TABLE DES MATIERES

RESUME EN FRANÇAIS AVEC MOTS CLES

Le marché de la prévoyance collective est ultra concurrentiel. Ses acteurs lancent des initiatives pour trouver de nouvelles sources de rentabilité. Par exemple, divers projets de transformation sont lancés pour réduire les coûts, optimiser les processus de gestion en intégrant de nouveaux flux de gestion (ex : la Déclaration Sociale Nominative). Sur la gestion de l'absentéisme, certains acteurs s'interrogent sur la compétitivité de leurs garanties de prévoyance obligatoire et la pertinence de leurs outils de tarification. C'est justement le cas de l'organisme assureur partenaire.

Cette recherche de compétitivité lui impose d'étudier l'opportunité d'actualiser sa grille tarifaire. Pour cela, il doit améliorer la connaissance de son portefeuille d'assurés et comprendre sa sinistralité.

Ce mémoire étudie l'opportunité d'actualiser la grille de tarif et d'orienter la prise de décision sur la pertinence d'introduire la notion de « classe de risque » pour segmenter les entreprises prospectées ou en portefeuille.

Nous démarrerons par une analyse descriptive de la population sous risque et du risque « incapacité ».

Ensuite nous utiliserons une méthode pour actualiser les lois de maintien et d'incidence en incapacité avec les données d'expérience (issues de la DSN et du système de gestion des sinistres).

Enfin, il sera l'occasion d'étudier l'opportunité d'affiner la loi d'incidence par recours à des méthodes de machine learning. Il sera également intéressant de s'interroger sur la pertinence de recourir à de l'open data pour segmenter.

Mots clés : incapacité, loi d'incidence, loi de maintien, Kaplan-Meier, Whittaker-Henderson, machine learning, tarification, prévoyance collective, Clustering, K-means, open data.

ABSTRACT

The group insurance market is highly competitive. These players are launching initiatives to find new sources of profitability. For example, various transformation projects are launched to reduce costs and optimize management processes by integrating new management flows (eg the DSN). On the management of absenteeism, some players wonder about the competitiveness of their guarantees. This is precisely the case with the partner insurer.

This search for competitiveness requires studying the advisability of updating the price list. To do this, the partner insurer must know his portfolio of policyholders and understand his claims experience.

This thesis studies the advisability of updating the price list and orienting decision-making on the relevance of introducing the notion of "risk class" to segment the subscribing companies.

The study will start with a descriptive analysis of the population at risk and the risk of "disability".

Then the study will propose a method to update the probability distribution of disability maintenance and incidence with the experience data (from the DSN and the claims management system).

Finally, it will be the opportunity to refine the probability distribution of incidence by using machine learning methods. It will also be interesting to question the relevance of using open data to segment.

Keywords: disability, probability distribution of incidence, probability distribution of maintenance, Kaplan-Meier, Whittaker-Henderson, machine learning, pricing, collective provident insurance, Clustering, Kmeans, open data.

REMERCIEMENTS

En tant que consultant en transformation dédié au secteur de l'assurance, j'ai souhaité compléter mon cursus en suivant la formation du CEA. Ce mémoire me permet de clôturer un long cycle de plusieurs années d'une densité professionnelle et personnelle assez élevée.

Tout d'abord, je remercie profondément les associés d'aVB Ophélie Viard, Virginie Gubler et Alexandre Gros sans qui ce projet d'envergure n'aurait tout simplement pas vu le jour. Cela me permet d'atteindre une nouvelle dimension dans mes aspirations personnelles et professionnelles.

Je remercie également les tuteurs de la mutuelle partenaire Nicolas Cadiou et Issam Tazrouti qui ont rendu possible la réalisation du mémoire de par la fourniture de données anonymisées et l'apport de leur expertise métier en matière de data particulièrement. Je tiens à remercier également les personnes de leur équipe Asmaa Benmalek et Samir Lazzali.

Je tiens à remercier mes collègues de promotion du CEA pour l'entraide durant les deux années académiques (en particulier notre petit groupe surnommé le « comex cea »).

Je tiens également à remercier mes amis des Mines de Saint-Etienne – sollicités en période de révision pour leur expertise respective : Oussama El Jani (finance), Anis Bensenane (ALM / actuariat), Pierre Attard, Yazan Markawabi et Kévin Pérez, PhD (data science).

Je remercie également Olivier Lopez pour sa disponibilité et ses précieux conseils. Je remercie également Sabahe Touat pour ses chaleureux encouragements.

Je tiens à remercier ma famille (ainsi que leurs prières) qui m'a soutenu durant ce projet et tout particulièrement ma femme Carine pour son extrême patience, sa compréhension, sa tolérance, son soutien inconditionnel dans mes choix et mes envies.

AVERTISSEMENTS

Pour des raisons de confidentialité, les résultats chiffrés ont été modifiés sans perte de généralité et une partie des échelles supprimée.

LEXIQUE

APET : **A**ctivité **P**rincipale de l'**Et**ablissement

Argmin : **arg**ument **min**imum d'une fonction. Il représente l'ensemble des points en lesquels une expression atteint sa valeur minimale

BCAC : **B**ureau **C**ommun d'**A**ssurances des **C**ollectives

CSP : **C**atégorie **s**ocio **p**rofessionnelle

CTIP : **C**entre **T**echnique des **I**nstitutions de **P**révoyance

DSN : **D**éclaration **s**ociale **n**ominative

FFA : **F**édération **F**rançaise de l'**A**ssurance

FNMF : **F**édération **N**ationale de la **M**utualité **F**rançaise

IJ : **I**ndemnités **j**ournalières

OA : **O**rganisme **A**ssureur

SMIC : **S**alaire **MI**nimum de **C**roissance

INTRODUCTION

En cas d'arrêt maladie, les assurés perçoivent généralement des indemnités de la Sécurité Sociale. Ces indemnités sont complétées par des organismes de prévoyance regroupés en trois fédérations (FNMF, CTIP et FFA).

D'après les chiffres de ces fédérations, le marché de la prévoyance complémentaire est en constante progression. Que ce soit le marché du collectif ou de l'individuel, cette tendance se confirme par une hausse de l'absentéisme et donc une hausse du besoin en protection sociale.

La forte croissance de la demande et la multiplicité des acteurs en font un marché très concurrentiel.

Nous avons eu l'opportunité de collaborer avec un organisme assureur qui souhaite accroître sa compétitivité et sa rentabilité sur le marché de la prévoyance collective obligatoire. Pour être compétitif et rentable, l'un des leviers est d'agir sur le tarif des garanties Incapacité de son portefeuille. Mais, agir sur le prix revient à approfondir la connaissance du portefeuille des assurés.

Le mémoire propose de s'intéresser à un portefeuille sur lequel la connaissance de la sinistralité est limitée.

Certaines typologies d'arrêts de travail ne sont pas prises en compte dans les modèles : les arrêts courts (antérieurs à la période de franchise) ne sont pas exploités.

Toutes les sources de données accessibles ne sont pas encore exploitées dans les modèles existants. Il s'agit par exemple : des données relatives au domaine « santé », des flux de gestion Prest'IJ, DSN (depuis 2017, la DSN permet l'observation de tous les arrêts qu'ils soient avant la période de franchise ou après la période d'indemnisation maximale de la mutuelle), des open data.

Pour la tarification, les tables réglementaires de maintien en incapacité et de sortie (décès, retraite, invalidité) sont utilisées. Ces tables sont pratiques et rapidement disponibles mais ne décrivent pas précisément la sinistralité du portefeuille. Ces tables sont construites sur la base de données issues d'un panel d'organismes assureurs. Pour des raisons de simplicité, seul le paramètre de l'âge est pris en compte (Kamega A., Planchet F., Wolfrum R. 2013). Par ailleurs et spécifiquement sur l'invalidité, il faudrait quarante ans d'historique pour disposer d'un recul suffisant.

Pour le portefeuille étudié, un outil de marché est utilisé pour déterminer un prix de référence. Il utilise des variables classiques pour la tarification de ses garanties incapacité (âge, répartition par genre et catégorie professionnelle).

Cet outil ne permet pas de rendre compte de la diversité des profils de risque des souscripteurs et toute nouvelle évolution est couteuse et complexe à faire évoluer.

Fort de ce constat, il convient de s'interroger sur les deux points suivants :

- L'opportunité de profiter des données d'expérience pour actualiser la tarification actuelle,
- L'opportunité de construire une grille de coefficient de minoration/majoration annexe à l'outil. Elle permettrait d'ajuster le prix en fonction du profil de risque de l'entreprise avec des variables de type démographique ou en lien avec le secteur d'activité.

Le mémoire est organisé de la manière suivante :

1) Mieux décrire le portefeuille et analyser les poches de sinistralité,
2) Reconstituer les lois d'incidence et de maintien en incapacité à partir des données d'expérience,
3) Modifier le modèle de tarification et retrouver un coût du risque plus juste. En effet, aucun levier sur la part des chargements n'est possible (le marché indique une part à environ 10% / 12% de la cotisation totale),
4) Etablir des classes de risques pour ajuster la tarification,
5) Constituer la grille de coefficient de minoration / majoration du tarif socle.

1. CONTEXTE DE L'ETUDE, PRESENTATION DES DONNEES ET DU SUJET

1.1. DEFINITION D'UN REGIME DE PREVOYANCE COLLECTIVE

Le Centre Technique des Institutions de Prévoyance (C.T.I.P) définit la prévoyance collective comme un système qui intervient en complément des prestations des régimes obligatoires de Sécurité sociale. La prévoyance collective est un système qui protège les assurés et leur famille des risques lourds (incapacité, invalidité, décès) et leurs ayants-droits de se couvrir contre les risques liés à la personne.

Ces risques sont ceux relatifs aux dommages corporels suite à une maladie ou un accident (hospitalisations, arrêts de travail, invalidité, décès).

Lorsque le risque se réalise, les prestations d'un régime de prévoyance permettent la compensation d'une partie des pertes de revenus en cas d'arrêt de travail, de décès (capital décès, rentes de conjoint et d'éducation, épargne retraite, dépendance).

Un régime de prévoyance collective s'instaure dans une entreprise ou au niveau d'une branche professionnelle. Plusieurs modalités existent pour sa mise en place.

1.2. LA MISE EN ŒUVRE D'UNE PREVOYANCE COLLECTIVE

1.2.1) Les cinq modalités de mise en place

Cette mise en place s'effectue par un échange entre les représentants des salariés et les représentants des employeurs.

Cette mise en place peut s'effectuer de cinq manières différentes :

1) Par application des textes d'accords collectifs ou de conventions collectives nationales (CCN)

Les dispositions décrites dans les textes fixent le cadre minimal du contrat de prévoyance (le contrat pouvant cependant offrir davantage de garanties).

2) Par accord d'entreprise

Cet accord résulte d'une négociation entre employeur et salariés. Le personnel doit obligatoirement être informé.

3) Par référendum

La mise en place du régime est démocratique (ratification à la majorité). Le personnel doit obligatoirement être informé.

Pour les modalités 1, 2 et 3, l'affiliation est obligatoire pour le salarié.

