantillons pour l'estimation de la sinistralité décès. *Mémoire*, 2010.
- [16] Jewell N.P. Wang M.-C. and Tsai W.-Y. Asymptotic properties of the product limit estimate under random truncation. *the annals of statistics*. *Cours*, 1986.
- [17] Wikipédia. <https://fr.wikipedia.org/wiki/Spline>.

Table des figures

1	Taux lissés définitifs pour la population femmes	6
2	Taux lissés définitifs pour la population hommes	6
3	Taux lissés définitifs pour la population femmes	7
4	Taux lissés définitifs pour la population hommes	8
5	Final smoothed rates for the female population	10
6	Final smoothed rates for the male population	10
7	Final smoothed rates for the managerial population	11
8	Final smoothed rates for the non-managerial population	12
9	Table de correspondance	21
14	Comparaison des répartitions des invalidités en portefeuille et dans la population active par tranches d'âge	25
15	Schéma représentant le phénomène de censure	27
16	Schéma représentant le phénomène de troncature	29
17	Répartition de l'exposition en fonction de l'âge	33
18	Répartition des sinistres en fonction de l'âge	34
19	Taux bruts d'entrée en invalidité pour l'ensemble de la population avec les estimateurs de Kaplan-Meier et de Hoem	35
20	Taux bruts estimés par Hoem et exposition en fonction de l'âge pour l'ensemble de la population	36
21	Lissage par splines cubiques pour la population femmes	39
22	Lissage par splines cubiques pour la population hommes	39
23	Lissage par Whittaker-Henderson avec variation du paramètre h pour la population femmes	42
24	Lissage par Whittaker-Henderson avec variation du paramètre z pour la population femmes	42
25	Lissage par Whittaker-Henderson avec variation du paramètre h pour la population hommes	43
26	Lissage par Whittaker-Henderson avec variation du paramètre z pour la population hommes	44
27	Lissage par moyenne mobile pour la population femmes	45
28	Lissage par moyenne mobile pour la population hommes	45
29	Lissage par moyenne mobile écrêtée pour la population femmes	46
30	Lissage par moyenne mobile écrêtée pour la population hommes	47
31	Forme des 3 noyaux utilisés	49

32	Comparaison des lissages par noyaux pour la population femmes	50
33	Comparaison des lissages par noyaux pour la population hommes	50
34	Taux lissés définitifs pour la population femmes	55
35	Taux lissés définitifs pour la population hommes	56
36	Comparaison des taux lissés des deux études pour la population femmes . . .	57
37	Comparaison des taux lissés des deux études pour la population hommes . .	58
38	Comparaison du nombre d'invalidités total prédit et du nombre d'invalidités total réel	58
39	Comparaison du tarif individuel et du tarif groupe pour la population femmes	60
40	Comparaison du tarif individuel et du tarif groupe pour la population hommes	61
41	Répartition de l'exposition des cadres en fonction de l'âge	64
42	Répartition des sinistres des cadres en fonction de l'âge	64
43	Répartition de l'exposition des non-cadres en fonction de l'âge	65
44	Répartition des sinistres des non-cadres en fonction de l'âge	65
45	Taux bruts d'entrée en invalidité pour la population cadres avec l'estimateur de Hoem	66
46	Taux bruts d'entrée en invalidité pour la population non-cadres avec l'estima- teur de Hoem	67
47	Lissage par Whittaker-Henderson avec variations du paramètre h pour la po- pulation cadres	68
48	Lissage par Whittaker-Henderson avec variations du paramètre z pour la po- pulation cadres	69
49	Lissage par Whittaker-Henderson avec variations du paramètre h pour la po- pulation non-cadres	69
50	Lissage par Whittaker-Henderson avec variations du paramètre z pour la po- pulation non-cadres	70
51	Taux lissés définitifs pour la population cadres	71
52	Taux lissés définitifs pour la population non-cadres	72
53	Comparaison du tarif individuel et du tarif groupe pour la population cadres	73
54	Comparaison du tarif individuel et du tarif groupe pour la population non-cadres	73
55	Répartition de l'exposition des femmes en fonction de l'âge	80
56	Répartition des sinistres des femmes en fonction de l'âge	80
57	Taux bruts d'entrée en invalidité pour l'ensemble de la population femmes avec les estimateurs de Kaplan-Meier et de Hoem	81
58	Répartition de l'exposition des hommes en fonction de l'âge	82

59	Répartition des sinistres des hommes en fonction de l'âge	82
60	Taux bruts d'entrée en invalidité pour l'ensemble de la population hommes avec les estimateurs de Kaplan-Meier et de Hoem	83
61	Taux bruts estimés par Hoem et exposition en fonction de l'âge pour la population femmes	84
62	Taux bruts estimés par Hoem et exposition en fonction de l'âge pour la population hommes	84
63	Taux bruts estimés par Hoem et exposition en fonction de l'âge pour la population cadres	85
64	Taux bruts estimés par Hoem et exposition en fonction de l'âge pour la population non-cadres	85
65	Comparaison des moyennes mobiles pour la population femmes	86
66	Comparaison des moyennes mobiles pour la population hommes	86
67	Lissage par noyau gaussien avec variation du paramètre h pour la population femmes	87
68	Lissage par noyau gaussien avec variation du paramètre h pour la population hommes	87
69	Lissage par noyau d'Epanechnikov avec variation du paramètre h pour la population femmes	88
70	Lissage par noyau d'Epanechnikov avec variation du paramètre h pour la population hommes	88
71	Lissage par noyau cubique avec variation du paramètre h pour la population femmes	89
72	Lissage par noyau cubique avec variation du paramètre h pour la population hommes	89
73	Test à distance finie	90
74	Test à distance finie population femme	91
75	Test à distance finie population homme	92
76	Test à distance finie population cadre	93
77	Test à distance finie population non cadre	93
78	Test SMR population femme	94
79	Test SMR population homme	95