4) Par décision unilatérale de l'employeur

Pour tous les salariés, le régime fiscal qui s'applique est celui de l'affiliation obligatoire. Le personnel doit obligatoirement être informé par un écrit. L'affiliation est obligatoire uniquement pour les nouveaux arrivants.

5) Par consentement individuel

Dans ce cas, le régime de prévoyance collective est facultatif. L'affiliation du salarié est facultative. De ce fait, l'employeur paie moins de charges patronales pour les salariés non assurés.

Après avoir vu les cinq modalités de mise en place, il convient de présenter en synthèse le cadre juridique.

1.2.2) Une mise en place encadrée juridiquement

La mise en place d'un régime de prévoyance implique le respect de plusieurs principes :

#	Principes	Descriptions
1	Devoir d'information	- Obligation d'information préalable des salariés avant l'adhésion, - Obligation chaque année d'envoi à l'entreprise d'un rapport sur les comptes du contrat.
2	Non sélection médicale	- Aucune différenciation d'âge entre les salariés ne doit être effectuée sur les tarifs, - Les questionnaires médicaux ne sont autorisés que dans quelques cas.
3	Participation uniforme	Contribution employeur fixée à un taux ou un montant uniforme : pour l'ensemble des salariés ou pour tous ceux appartenant à la même catégorie.
4	Caractère « obligatoire »	Lorsqu'un contrat collectif à adhésion obligatoire est mis en place au sein de l'entreprise, l'ensemble des salariés y est affilié sauf facultés de dispense d'adhésion au choix du salarié.
5	Changement d'assureur (résiliation)	- Pour les assurés en cours d'indemnisation, maintien des rentes en cours à un niveau au moins égal à celui de la dernière prestation (Loi Evin art.7) - Pour les assurés en Incapacité de travail ou Invalidité (Loi Evin art.7.1), maintien de la garantie décès en cas de rente.

1.3. LES GRANDES DATES DE LA REGLEMENTATION PREVOYANCE

Dates	Réglementation	Précisions
1947	CCN	Mise en place d'une obligation de garantie Décès pour les cadres à la charge seule de l'employeur
1978	Loi de mensualisation	Maintien de salaire partiel obligatoire en cas d'arrêt de travail pour maladie, accident de la vie courante ou de la vie professionnelle, par l'employeur en complément des IJ SS, à partir d'une année d'ancienneté, financée : - Via sa propre trésorerie, - Via des cotisations versées à un organisme assureur.

Dates	Réglementation	Précisions
1989	Loi Evin	Obligation de maintien des prestations en cas de changement d'assureur
2007	Contrat en déshérence	Obligation de recherche active de bénéficiaires
2012	Catégories objectives	Durcissement des critères de définition des catégories de personnel (définition des catégories de salariés avec des critères objectifs (circulaire DSS du 25/9/2013), égalité de traitement de tous les salariés d'une même catégorie de personnels (jurisprudence de janvier 2015))
2014	Contrat en déshérence – Loi Eckert	Renforcement du dispositif de contrôle autour des contrats en déshérence

Par ailleurs, les réformes des retraites impactent directement les organismes assureurs. Ils sont tenus de couvrir les assurés en arrêt de travail jusqu'à leur départ en retraite.

Il convient désormais de s'intéresser aux garanties que recouvre la prévoyance collective.

1.4. LES GARANTIES DE LA PREVOYANCE COLLECTIVE

Les trois risques principaux couverts en prévoyance sont :

- L'incapacité de travail : cette garantie permet au salarié en arrêt de travail de percevoir des indemnités journalières complémentaires à celles de la Sécurité sociale et du revenu versé par l'employeur.
- L'invalidité : cette garantie permet le versement d'une rente d'invalidité. Cette rente se substitue partiellement voire en totalité à la perte de revenu du salarié en invalidité. Elle est versée en complément de la pension d'invalidité de la Sécurité sociale,
- Le décès : ces garanties décès prévoient le versement de rente ou capital aux bénéficiaires de l'assuré décédé (rente conjoint, rente éducation…).

En complément des garanties de prévoyance, les contrats de prévoyance collective peuvent être assortis :

- De garanties complémentaire santé : il s'agit de remboursements complémentaires à la sécurité sociale pour des soins de ville, des frais d'hospitalisation, dentaires et optiques. Ces remboursements sont en faveur de l'assuré et/ou de ses bénéficiaires.
- De produit d'épargne retraite collective : il s'agit de dispositif de retraite supplémentaire en complément des pensions versées par les régimes obligatoires de retraite. Ces régimes sont dits « par capitalisation » en opposition à « par répartition ».
- De garantie liée à la dépendance : non encore identifié par la Sécurité Sociale comme une famille de risque, les organismes assureurs proposent des garanties qui prévoient le versement d'une rente viagère en cas d'assuré en dépendance.

En réalité, les garanties de l'arrêt de travail sont multiples.

1.5. LES GARANTIES EN CAS D'ARRET DE TRAVAIL

Les garanties complémentaires liées à l'arrêt de travail sont l'incapacité et l'invalidité.

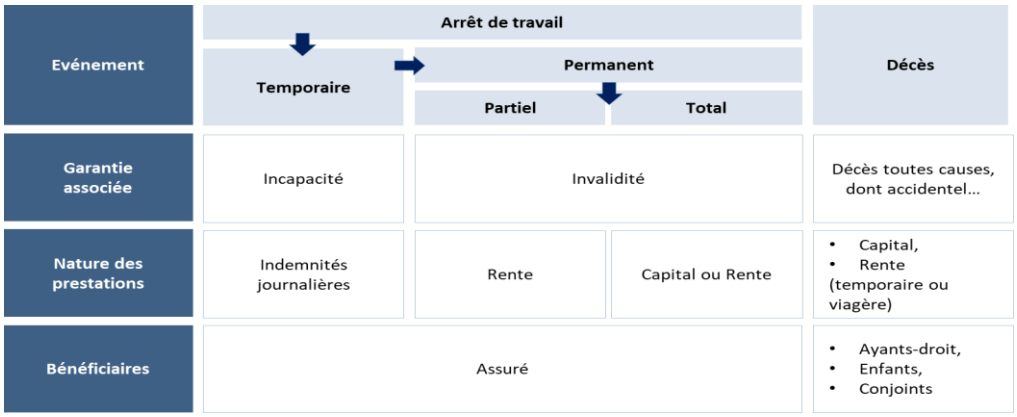

Figure 1. Les garanties de l'arrêt de travail

Après un accident du travail ou une maladie professionnelle indemnisé(e) (ATMP), plusieurs cas de figure sont possibles suivant l'âge et l'état de santé de l'assuré :

- La guérison : dans ce cas, le salarié est en mesure de reprendre une activité professionnelle. Elle est formalisée par un certificat médical final de guérison adressé à la CPAM lorsqu'aucune incapacité permanente n'est signalée,
- Le départ à la retraite : en tout état de cause à compter de la date d'effet de la pension servie par le régime général de la Sécurité sociale ou par un régime de retraite complémentaire (ex : AGIRC-ARRCO),
- La consolidation.
    - Elle est observée lorsque l'absence de guérison est constatée par le médecin. La lésion a pris un caractère permanent sinon définitif. Elle est formalisée par un certificat de consolidation adressé à la CPAM.
    - Le médecin établit la date de consolidation et fixe le taux d'Incapacité Permanente Partielle (taux d'IPP). Si les séquelles persistent, un reclassement professionnel est à envisager. Dans ce cas, l'assuré passe en incapacité permanente acté au niveau de l'Assurance Maladie.

Lorsqu'une personne est victime d'une maladie ou d'un accident d'origine non professionnelle, qui réduit d'au moins 2/3 sa capacité de travail ou de gain, elle peut être reconnue invalide. Il existe trois catégories :

- Catégorie 1 - Personne capable d'exercer une activité rémunérée mais dont la capacité de gain est réduite de plus des deux tiers,
- Catégorie 2 - Incapacité absolue d'exercer une profession quelconque,
- Catégorie 3 - Personne absolument incapable d'exercer une profession quelconque, et nécessitant l'assistance d'une autre personne pour effectuer les actes ordinaires de la vie.

Le calcul de la pension est fonction de la base du salaire annuel moyen perçu pendant les dix meilleures années d'activité dans la limite des planchers et plafonds (limités à la T1, décrits et fixés par la Sécurité Sociale chaque année).

Les garanties complémentaires liées à l'arrêt de travail sont vendues par plusieurs typologies d'acteurs.

1.6. LE MARCHE DE L'ARRET DE TRAVAIL EN FRANCE

Les quatre types d'acteurs suivants sont habilités à proposer des couvertures de prévoyance :

1) Les institutions de prévoyance dépendent du code de la sécurité sociale. Elles sont pilotées par un conseil d'administration paritaire,
2) Les mutuelles sont régies par le code de la mutualité et pilotées par un conseil d'administration élu par les sociétaires,
3) Les sociétés d'assurance mutuelle diffèrent des mutuelles par le fait qu'elles dépendent du code des assurances,
4) Les compagnies d'assurance dépendent du code des assurances et sont pilotées par un conseil d'administration élu par les actionnaires.

Ces acteurs se partagent le marché :

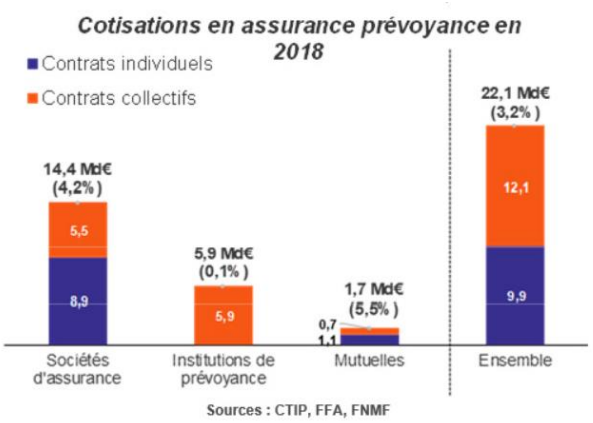

Figure 2. Cotisations en assurance prévoyance en 2018

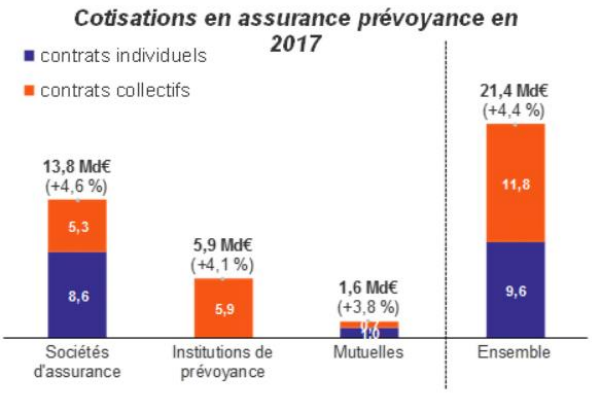

Figure 3. Cotisations en assurance prévoyance en 2017

Globalement, le marché de prévoyance progresse de 3,2% par rapport à 2017. Mais selon le CTIP, la FFA ou la FNMF, il n'est pas possible de distinguer avec précision « les facteurs de croissance » parmi :

- La hausse générale des cotisations,
- L'augmentation globale du portefeuille d'assurés.

Par rapport à 2017, la répartition du marché selon les acteurs reste stable pour les institutions de prévoyance et les mutuelles. La croissance du marché profite aux sociétés d'assurance.

En prévoyance, la part de marché du collectif est stable et représente plus de 55%. Mais cette proportion diffère selon les acteurs :

- Les institutions de prévoyance sont historiquement adossées à des caisses de retraite complémentaire et bénéficiaient de clauses de désignations. De ce fait, le marché du collectif représente la quasi intégralité de leur portefeuille. Depuis la fin des désignations en 2013, la part de marché des institutions de prévoyance reste finalement stable,
- Les mutuelles et les sociétés d'assurances se partagent un peu moins de la moitié du marché du collectif mais détiennent la quasi-totalité du marché de l'individuel.

D'un côté, l'organisme assureur évolue donc dans un contexte fortement concurrentiel avec un objectif de gagner des parts de marché.

De l'autre, il a pour objectif de compléter au plus juste les prestations versées par la Sécurité sociale.

1.7. ARTICULATION AVEC LA GARANTIE DE LA SECURITE SOCIALE

1.7.1) Processus général de versement des indemnités journalières de base et complémentaires

Des conditions administratives et médicales doivent être respectées pour l'ouverture de droits consécutive à un arrêt de travail (AT) :

- Constatation médicale de l'incapacité à continuer ou à reprendre le travail par les prescripteurs autorisés,
- Un plafond minimum d'heures travaillées en fonction de la durée de l'arrêt de travail.

En cas de maladie nécessitant un Arrêt de Travail : la Sécurité Sociale verse au salarié des Indemnités Journalières, à compter du 4e jour d'arrêt de travail.

Au niveau complémentaire, la garantie « incapacité de travail » prévoit le versement au salarié d'Indemnités Journalières Complémentaires (IJC), complémentaires à celles de la Sécurité sociale.

En cas d'accident du travail, de maladie professionnelle, ou de congé maternité ou paternité, le versement de ces indemnités journalières suit un régime spécifique, avec un délai de carence et un plafond de versement différents.

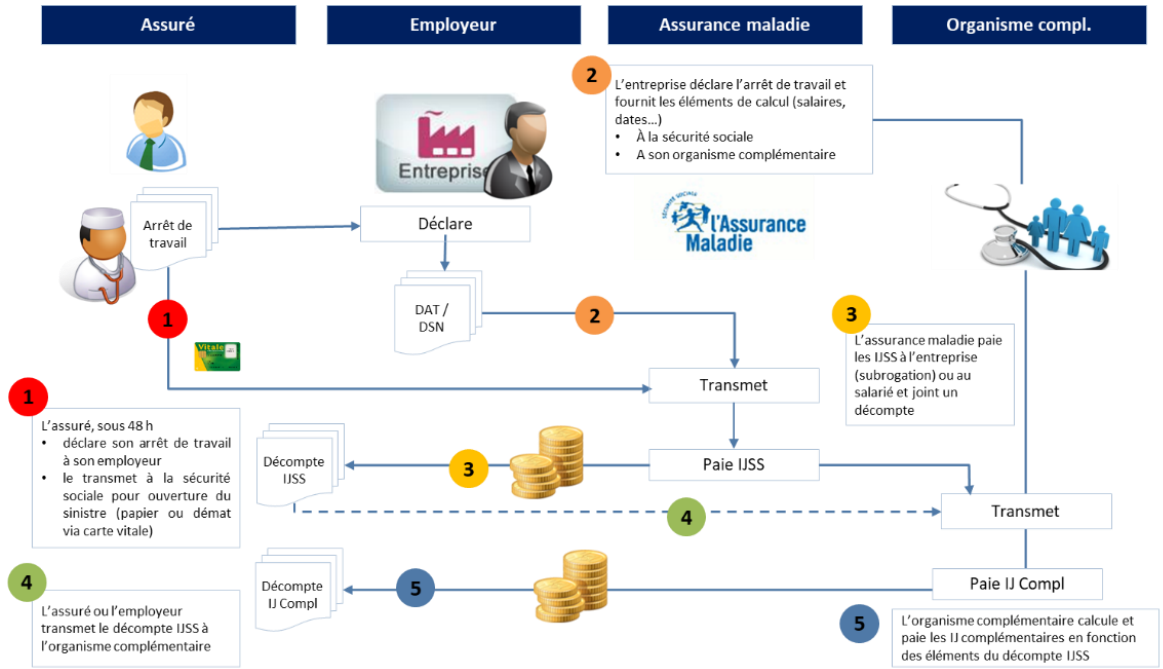

Figure 4. Fonctionnement des Indemnités journalières

1.7.2) Principes de versement des indemnités journalières

a) Premier principe

Les indemnités journalières versées par la SS (IJSS) sont égales à 50% du gain journalier de base (plafonné à 1,8 SMIC) après une carence de 3 jours.

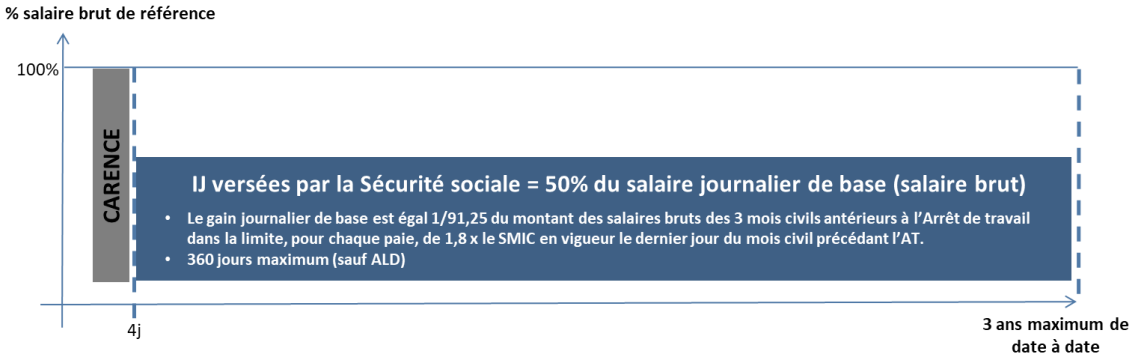

Figure 5. Versement des IJ - 1er principe

b) Deuxième principe

A partir du 8ème jour d'arrêt de travail, l'obligation de maintien de salaire prévue par la CCN ou la loi de mensualisation pèse sur l'employeur.

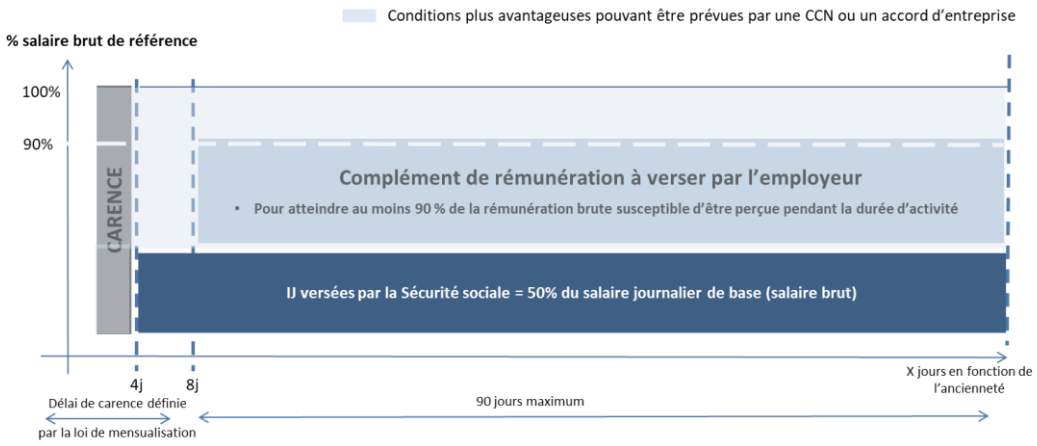

Figure 6. Versement des IJ - 2ème principe

La convention collective ou un accord d'entreprise peut prévoir :

- Une réduction / suppression du délai de carence, i.e. que l'entreprise complète les IJ avant le 8e jour d'AT (dès le 1er jour par exemple),
- Un niveau d'indemnisation plus élevé (par exemple, 100 %, puis 80 %),
- Des périodes de versement du complément de rémunération plus longues que les périodes prévues par la loi de mensualisation.

c) *Troisième principe*

Cette obligation baisse à une date en fonction de l'ancienneté du salarié.

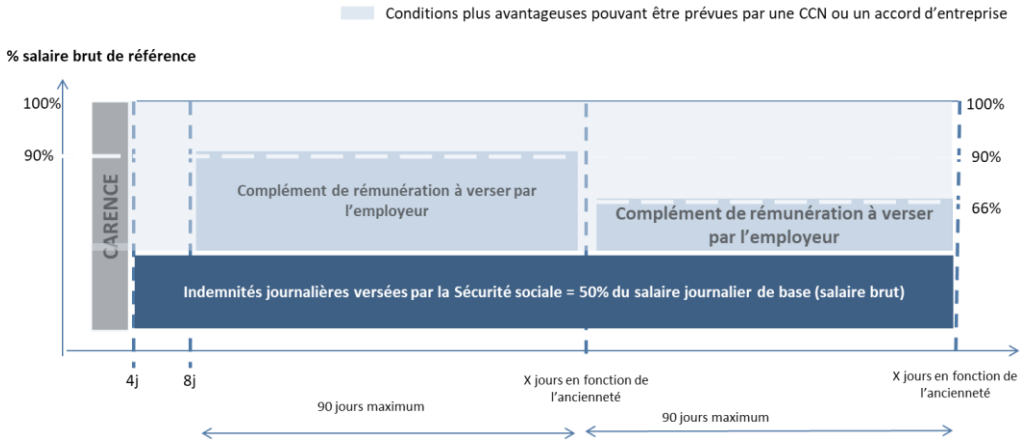

Figure 7. Versement des IJ - 3ème principe

La loi n'impose pas aux employeurs la souscription d'une assurance. Le risque pour l'employeur étant de financer l'incapacité lui-même.

Ces trois principes mettent en exergue l'implication d'au moins quatre acteurs « employeur-SS-assureur-assuré » dans le processus de versement d'indemnités journalières. Cette relation nécessite d'échanger des flux d'informations sous un format idéalement compréhensible par les organismes de la sphère sociale : la DSN.

1.8. LA DSN (DECLARATION SOCIALE NOMINATIVE) - VEHICULE DES INFORMATIONS

1.8.1) L'un des grands chantiers de transformation de la sphère sociale

A compter de la loi n° 2012-387 du 22 mars 2012 relative à la simplification du droit et à l'allégement des démarches administratives, les entreprises de la sphère sociale se sont mises en ordre de marche pour instaurer le dispositif DSN. Il s'agit d'un projet de transformation majeur de la sphère sociale.

Depuis 2017, ce dispositif vise à simplifier les démarches administratives des entreprises et à automatiser le processus de collecte d'informations pour les consommateurs de ces données.

A partir du logiciel de paie de l'entreprise, la DSN permet en une seule transmission mensuelle dématérialisée de remplacer la plupart des déclarations sociales pour les entreprises (ex : DADSU).