Annexes

Estimation des taux bruts

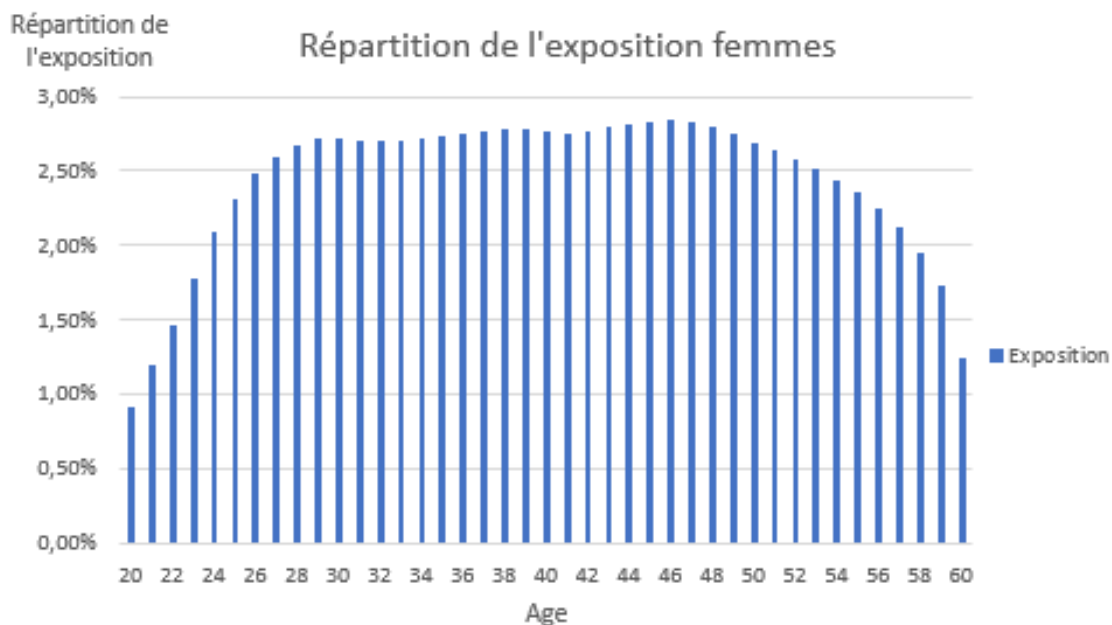


FIGURE 55 – Répartition de l'exposition des femmes en fonction de l'âge

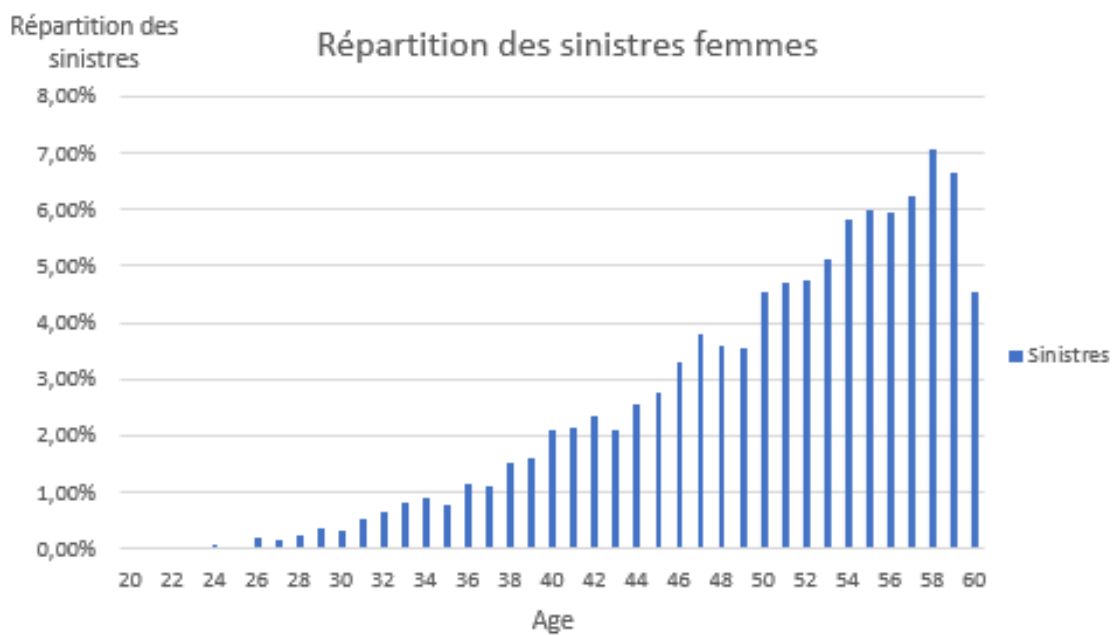


FIGURE 56 – Répartition des sinistres des femmes en fonction de l'âge

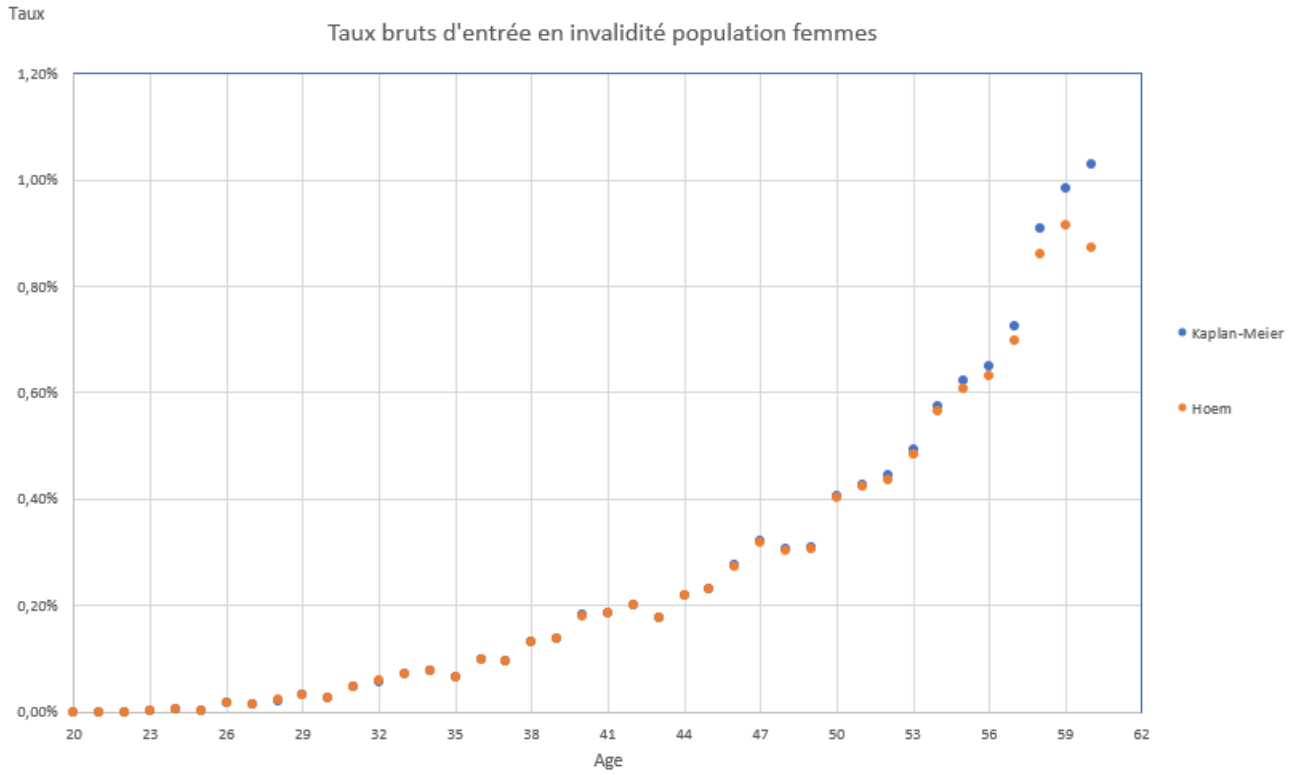


FIGURE 57 – Taux bruts d'entrée en invalidité pour l'ensemble de la population femmes avec les estimateurs de Kaplan-Meier et de Hoem

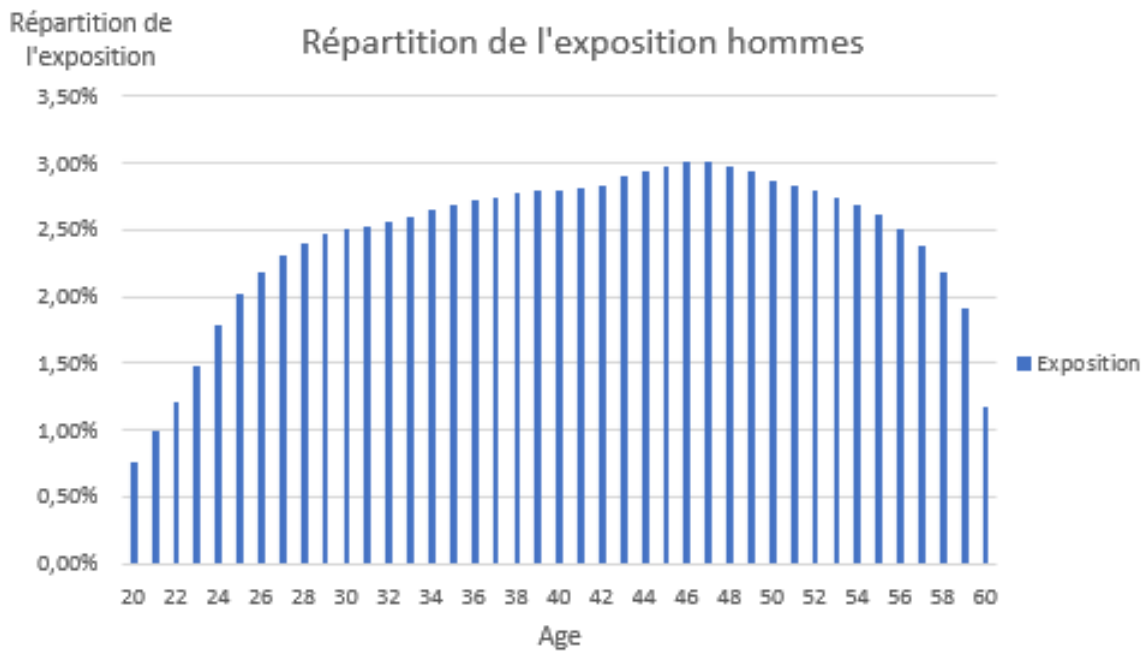


FIGURE 58 – Répartition de l'exposition des hommes en fonction de l'âge

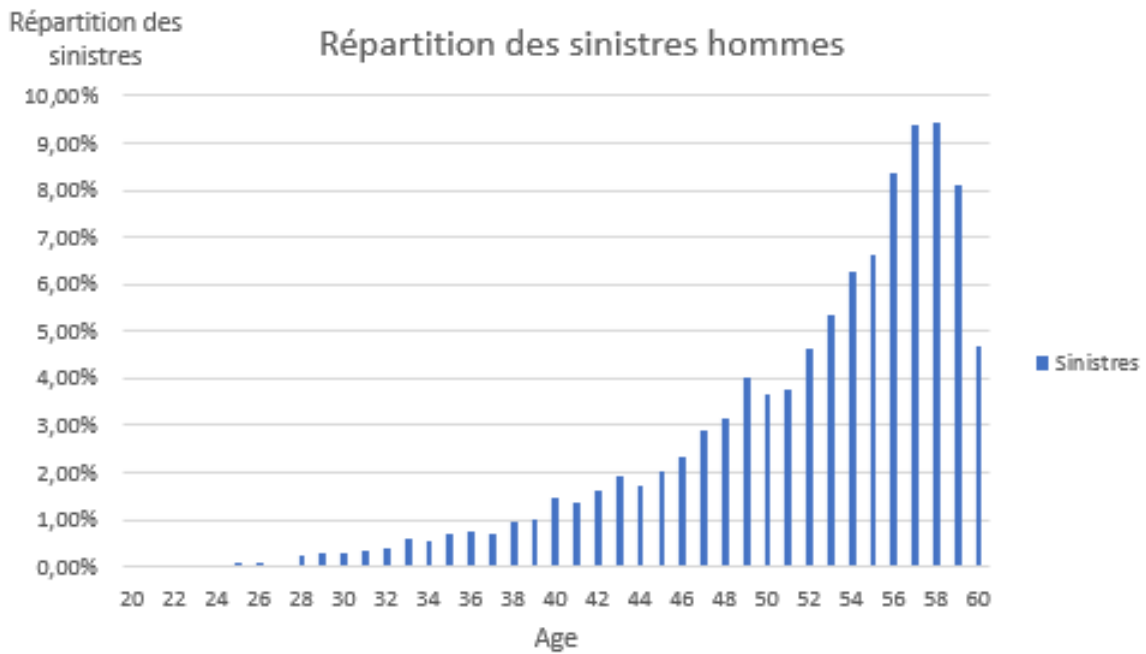


FIGURE 59 – Répartition des sinistres des hommes en fonction de l'âge

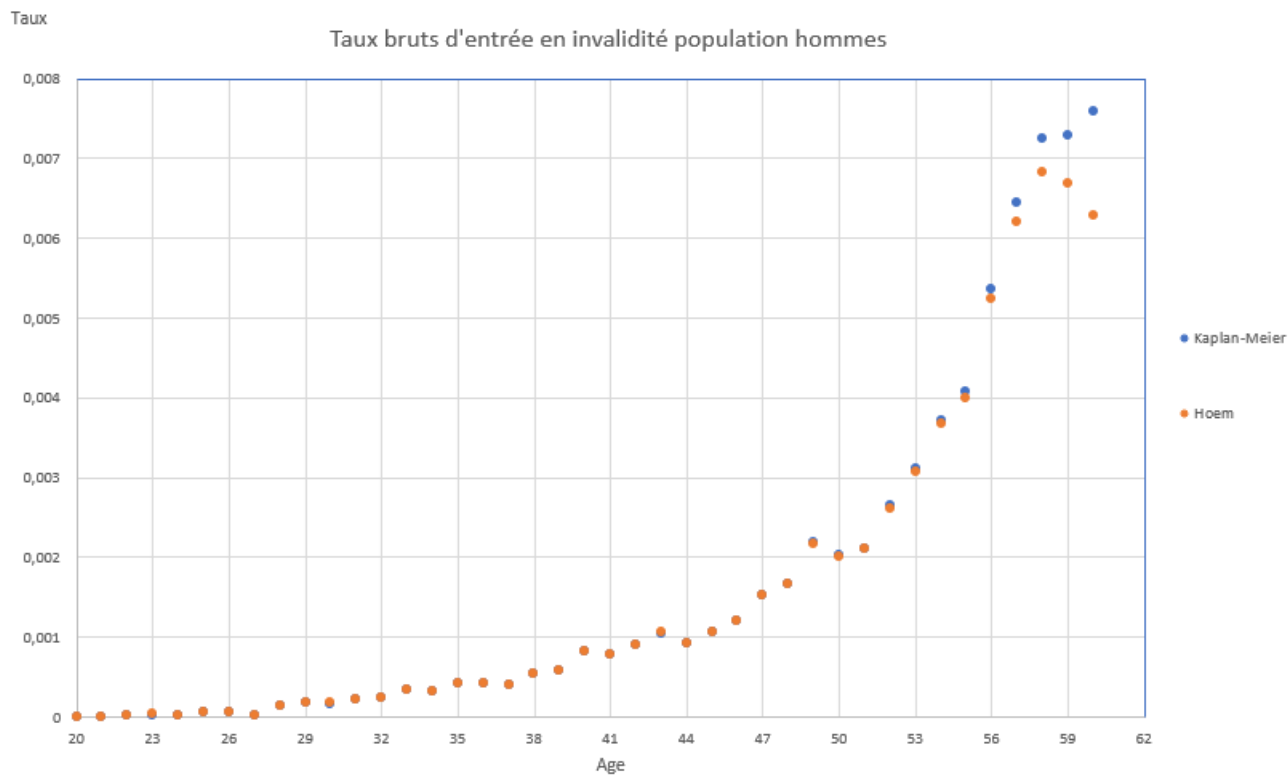


FIGURE 60 – Taux bruts d’entrée en invalidité pour l’ensemble de la population hommes avec les estimateurs de Kaplan-Meier et de Hoem