Les données de la DSN sont transmises aux organismes collecteurs de cotisations sociales :

- Les organismes de la sphère ACOSS ou MSA (Pôle Emploi, CNAV, CNAM…),
- Les organismes des régimes complémentaires (Agirc-Arrco, Délégataires de gestion, Institutions de prévoyance, Mutuelles, etc.).

L'instauration de la DSN profite spécifiquement au secteur de la prévoyance. En effet, les informations sont disponibles :

- Tête par tête : les rémunérations brutes versées aux salariés, sur lesquelles sont calculées les cotisations sociales, ainsi que les droits des salariés (retraite, assurance maladie...), les caractéristiques de l'emploi occupé et du contrat de travail, période d'emploi, etc.).
- A une fréquence mensuelle : ce qui constitue du « temps réel » en comparaison avec la DADSU annuelle,
- Chaque évènement dans la carrière d'un salarié est connu : dès le premier jour d'arrêt de travail, l'ensemble des engagements dus sera connu en temps réel.

Spécifiquement sur le volet tarification / provisionnement des garanties incapacité, les actuaires exploitent les tables de référence du BCAC. Ils appliquent ensuite des coefficients de minoration / majoration en fonction du profil de la population sous risque.

Depuis 2017, les DSN fournissent des données d'expérience. Cela offre donc de nouvelles perspectives dans les modèles. Toutefois, comme tout démarrage de dispositif, les données sont à fiabiliser.

1.8.2) Structure de la DSN

Les données de la DSN sont hiérarchisées et suivent une logique en structure, groupe, bloc, rubrique.

La Norme d'Echanges Optimisée des Données Sociales (NEODeS) formalise les règles de constitution d'une DSN. Cette norme est mise à jour annuellement.

La DSN est composée des cinq structures suivantes :

1) « En-tête »

Cette structure contient principalement les blocs de données relatives à l'identification de l'émetteur de la DSN, du logiciel utilisé, de la norme DSN utilisée, du point de dépôt.

2) « Déclaration »

Cette structure contient les blocs relatifs à l'identification du déclarant, du destinataire et de la déclaration elle-même (mois de paie, numéro d'ordre, de fraction…). En particulier, cette rubrique permet de signaler les évènements (fin de contrat, arrêt de travail, reprise suite à un arrêt de travail…).

3) « Paie et RH »

Cette structure contient le plus grand nombre de blocs. Les blocs permettent d'identifier l'entreprise, l'établissement et spécifiquement de déclarer les adhésions à des organismes complémentaires. Le bloc « Adhésion Prévoyance » est dédié à cela. La structure contient également les blocs de données relatifs aux individus, au contrat de travail, à la rémunération, au temps d'activité, aux régularisations, aux ayants droits, au paiement des cotisations, aux éléments de versement, aux cotisations individuelles.

4) « Véhicule technique ». Cette structure regroupe des données à finalité fiscale.
5) « Totaux ». Cette structure dénombre le nombre de rubriques et de DSN.

1.8.3) Exemple : les données de la DSN pertinentes de l'arrêt de travail

Pour étudier l'arrêt de travail, les données relatives à l'entreprise, l'adhésion, l'individu, le contrat de travail et à l'arrêt sont hiérarchisées en 4 niveaux de la manière suivante dans une DSN :

Niveau 1 Structure	Niveau 2 Groupe	Niveau 3 Bloc	Niveau 4 Rubrique
S20 Déclaration			
	S20.G00		
		S20.G00.07 - Contact chez le déclaré (1,2)	
S21 Paie et RH			
	S21.G00		
		S21.G00.06 - Entreprise (1,1)	
		S21.G00.15 - Adhésion Prévoyance (0,*)	
		S21.G00.30 - Individu (1,1)	
		S21.G00.40 - Contrat (contrat de travail, convention, …) (1,1)	
		S21.G00.60 - Arrêt de travail (1,1)	Rubriques cf. ci-dessous

Le bloc arrêt de travail (S21.G00.60) contient plusieurs rubriques :

Libellé de la rubrique	Référence de la rubrique
Motif de l'arrêt	*S21.G00.60.001*
Date du dernier jour travaillé	*S21.G00.60.002*
Date de fin prévisionnelle	*S21.G00.60.003*
Subrogation	*S21.G00.60.004*
Date de début de subrogation	*S21.G00.60.005*
Date de fin de subrogation	*S21.G00.60.006*
IBAN	*S21.G00.60.007*
BIC	*S21.G00.60.008*
Date de la reprise	*S21.G00.60.010*
Motif de la reprise	*S21.G00.60.011*
Date de l'accident ou de la première constatation	*S21.G00.60.012*
SIRET Centralisateur	*S21.G00.60.600*

Après avoir explicité les caractéristiques de l'arrêt de travail, il convient de s'intéresser aux données du portefeuille étudié.

2. LES DONNEES

2.1. LE PERIMETRE

Le portefeuille étudié est constitué de contrats collectifs de prévoyance spécifiquement sur les garanties incapacité.

Niveau de granularité	Inclusion	Exclusion
Individus (sinistres et affiliés)	- Les individus étudiés sont âgés de 18 à 60 ans, - Cadres / Agent de maitrise / Employés / Ouvriers, - Individus en activité au 1er janvier 2017, - Les salariés ayant été actifs sur la période d'observation.	- Les agriculteurs et agents de la fonction publique (pour cause d'intégrité des données), - Les individus avec genre non renseigné (pour cause de représentativité <0.01% du périmètre), - Tout individu dont le sinistre est survenu le même jour qu'un sinistre tronqué mais n'ayant pas dépassé le début de la période d'observation.
Entreprises	Entreprises adhérentes à un contrat collectif de prévoyance	Les entreprises de 50 salariés sans aucun arrêt de travail observé sur 3 ans sont exclues du périmètre (la volumétrie n'est pas disponible. En effet, dans ces cas, les équipes de gestion n'enregistraient pas les affiliations prévoyance).
Garanties	- Indemnités Journalières (IJ1) - IJ courte durée (mensu - compl. mensu) (IJ2) - IJ courte pour les – 200H (IJ4) - Rachat de franchise (IJF) - Rente Incapacité Permanente (RI1) - Rente d'Incapacité professionnelle (RI2)	- Garantie maternité, - Garantie paternité, - Accident de service, - Cotisation additionnelle maladie, - IJ pour encaissement de cotisation, - IJ hospitalisation.

2.2. LES DEUX SOURCES DE DONNEES

Deux bases de données sont fournies :

- Une base de données issue de la DSN pour construire la loi d'incidence en incapacité,
- Une base de données issue du système de gestion des sinistres pour reconstruire la loi de maintien. Depuis 2017, le système de gestion est également alimenté des données DSN. Cela constitue un apport pour la loi de maintien notamment sur les franchises courtes. En effet :
    - Le système enregistre l'information de l'arrêt de travail même s'il n'est pas encore indemnisé par l'OA,
    - Seuls les sinistres indemnisés étaient enregistrés.

#	Source	Périodes d'observation
1	Base de stockage des DSN	Concaténation de toutes les DSN du mois de déclaration de janvier 2020 du portefeuille Des affiliations ouvertes depuis 1980 Des arrêts observés depuis janvier 2017
2	Base de données issue du système de gestion des sinistres	Du 01 Janvier 2010 jusqu'au 31 mars 2019

L'arrivée de la DSN constitue une opportunité pour éviter « l'affiliation au sinistre » - pratique de gestion qui permettait de limiter la charge en gestion : les personnes indemnisées sont affiliées au contrat au moment de la déclaration de sinistre. Les informations nominatives étaient transmises annuellement via les DADSU – supports qui comprenaient l'ensemble des mouvements de personnels. Depuis 2017, la DSN permet l'affiliation systématique modulo le retard de gestion et le traitement des rejets DSN.

Pour des raisons de confidentialité, les données extraites sont anonymisées et restituées sous format .csv pour permettre un travail à distance (dans le cas contraire, une table SAS aurait été privilégiée).

Les extractions sont réalisées en janvier 2020.

2.3. LES VARIABLES

2.3.1) Les variables de la base de données DSN

Variables	Libellé base de données	Précisions / valeurs possibles
Identifiant de l'individu	ID_INDIVIDU	-
Date de début de contrat	DebutContrat	-
Nature du contrat de travail	NatureContrat	-
Modalité du contrat de travail	Modalite	-
Quotité du contrat de travail	QuotiteContrat	-
Quotité de référence du contrat de travail	QuotiteReference	-
Date de début d'arrêt de travail	DernierJourTravaille	-
Date de fin prévisionnelle d'arrêt de travail	DateFinPrevisionnelle	-
Date de fin réelle d'arrêt de travail	RepriseDate	-
Identifiant de l'entreprise	Siren	-
Numéro d'établissement	NIC	-
Date de naissance de l'individu	DtNaissance	-
Genre de l'individu	Genre	-
Catégorie socioprofessionnelle	CSP	10 valeurs possibles (Cf. annexes)
Code activité principale de l'établissement	APET	-

2.3.2) Les variables de la base de données de gestion des sinistres

Variables	Libellé base de données	Précisions / valeurs possibles
Numéro de dossier	NU_DOSS	
Identifiant de l'individu sinistré	ID_PER_ASS	
Date de survenance	DT_SURV	
Date de début d'indemnisation	DT_DEB_PREST	
Date de fin d'indemnisation	DT_FIN_PREST	
Motif de l'arrêt	EVT_LIB_CAU	
Code garantie	C_GAR	
Ville	L_VILLE	
Code postal de résidence	C_POST	
Identifiant de l'entreprise	C_SIREN	
Situation familiale	C_SIT_FAM	C pour célibataire, M comme marié ou en couple
Libellé de la profession	L_PROF	
Code profession et catégorie socioprofessionnelle	C_CSP	
Date de naissance	DT_NAIS_PER	
Genre	C_SEXE	Deux valeurs possibles : homme ou femme
Effectif dans l'entreprise	NB_EFFECTIF	
Montant d'IJ	MT_PRES_BRUT	
Code banque	PER_NOM_BAN	Banque du bénéficiaire de la prestation
Nombre d'enfant	NB_ENF	Ce chiffre ne semble pas fiable (de nombreuses valeurs à 99)
Indicateur « avec ou sans enfant »	Valeur calculée	Indicateur construit à partir de la variable NB_ENF
Age à la survenance	Valeur calculée	Variable construite par différence entre la date d'extraction et la date de naissance
Durée de l'arrêt de travail	Valeur calculée	Variable construite par différence entre la date de fin d'indemnisation et la date de survenance
Année de survenance	Valeur calculée	
Mois de survenance	Valeur calculée	
Age à la survenance	Valeur calculée	Variable construite par différence entre la date de survenance et la date de naissance
Franchise	Valeur calculée	Différence entre les dates de début d'indemnisation et de survenance

Ces bases de données font l'objet de retraitements pour permettre de mener notre analyse.

2.4. LES PRINCIPAUX RETRAITEMENTS ET VOLUME DE DONNEES

La base de données « DSN » pour le calcul de la loi d'incidence ainsi que la base de données de gestion des sinistres font l'objet de retraitements. Pour cela, des hypothèses métiers sont nécessairement prises.