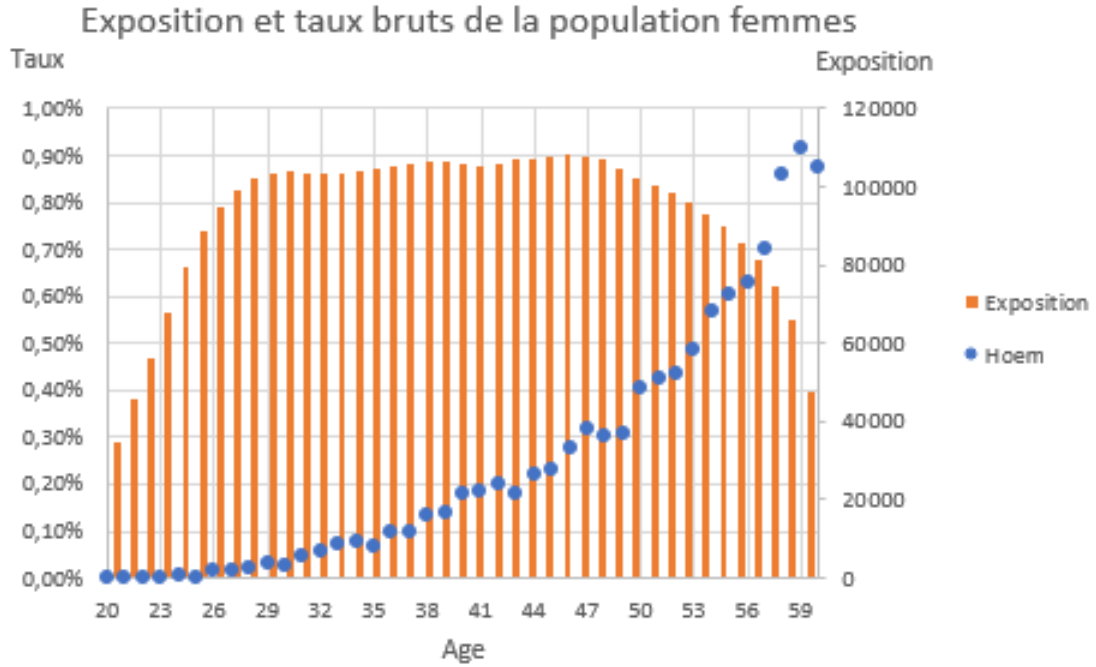


FIGURE 61 – Taux bruts estimés par Hoem et exposition en fonction de l’âge pour la population femmes

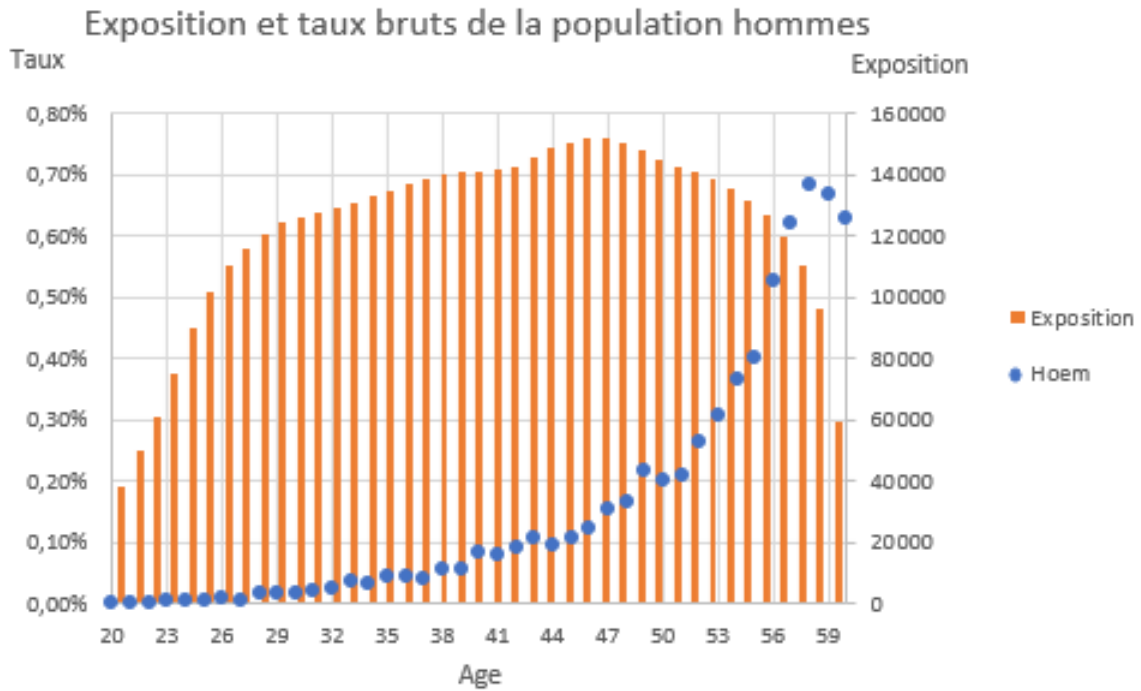


FIGURE 62 – Taux bruts estimés par Hoem et exposition en fonction de l’âge pour la population hommes

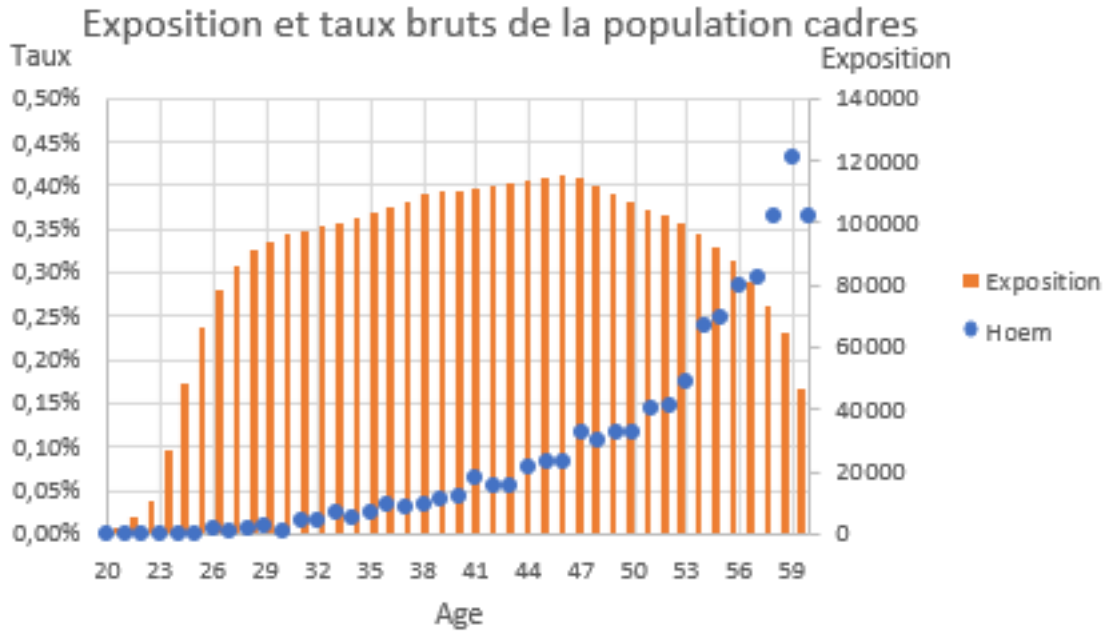


FIGURE 63 – Taux bruts estimés par Hoem et exposition en fonction de l’âge pour la population cadres

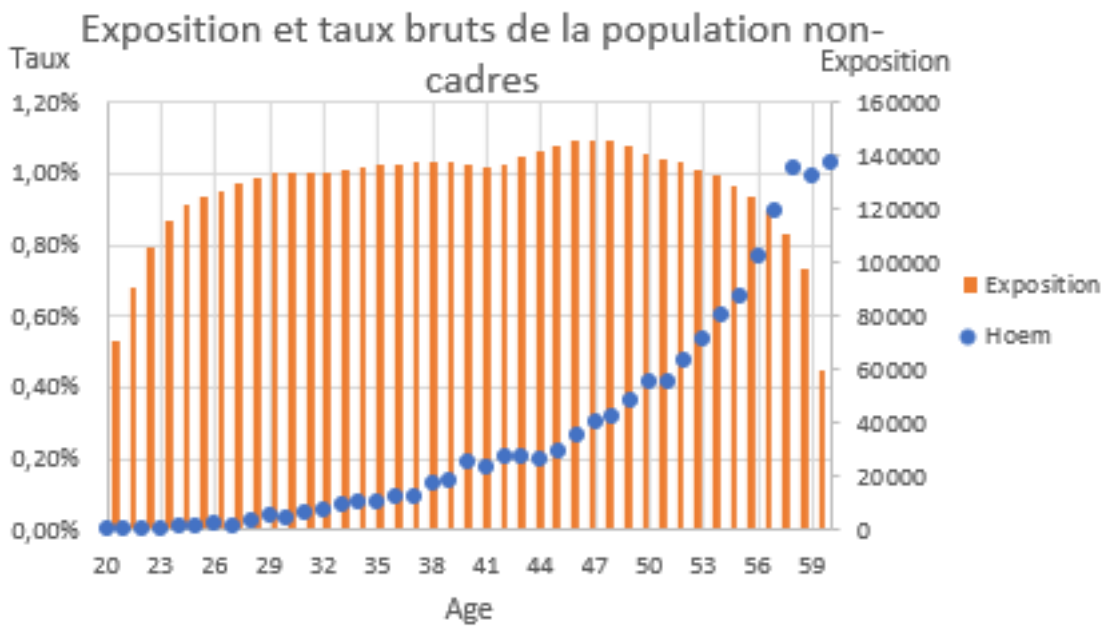


FIGURE 64 – Taux bruts estimés par Hoem et exposition en fonction de l’âge pour la population non-cadres