2.4.1) Les hypothèses métiers prises en compte

a) La prise en compte des rechutes

Lorsqu'un actif ayant déjà eu un arrêt reprend son travail et a un nouvel arrêt de travail en lien avec le premier, alors il s'agit d'une rechute.

Dans la base de données et pour un même individu, un cas de rechute est détecté de la manière suivante :

- 2 arrêts consécutifs,
- Chacun des 2 arrêts dure plus de 2 semaines,
- Le délai entre les 2 arrêts est de moins de 2 mois.

Compte-tenu de l'indisponibilité de la donnée dans le SI de gestion, il n'est pas possible de vérifier l'identité de la cause médicale entre ces 2 arrêts.

Figure 8. Condition de rattachement de deux arrêts successifs

b) Les motifs d'entrée et de sortie en incapacité

Grâce à l'exploitation de la DSN, les motifs d'arrêt de travail sont normalisés et permettent de mieux segmenter. Lesdits motifs issus de la DSN sont :

- 01 – maladie,
- 02 – maternité,
- 03 - paternité / accueil de l'enfant,
- 04 - congé suite à un accident de trajet,
- 05 - congé suite à maladie professionnelle,
- 06 - congé suite à accident de travail ou de service,
- 07 - femme enceinte dispensée de travail,
- 08 - temps partiel thérapeutique,
- 09 – adoption,
- 10 - [FP] Congé suite à une maladie imputable au service,
- 11 - [FP] Congé de maladie des victimes ou réformés de guerre (art 41),
- 12 - [FP] Congé de longue durée,
- 13 - [FP] Congé de longue maladie.

NB : tous les motifs ne donnent pas droit au versement de prestation par un organisme complémentaire.

Les motifs de sortie en incapacité sont : observation censurée, guérison, entrée en invalidité et décès.

c) Les assurés inclus dans le portefeuille

Tous les actifs de moins de 60 ans en activité professionnelle au 1er Janvier 2017 sont inclus. Au-delà de 60 ans, il n'est pas pertinent d'analyser :

- L'actif peut demander sa liquidation retraite sans adresser d'arrêt de travail,
- Le volume de données est faible. Ce qui ne garantit pas la pertinence des taux d'entrée en incapacité.

Pour ces actifs, les éventuels arrêts de travail antérieurement à cette date sont exclus car non disponible dans l'extraction.

Toutes les périodes d'incapacité au statut « annulé » sont retirées de la base de données.

2.4.2) Retraitements de la base de données « DSN »

Il est nécessaire de retraiter la base de données compte-tenu de deux sources majeures d'erreurs :

- Source 1 – Les erreurs de déclaration de l'entreprise,
- Source 2 - Des dysfonctionnements dans le processus d'intégration des données DSN dans le système d'information de l'entreprise.

Ces erreurs sont :

- la redondance d'information dans la base de données. Par exemple : des périodes d'arrêt incluses dans d'autres lignes du fichier, des doublons… Le risque est de surestimer le nombre d'arrêt pour un individu,
- des informations erronées à supprimer – exemple : pour un même individu, une date de fin de période d'arrêt antérieure à la date de début d'arrêt,
- des individus sans date de début de contrat. Cela ne permet pas de reconstituer l'exposition au risque,
- des individus avec un arrêt sans date de début d'arrêt. Cela ne permet pas de reconstituer la durée d'arrêt,
- des individus avec plusieurs dates de début de contrat. Cela peut biaiser le calcul de l'exposition,
- des individus avec plusieurs dates de naissance. Cela peut biaiser la segmentation des individus par âge à la survenance ou âge à la souscription.

Plusieurs fonctions et méthodes sont créées / utilisées pour retraiter les données de la base. Il s'agit d'obtenir une base avec :

- une ligne par individu pour les individus sans ou avec un seul arrêt,
- une ligne par arrêt pour les individus avec strictement plus d'un arrêt.

Les traitements sont réalisés avec le langage informatique Python.

a) Retraitements génériques des individus

Les retraitements effectués sont :

- Retenir une seule date de début de contrat par individu pour ceux ayant plusieurs dates. A défaut de disposer de la date d'enregistrement de l'observation, la date de début de contrat la plus récente est retenue. Malgré la sous-estimation induite de l'exposition au risque, cela permet d'être plus prudent dans le calcul de l'incidence en arrêt. La fonction $lister\_minimum$ est créée à cet effet. Illustration :

N° Ligne	ID	DébutContrat	DébutContrat fourni par la fonction
1	1	20/12/2018	20/12/2018
2	1	01/10/2019	20/12/2018

- Retenir une seule date de naissance par individu pour ceux ayant plusieurs dates. A défaut de disposer de la date d'enregistrement de l'observation, la date la plus récente est retenue. La fonction $lister\_minimum$ est également utilisée,
- Imputer la médiane des dates de début de contrat de la base aux individus sans date de contrat renseignée et dont un arrêt est déclaré (< 0,1% du portefeuille). Cela permet de ne pas sous-estimer le nombre d'arrêt et d'être prudent sur le taux d'incidence en arrêt,
- Retirer les individus sans date de naissance (0,5% du portefeuille),
- Retirer les individus avec un âge à la souscription inférieur à 18 ans et supérieur à 65 ans.

	Nombre de lignes		Nombre d'individus	
	En nombre	% par rapport au départ	En nombre	% par rapport au départ
Avant retraitement	1 091 558	100%	891 638	100%
Après retraitement	880 076	80,6%	880 076	98,7%
Après filtrage des âges aberrants à la souscription	857 701	78,6%	857 701	96,2%

b) Retraitements spécifiques aux individus avec au moins un arrêt

Pour les besoins de l'étude, la date de fin d'arrêt théorique est modélisée par la variable calculée « $FinModel$ ». Elle est déterminée de la manière suivante :

$$FinModel = \begin{cases} RepriseDate \text{ si renseignée} \\ DateFinPrevisionnelle \text{ sinon} \end{cases}$$

Une première fonction est créée pour distinguer les individus dont le nombre d'arrêt est :

- Soit égal à un,
- Soit strictement supérieur à un.

Pour les individus ayant un arrêt, deux traitements sont réalisés :

- Un pour retirer les valeurs manquantes ($FinModel\ ou\ DernierJourTravail$ vide),

- Une fonction $controler\_date\_fin\_avant\_debut$ pour identifier les lignes dont les dates de fin d'arrêt sont antérieures à la date de début.

Pour les individus ayant strictement plus d'un arrêt :

- Les deux fonctions précédentes sont également utilisées,
- D'autres fonctions sont créées :
    - Une fonction $identifier\_ligne\_en\_inclusion$ pour identifier les lignes dont la période d'arrêt est incluse dans une autre :

Illustration du fonctionnement la fonction :

N° Ligne	ID	DernierJourTravail	FinModel	Indicateur d'Inclusion fourni par la fonction	Commentaires
1	1	20/12/2018	01/02/2019	non_inclus	
2	1	20/12/2018	01/01/2019	inclus	(avec ligne 1)
3	1	01/04/2019	30/06/2019	inclus	(avec ligne 4)
4	1	15/03/2019	15/07/2019	non_inclus	

    - Une fonction $identifier\_lignes\_a\_rattacher$ pour identifier les périodes continues à rattacher.

N° Ligne	ID	DernierJourTravail	FinModel	Indicateur d'Inclusion fourni par la fonction	Commentaires
1	1	20/12/2018	01/02/2019	A_rattacher0	Le nombre en guise de suffixe joue le rôle de numéro d'ordre pour le rattachement des périodes
2	1	02/02/2019	01/03/2019	A_rattacher0	
3	1	01/04/2020	30/06/2020	A_rattacher1	
4	1	10/07/2020	15/08/2020	A_rattacher1	
5	1	01/09/2019	01/10/2019	RAS	Rien à signaler

    - Une fonction $identifier\_rechutes$ pour identifier les périodes d'arrêt consécutives séparées de 15 jours,
    - Une fonction $rattacher\_lignes$ pour modifier les lignes « à rattacher ». Illustration avec l'exemple précédent :

N° Ligne	ID	DernierJourTravail	FinModel	Indicateur d'Inclusion	Commentaires
1	1	20/12/2018	01/03/2019	A_rattacher0	Pour les lignes « A_rattacher », les valeurs des dates sont modifiées
2	1	**20/12/2018**	01/03/2019	A_rattacher0	
3	1	01/04/2020	15/08/2020	A_rattacher1	
4	1	**01/04/2020**	15/08/2020	A_rattacher1	
5	1	01/09/2019	01/10/2019	RAS	Rien à signaler

La méthode . $drop()$ est utilisée pour supprimer les lignes souhaitées - repérées par leur index.

Finalement, 92,1% des individus avec un arrêt sont conservés.

	Nombre de lignes	Nombre d'individus

	En nombre	% par rapport au départ	En nombre	% par rapport au départ
Avant retraitement	114 496	100%	99 230	100%
Après retraitement	104 060	90,9%	91 369	92,1%

2.4.3) Retraitement des données de gestion des sinistres

La base de données à disposition contient l'exhaustivité des sinistres indemnisés du portefeuille. Les principaux retraitements consistent à :

- conserver uniquement les sinistres relatifs aux garanties incapacité,
- contrôler les différentes dates (date de fin d'indemnisation nécessairement postérieure à la date de début).

Nombre de sinistres indemnisés	En nombre	% par rapport au départ
Avant retraitement	1 198 500	100%
Après retraitement	1 135 806	94,8%

Les bases de données sont retraitées et prêtes pour être analysées. Une partie significative des données est conservée. Il est désormais possible de réaliser l'étude descriptive du portefeuille et de sa sinistralité.

3. ANALYSE DESCRIPTIVE DU PORTEFEUILLE ET DE LA SINISTRALITE

3.1. LES CARACTERISTIQUES DES ASSURES SOUS RISQUES

Le jeu de données disponible est restreint à un mois de DSN. Pour des raisons liées à la sensibilité des données :

- L'organisme assureur a transmis une vision partielle des données.
- Les informations relatives aux garanties souscrites et aux franchises souscrites n'ont pas pu être transmises.

Le portefeuille contient 880 076 individus. La répartition selon leur genre est réalisée au prorata de l'exposition. L'exposition représente le nombre de jours de présence dans le portefeuille.

La part de femme se calcule $PartFemme = \frac{Exposition\ des\ femmes}{Somme\ des\ expositions\ des\ hommes\ et\ femmes}$.

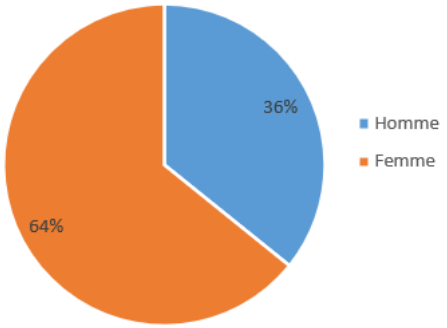

Figure 9. Répartition par genre

Le portefeuille est majoritairement constitué de la catégorie socio professionnelle n°6 *(le pourcentage n'est pas indiqué pour préserver la confidentialité)*. La codification des CSP est disponible en annexes au paragraphe §8.2 CSP).