Lissage des taux

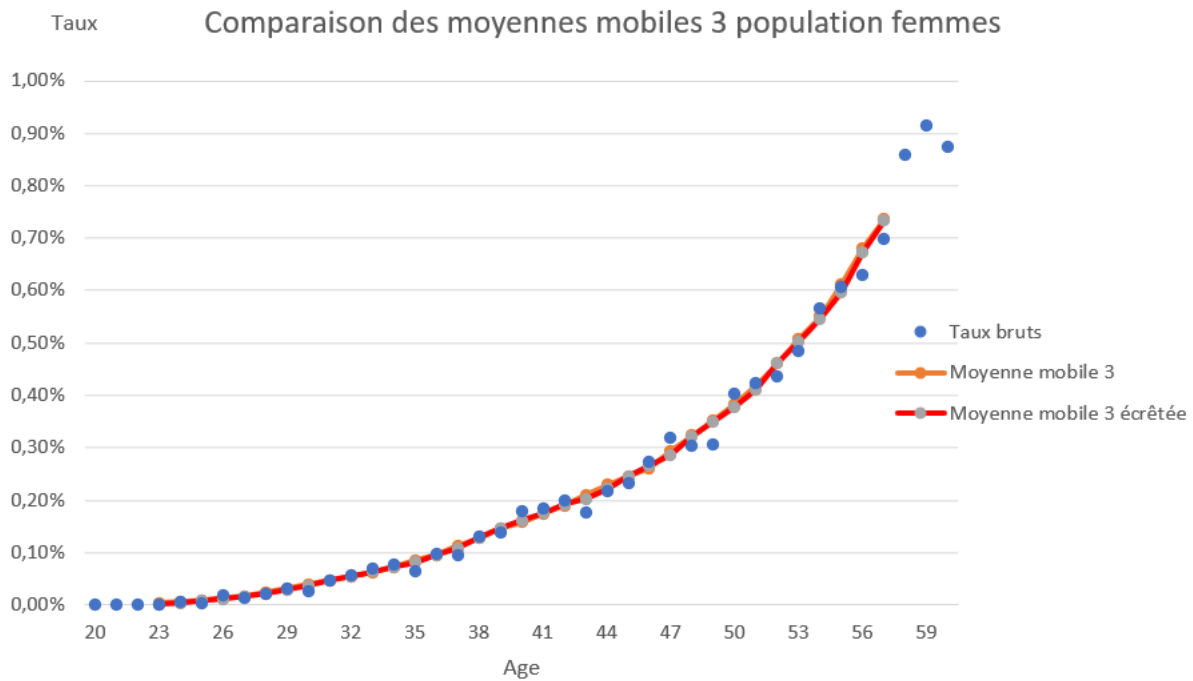


FIGURE 65 – Comparaison des moyennes mobiles pour la population femmes

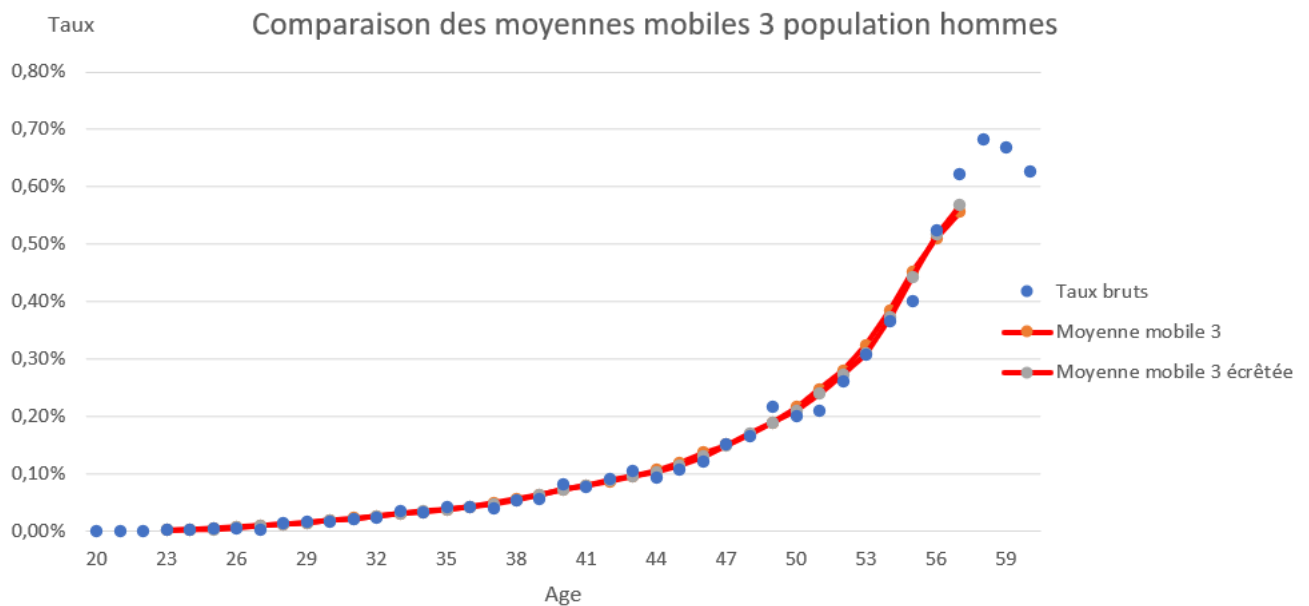


FIGURE 66 – Comparaison des moyennes mobiles pour la population hommes

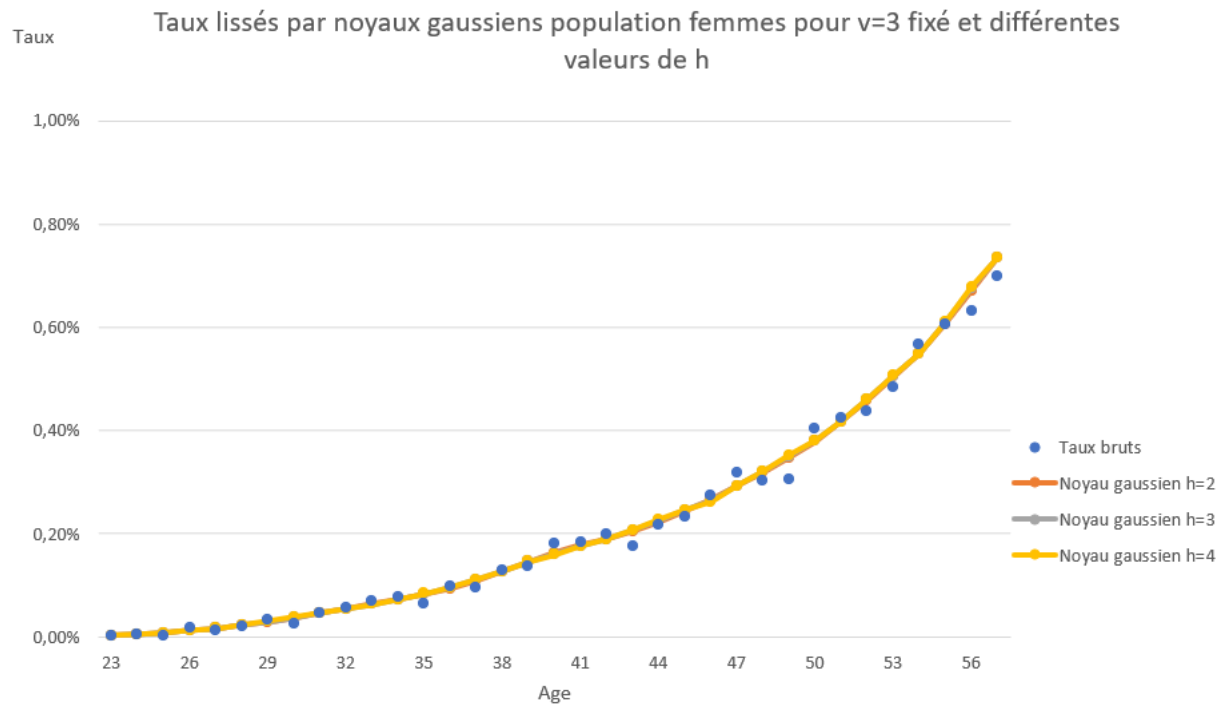


FIGURE 67 – Lissage par noyau gaussien avec variation du paramètre h pour la population femmes

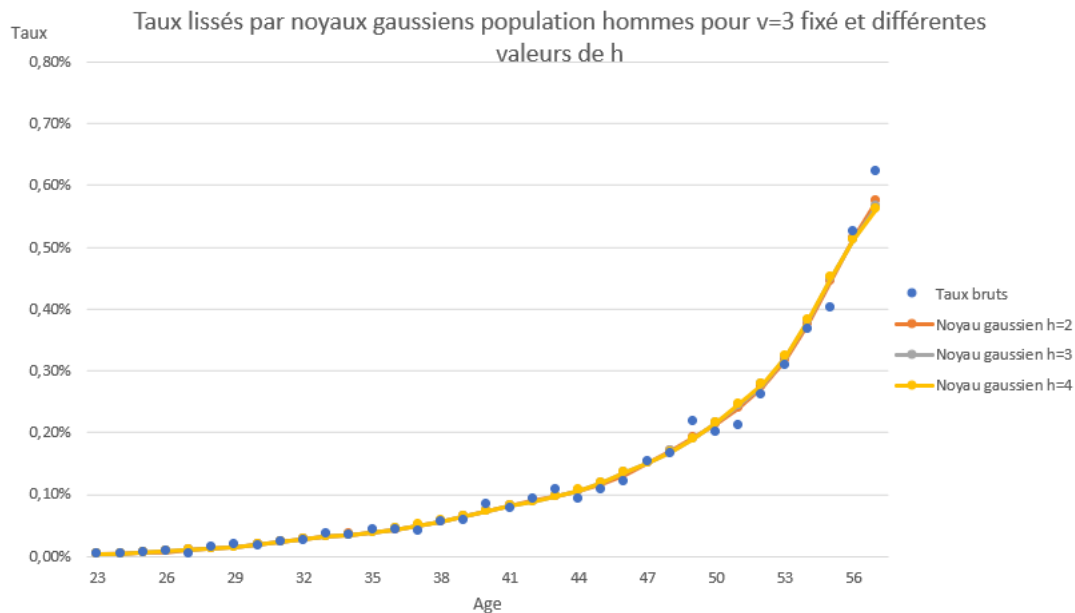


FIGURE 68 – Lissage par noyau gaussien avec variation du paramètre h pour la population hommes

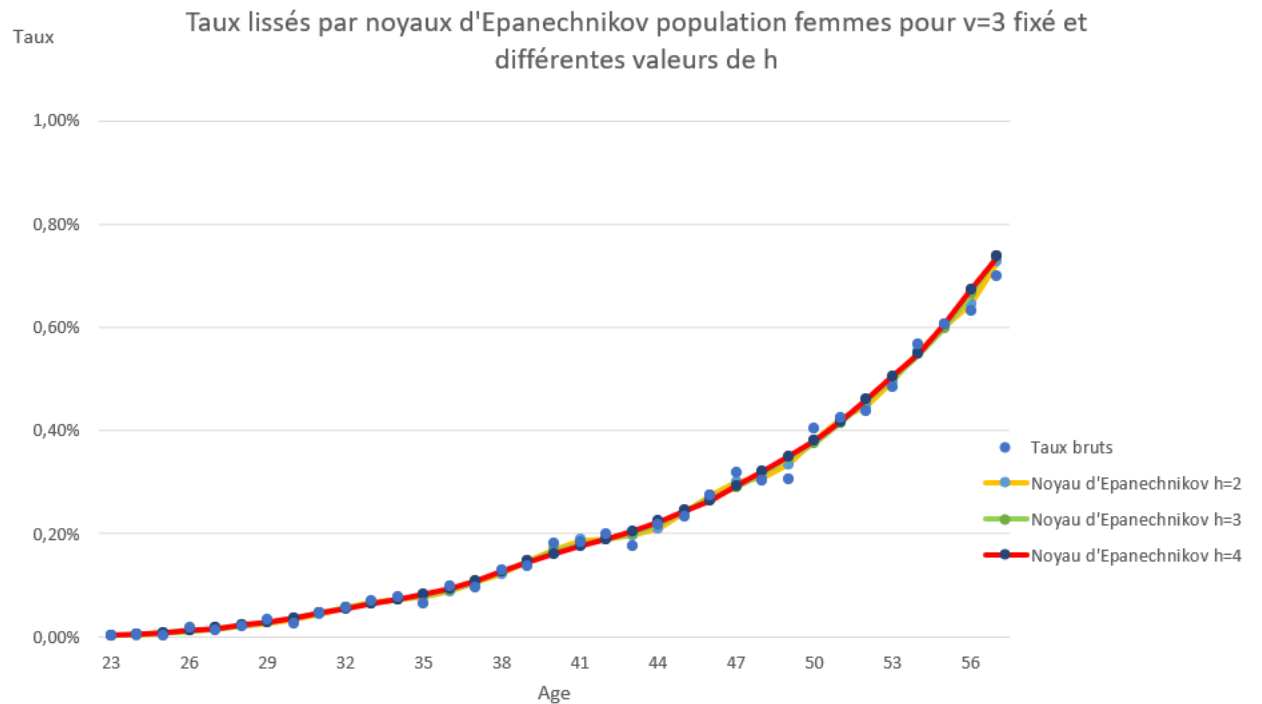


FIGURE 69 – Lissage par noyau d'Epanechnikov avec variation du paramètre h pour la population femmes

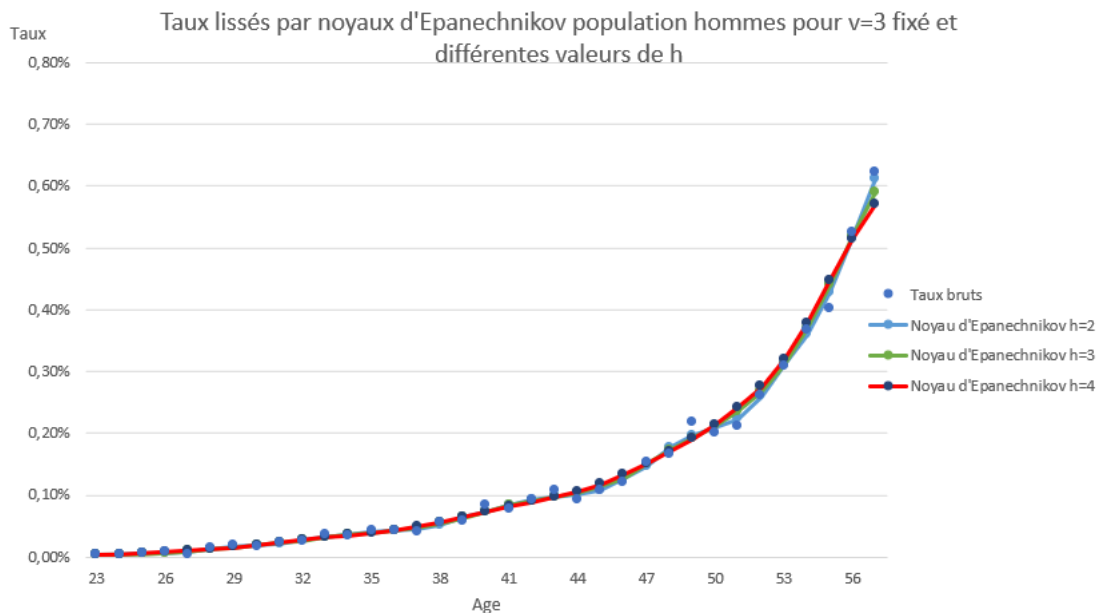


FIGURE 70 – Lissage par noyau d'Epanechnikov avec variation du paramètre h pour la population hommes

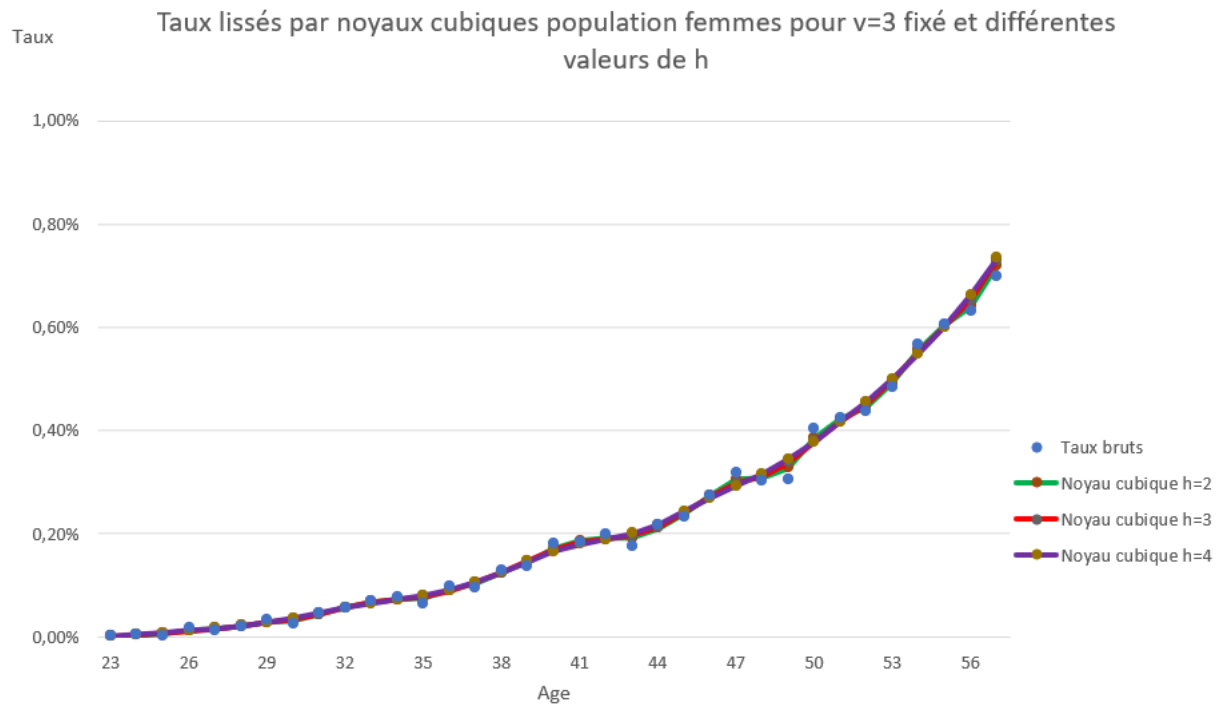


FIGURE 71 – Lissage par noyau cubique avec variation du paramètre h pour la population femmes