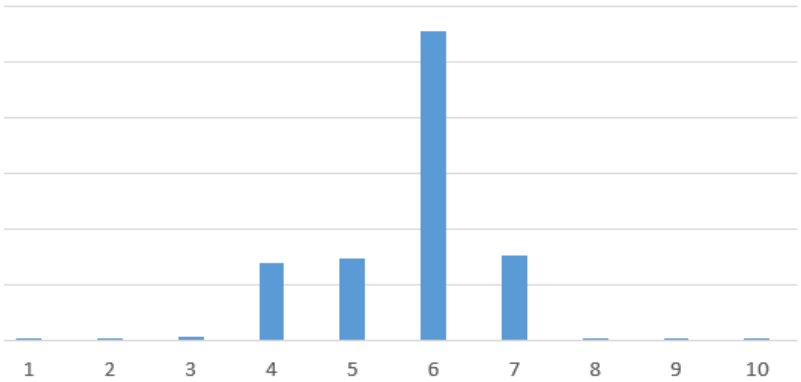

Figure 10. Répartition par catégorie socio professionnelle

La population des 20-40 ans est la plus représentée.

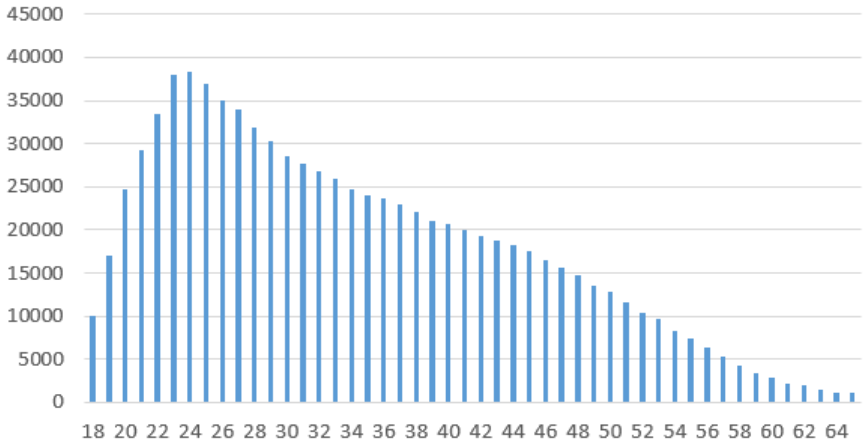

Figure 11. Répartition des assurés selon l'âge à la souscription

2,5% des valeurs sont retirées du portefeuille. Il s'agit des individus avec des âges à la souscription inférieurs à 18 ans ou supérieurs à 65 ans.

Nombre d'individus	En nombre	% par rapport au départ
Après retraitement	880 076	100%
Après filtrage des âges non pertinents à la souscription	857 701	97,5%

Pour le calcul de l'incidence en incapacité à utiliser pour la tarification, l'hypothèse retenue est la suivante : 100% du portefeuille a souscrit aux garanties « incapacité ».

Malgré le volume restreint de données à disposition, il s'avère que l'OA réalise ses propres analyses statistiques. Il convient de tester la représentativité de l'échantillon à disposition vis-à-vis du portefeuille de l'OA. Le test de représentativité est réalisé à minima sur le genre.

Pour vérifier si l'échantillon est représentatif de la distribution du genre, il convient de calculer l'intervalle de fluctuation au seuil de 95% et la fréquence de femmes dans l'échantillon.

Soit X la variable aléatoire donnant le nombre de femmes du portefeuille de l'OA dans un échantillon aléatoire de taille $n = 880\,076$ individus.

Chaque individu est soit un homme ou une femme. X suit une loi binomiale de paramètre $n = 880\,076$ et de probabilité $p = 68\%$ (cette probabilité est communiquée par l'OA).

L'intervalle de fluctuation I s'écrit $[p - 1{,}96 * \sqrt{\frac{p*(1-p)}{n}}\ ;\ p + 1{,}96 * \sqrt{\frac{p*(1-p)}{n}}]$. Il s'agit de l'intervalle de fluctuation asymptotique au seuil 95% de la fréquence de femmes dans l'échantillon.

En pratique $I = [0{,}679; 0{,}681]$

Et la fréquence f de femmes dans l'échantillon est égale $f = \frac{nombre\ de\ femme*exposition}{somme\ des\ expositions\ hommes\ et\ femmes} = 64\%$.

f n'appartient pas à l'intervalle de fluctuation. L'échantillon n'est pas représentatif du genre.

Par ailleurs, l'échantillon à disposition n'est pas homogène sur les années d'exposition : les individus avec un contrat démarrant en 2019 représentent 25% du portefeuille total. Ils sont sur-représentés dans l'échantillon.

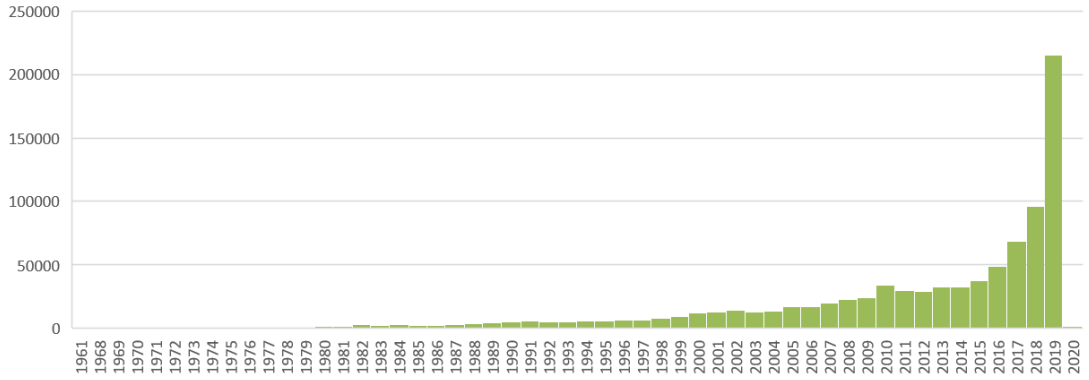

Figure 12. Répartition des individus par date de début de contrat

L'augmentation du nombre de contrats entre 2018 et 2019 peut s'expliquer par le recours à des contrats courts type CDD en lien avec des politiques volontaristes en faveur de l'action sociale et de l'accès à l'emploi.

La typologie de contrat CDD / CDI, etc. n'était pas disponible dans la base de données. Elle aurait pu permettre d'affiner la répartition des individus par contrat.

Une fois la population sous risque analysée, il convient d'étudier les données sur la sinistralité.

3.2. LA REPARTITION DES SINISTRES

L'étude des caractéristiques des sinistres indemnisés s'appuie sur la base de gestion des sinistres d'historique de dix ans de données. Le périmètre étudié représente 1 135 806 sinistres indemnisés.

79% des individus indemnisés sont des femmes. 52% des personnes indemnisées sont célibataires, près de 40% sont mariées, pacsés ou en union libre. L'information de la situation maritale est recueillie au moment de l'affiliation (lorsqu'elle est réalisée) et n'est pas toujours mise à jour à l'enregistrement de la déclaration de sinistre.

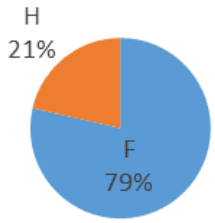

Figure 13. Répartition des sinistres par genre

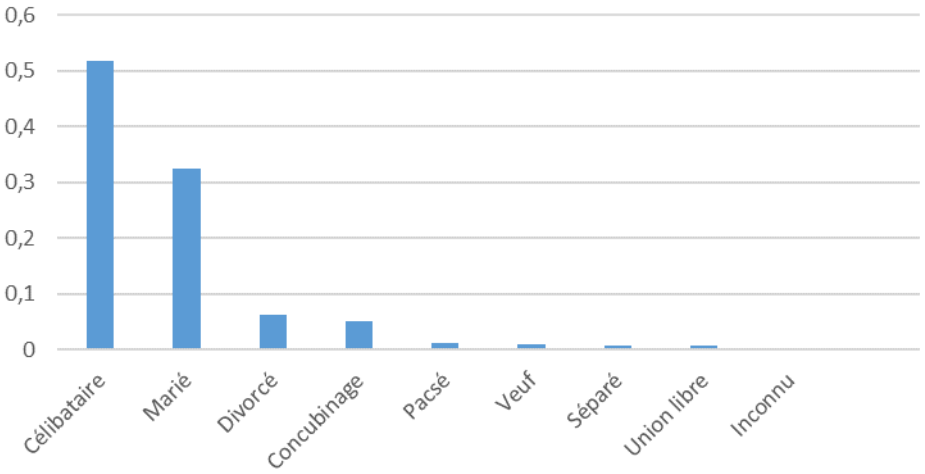

Figure 14. Répartition selon la situation maritale

Les arrêts sont principalement pour « maladie » sans précision. Les garanties les plus consommées du portefeuille sont les IJ1 et IJ2.

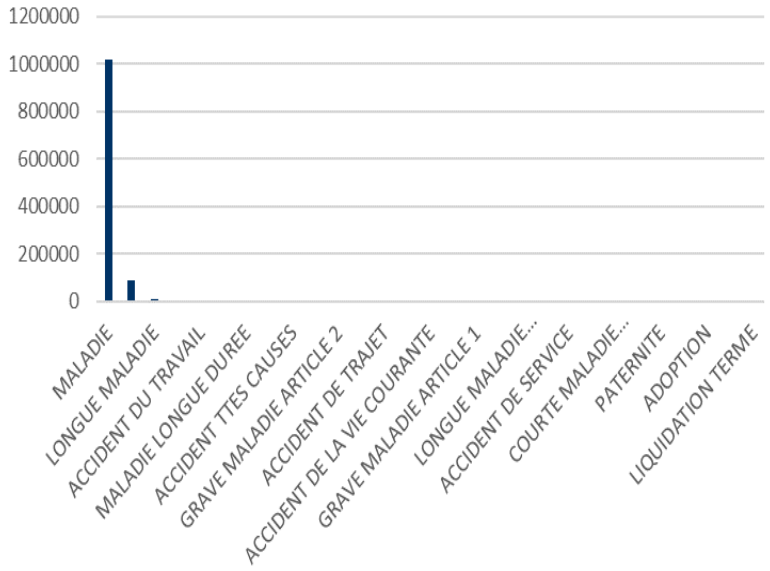

Figure 15. Répartition par cause d'arrêt

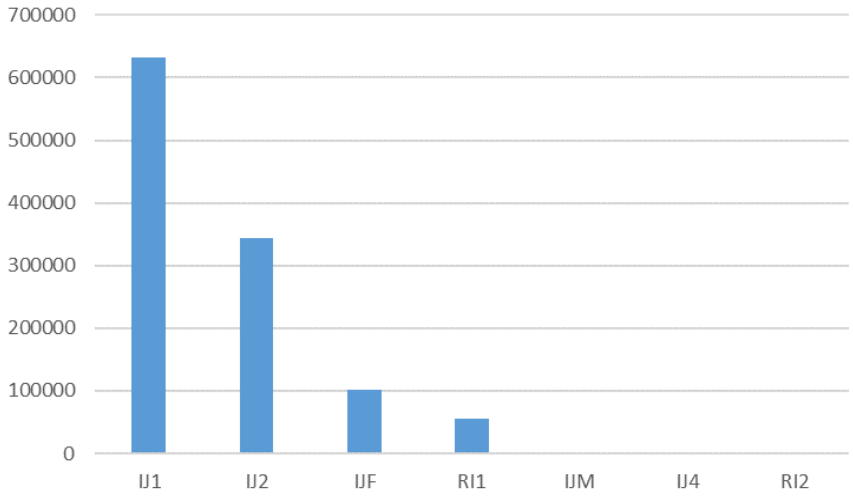

Figure 16. Répartition selon les garanties

Le graphe ci-dessous présente la répartition des CSP dans la base des sinistres.

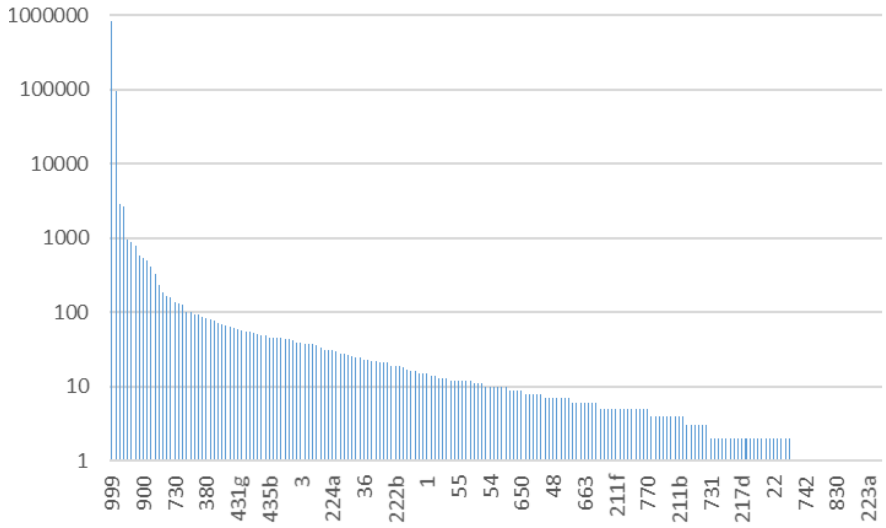

Figure 17. Répartition selon le code CSP

Ce graphe n'est finalement pas exploitable puisque la majorité des CSP est codifiée à 999. Il ne peut donc pas être envisagé d'exploiter cette information. Elle est disponible dans le support DSN (pour l'étude de l'incidence).

La nomenclature de l'OA décrit la codification de la CSP de la manière suivante :

- Chiffre commençant par 1 ou 2 : agriculteurs, artisans, commerçants et chefs d'entreprises,
- Chiffre commençant par 3 : les cadres,
- Chiffre commençant par 4 : les techniciens, agents de maîtrise et autres professions intermédiaires,
- Chiffre commençant par 5 ou 6 : les employés ou ouvriers.

Les populations en arrêt sont surtout les 30 – 35 ans et les individus situés entre 40 – 60 ans avec un pic à 50 ans.

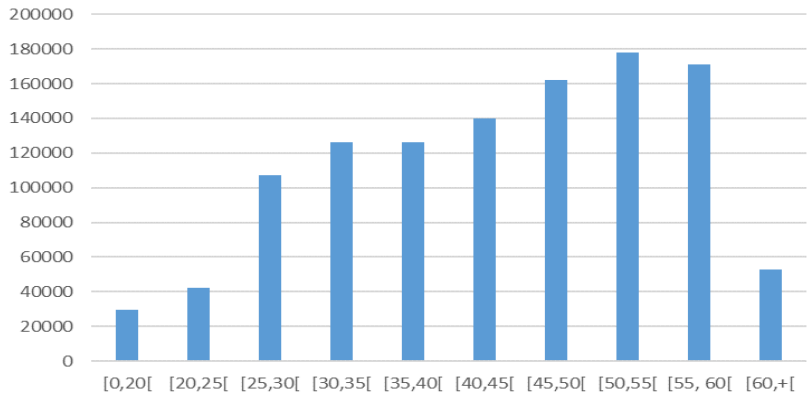

Figure 18. Répartition selon la tranche d'âge à la survenance

Les zones géographiques les plus touchées sont Paris (75), le département Nord Pas de Calais (62), la Gironde (33) et le Rhône (69) et l'Isère (38).

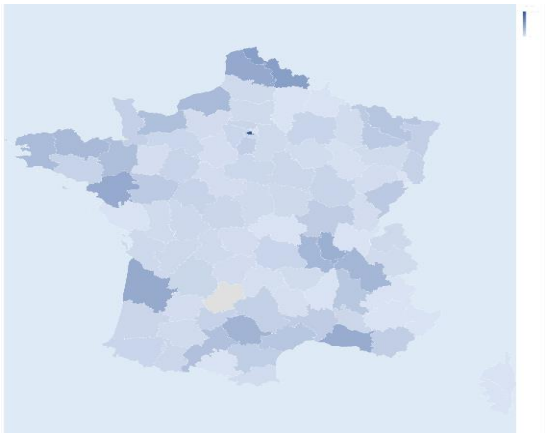

Figure 19. Répartition des sinistres du portefeuille selon les départements

Au fur et à mesure des années, la dérive de la sinistralité se confirme. Pour rappel, l'année 2019 n'est pas observée dans sa totalité.

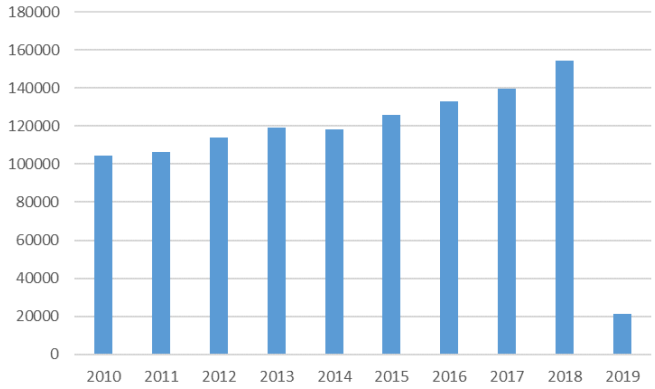

Figure 20. Répartition par année de survenance

Les sinistres sont plus nombreux pendant les mois d'automne et surtout en hiver. Chaque année, la tendance semble se confirmer.

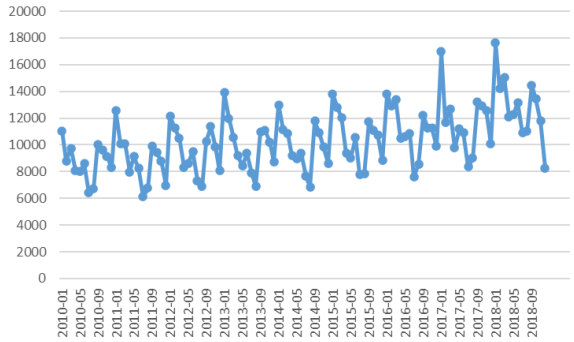

Figure 21. Répartition mensuelle

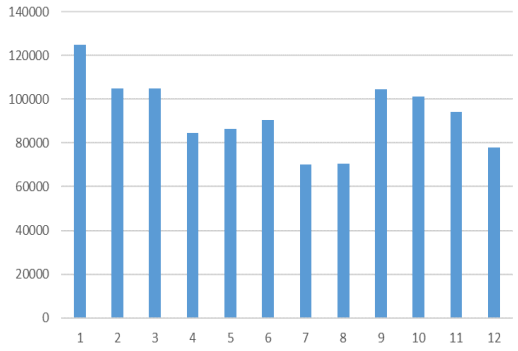

Figure 22. Répartition mensuelle cumulée

Les arrêts maladie démarrent le lundi généralement et a priori en début de mois.

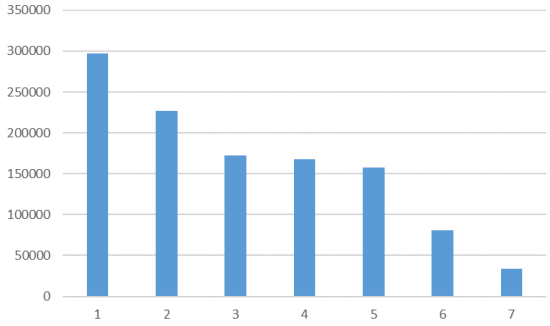

Figure 23. Répartition selon le jour de démarrage dans la semaine

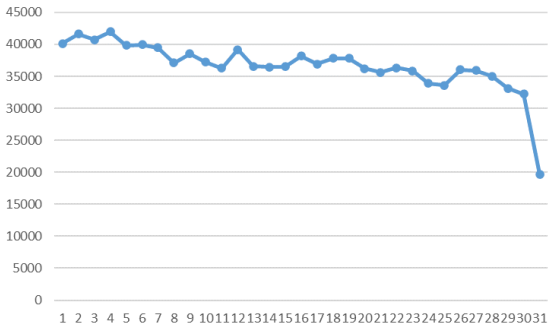

Figure 24. Répartition selon le jour de démarrage dans le mois

La répartition du nombre d'arrêts par tranche de durée est représentée dans le graphe ci-dessous : de nombreux arrêts courts sont observés.

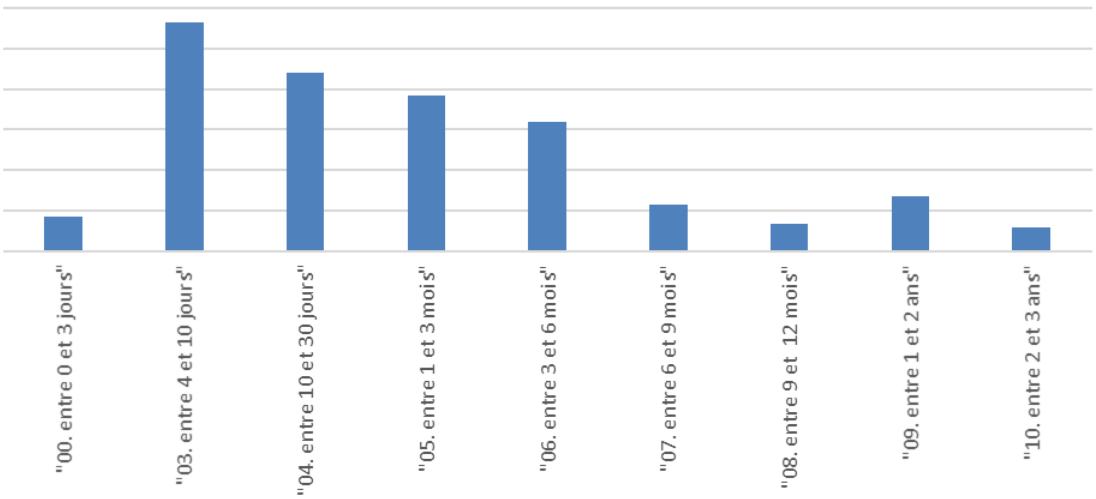

Figure 25. Répartition du nombre des arrêts par tranche de durée

D'après les données à disposition, les deux graphes ci-dessous montrent que la durée d'indemnisation par l'OA est significative par rapport à la durée d'incapacité. Elle s'élève en moyenne à 83% en passant par un minimum de 65% (pour les durées d'incapacité de 90 jours). Les maximums sont atteints à la fois pour les très courtes durées et les plus longues (~2 ans).