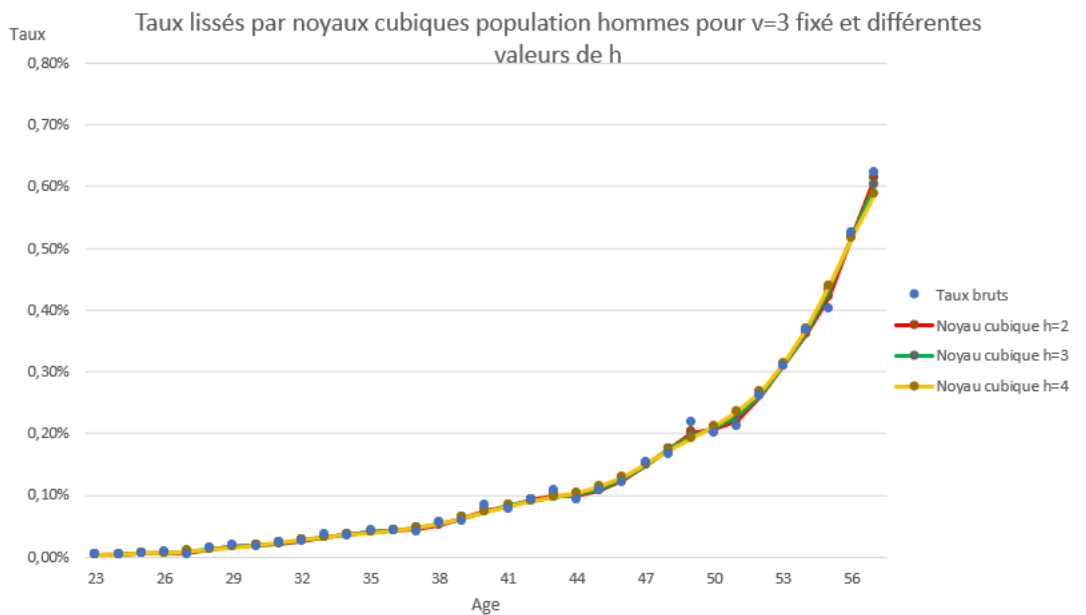


FIGURE 72 – Lissage par noyau cubique avec variation du paramètre h pour la population hommes

Tests de lissages

k	P(X=k)	F(X≤k)
0	9,09E-13	9,09E-13
1	3,64E-11	3,73E-11
2	7,09E-10	7,47E-10
3	8,99E-09	9,73E-09
4	8,31E-08	9,29E-08
5	5,98E-07	6,91E-07
6	3,49E-06	4,18E-06
7	1,7E-05	2,11E-05
8	6,99E-05	9,11E-05
9	0,000249	0,00034
10	0,000771	0,001111
11	0,002103	0,003213
12	0,005081	0,008295
13	0,010944	0,019239
14	0,021107	0,040345
15	0,036585	0,07693
16	0,057164	0,134094
17	0,080702	0,214795
18	0,103119	0,317914
19	0,119401	0,437315
20	0,125371	0,562685
21	0,119401	0,682086
22	0,103119	0,785205
23	0,080702	0,865906
24	0,057164	0,92307
25	0,036585	0,959655
26	0,021107	0,980761
27	0,010944	0,991705
28	0,005081	0,996787
29	0,002103	0,998889
30	0,000771	0,99966
31	0,000249	0,999909
32	6,99E-05	0,999979
33	1,7E-05	0,999996
34	3,49E-06	0,999999
35	5,98E-07	1
36	8,31E-08	1
37	8,99E-09	1
38	7,09E-10	1
39	3,64E-11	1
40	9,09E-13	1

FIGURE 73 – Test à distance finie

Types de lissages	Nombre de changements de signes
Splines	23
W-H, h=1, z=2	21
W-H, h=100, z=2	6
W-H, h=1000, z=2	4
W-H, h=1, z=3	23
W-H, h=100, z=3	19
W-H, h=1000, z=3	18
W-H, h=1, z=4	24
W-H, h=100, z=4	20
W-H, h=1000, z=4	20
Moyenne mobile	22
Moyenne mobile écrêtée	22
Noyau gaussien h=2	22
Noyau gaussien h=3	22
Noyau gaussien h=4	22
Noyau d'Epanechnikov h=2	28
Noyau d'Epanechnikov h=3	24
Noyau d'Epanechnikov h=4	22
Noyau cubique h=2	28
Noyau cubique h=3	26
Noyau cubique h=4	22

FIGURE 74 – Test à distance finie population femme

Types de lissages	Nombre de changements de signes
Splines	9
W-H, h=1, z=2	15
W-H, h=100, z=2	2
W-H, h=1000, z=2	2
W-H, h=1, z=3	18
W-H, h=100, z=3	15
W-H, h=1000, z=3	6
W-H, h=1, z=4	20
W-H, h=100, z=4	18
W-H, h=1000, z=4	13
Moyenne mobile	18
Moyenne mobile écrêtée	18
Noyau gaussien h=2	18
Noyau gaussien h=3	18
Noyau gaussien h=4	18
Noyau d'Epanechnikov h=2	28
Noyau d'Epanechnikov h=3	18
Noyau d'Epanechnikov h=4	18
Noyau cubique h=2	28
Noyau cubique h=3	22
Noyau cubique h=4	20

FIGURE 75 – Test à distance finie population homme

Types de lissages	Nombre de changements de signes
W-H, h=1, z=2	23
W-H, h=100, z=2	4
W-H, h=1000, z=2	2
W-H, h=1, z=3	26
W-H, h=100, z=3	19
W-H, h=1000, z=3	14
W-H, h=1, z=4	25
W-H, h=100, z=4	20
W-H, h=1000, z=4	21

FIGURE 76 – Test à distance finie population cadre

Types de lissages	Nombre de changements de signes
W-H, h=1, z=2	17
W-H, h=100, z=2	4
W-H, h=1000, z=2	2
W-H, h=1, z=3	20
W-H, h=100, z=3	13
W-H, h=1000, z=3	6
W-H, h=1, z=4	22
W-H, h=100, z=4	18
W-H, h=1000, z=4	18

FIGURE 77 – Test à distance finie population non cadre

Types de lissages	Nombre de changements de signes
Splines	1,000809157
W-H, h=1, z=2	1
W-H, h=100, z=2	1
W-H, h=1000, z=2	1
W-H, h=1, z=3	1
W-H, h=100, z=3	1
W-H, h=1000, z=3	1
W-H, h=1, z=4	1
W-H, h=100, z=4	1
W-H, h=1000, z=4	1
Moyenne mobile	1,195619174
Moyenne mobile écrêtée	1,206801279
Noyau gaussien h=2	1,20305687
Noyau gaussien h=3	1,199159731
Noyau gaussien h=4	1,197648941
Noyau d'Epanechnikov h=2	1,216661085
Noyau d'Epanechnikov h=3	1,209017669
Noyau d'Epanechnikov h=4	1,201165623
Noyau cubique h=2	1,218283419
Noyau cubique h=3	1,213608644
Noyau cubique h=4	1,208390511

FIGURE 78 – Test SMR population femme

Types de lissages	Nombre de changements de signes
Splines	1,00440634
W-H, h=1, z=2	1
W-H, h=100, z=2	1
W-H, h=1000, z=2	1
W-H, h=1, z=3	1
W-H, h=100, z=3	1
W-H, h=1000, z=3	1
W-H, h=1, z=4	1
W-H, h=100, z=4	1
W-H, h=1000, z=4	1
Moyenne mobile	1,255748583
Moyenne mobile écrêtée	1,273118894
Noyau gaussien h=2	1,264987598
Noyau gaussien h=3	1,260190638
Noyau gaussien h=4	1,258303892
Noyau d'Epanechnikov h=2	1,280459123
Noyau d'Epanechnikov h=3	1,272727732
Noyau d'Epanechnikov h=4	1,262767865
Noyau cubique h=2	1,281774751
Noyau cubique h=3	1,277312281
Noyau cubique h=4	1,271517729

FIGURE 79 – Test SMR population homme