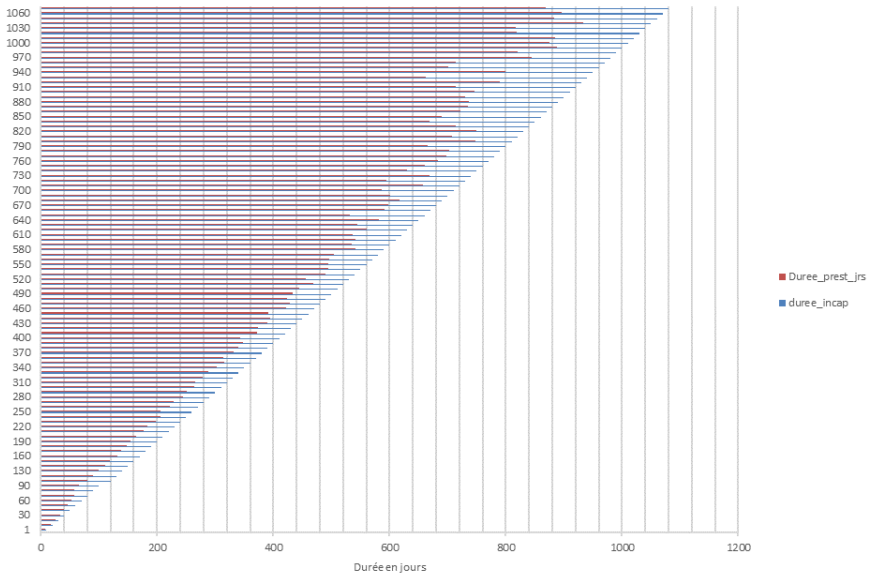

Figure 26. Durées moyennes d'indemnisation (en rouge) pour chaque durée d'incapacité (en bleu et par pas de 10 jours)

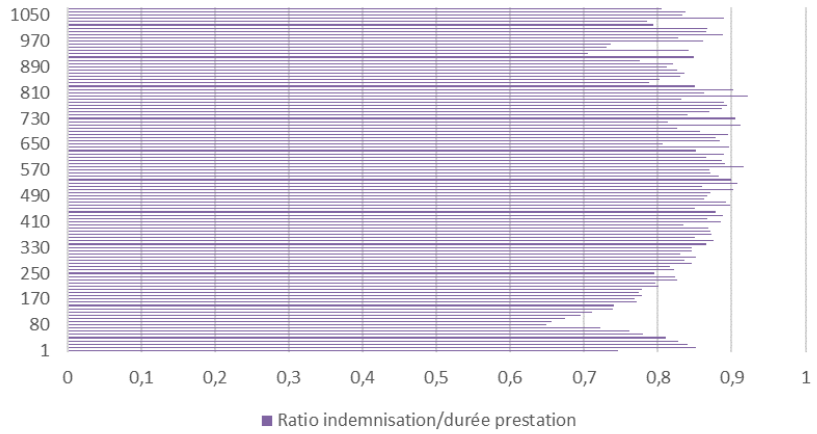

Figure 27. Rapport entre la durée d'indemnisation par l'OA et la durée d'incapacité (par pas de 10 jours)

Les deux graphes ci-dessous montrent que le portefeuille sélectionné est exposé :

- Soit à de nombreux petits arrêts pour 33% des cas (tranche 00),
- Soit à des arrêts dont les montants sont plus conséquents (tranches 10 et 11).

Les tranches 01 à 09 du graphe « répartition selon les montants » sont constituées par pas de 100€. Les tranches 10 et 11 sont par pas de 1 000€.

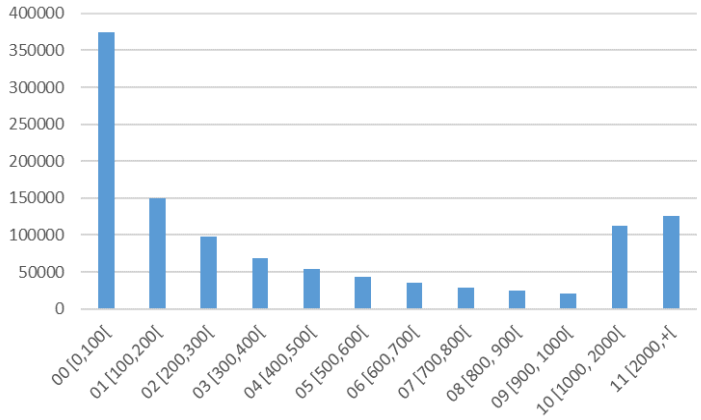

Figure 28. Répartition des arrêts selon les montants d'indemnisation

Le graphe ci-dessous montre que l'organisme assureur doit faire face à un nombre de sinistres :

- Moyen pour les durées comprises entre 3 mois et 1 an,
- Elevé pour :
    - Les durées entre 1 et 3 mois (tranches 6 à 11),
    - Les très grandes durées,
- Très élevé pour les courtes durées (tranches de 1 à 4).

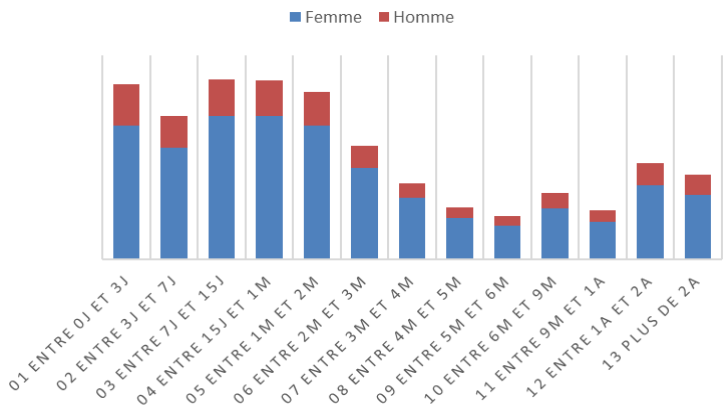

Figure 29. Répartition selon les durées d'indemnisation

Théoriquement, les sinistres sont indemnisés après un délais de franchise. Cependant le graphe ci-dessous montre que les sinistres sont indemnisés majoritairement dès le troisième jour après la franchise.

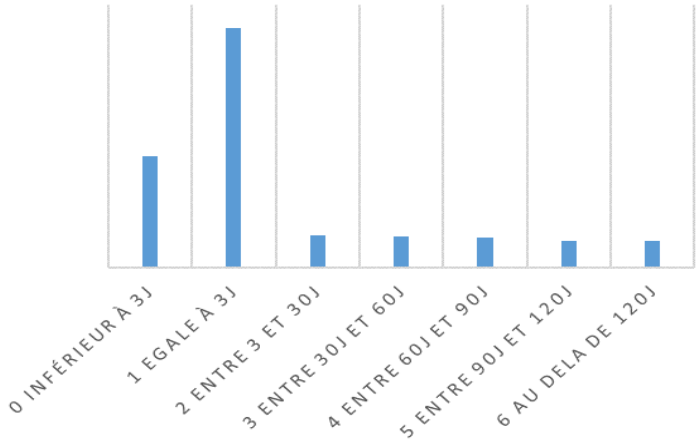

Figure 30. Répartition selon les délais de carence

Les analyses descriptives et de sinistralité sont réalisées. Il convient désormais de reconstruire les lois de probabilité d'incidence et de maintien en incapacité. Ces lois permettent de modéliser le risque incapacité.

La loi d'incidence a pour objectif de modéliser pour chaque âge la probabilité de tomber en arrêt de travail.

Quant à la loi de maintien, elle modélise également pour chaque âge le phénomène de durée en arrêt de travail sachant l'historique.

La connaissance de ces deux lois permettra de construire le modèle de tarification.

4. RECONSTRUCTION EMPIRIQUE DES LOIS DE MAINTIEN ET D'INCIDENCE

4.1. CONSTRUCTION DE LA LOI DE MAINTIEN

Cette partie est consacrée à la modélisation de la loi de maintien avec les données d'expérience. Tout d'abord, les clés de lecture d'une table de maintien sont rappelées. Ensuite, une proposition d'estimation de cette loi ainsi qu'une analyse comparative avec les données du marché sont réalisées.

4.1.1) Grille de lecture d'une table de maintien

a) Fonction et probabilités de maintien

Soit $\tilde{T}_x$ la variable aléatoire qui désigne la durée de maintien en incapacité d'un individu d'âge x. Soit t un réel strictement supérieur à 0. Les trois quantités nécessaires à la modélisation sont les suivantes :

Quantités	Détails	Définitions
$S_x(t)$	$= P(\tilde{T}_x > t)$	Probabilité pour un individu d'âge x en incapacité de rester dans cet état au moins t jours.
${}_1q_{x,t}$	$P(t < \tilde{T}_x < t+1 \mid \tilde{T}_x > t) = 1 - \frac{S_x(t+1)}{S_x(t)}$	Probabilité pour un individu d'âge x de quitter l'état d'incapacité entre la période t et $t+1$ sachant qu'il était en incapacité à t.
${}_1p_{x,t}$	$= \frac{S_x(t+1)}{S_x(t)}$	Probabilité qu'un individu d'âge x reste en incapacité entre t et $t+1$ sachant qu'il était en incapacité à t.

b) Lecture des tables de maintien

Pour lire les trois quantités précédentes, il est d'usage d'exploiter une table de maintien. Il s'agit d'un tableau à deux entrées :

	Entrée 2 - Ancienneté dans l'état incapacité			
Entrée 1 : Age d'entrée en incapacité	0	1	t	…
x	$l_{x,0}$	$l_{x,1}$	$l_{x,t}$	…
$x+1$	$l_{x+1,0}$	$l_{x+1,1}$	$l_{x+1,t}$	…
…	…	…	…	…

Les $l_{x,t}$ représentent l'effectif pour un âge et une ancienneté donnés.

Pour chaque âge x :

- Il est d'usage de partir de 10 000 individus pour l'ancienneté 0,
- Plus l'ancienneté en incapacité croit, plus l'effectif en incapacité décroit compte-tenu des sorties.

Par lecture directe de la table de maintien, il est possible de calculer les trois quantités suivantes :

- $S_x(t) = \frac{l_{x,t}}{l_{x,0}}$,
- ${}_1p_{x,t} = \frac{l_{x,t+1}}{l_{x,t}}$,
- ${}_1q_{x,t} = \frac{d_{x,t}}{l_{x,t}}$ avec $d_{x,t} = l_{x,t} - l_{x,t+1}$ qui désigne le nombre de sorties de l'état d'incapacité.

4.1.2) Estimateur de Kaplan Meier

L'objet est d'estimer $S_x(t)$, la probabilité de maintien en incapacité pour un individu d'âge x.

Pour cela, l'estimateur de Kaplan Meier est généralement utilisé pour les raisons suivantes :

- Il est non paramétrique. Cela permet d'éviter d'avoir un a priori sur la distribution des variables aléatoires de durée,
- Il est robuste et plébiscité par la communauté des actuaires : l'objet principal du mémoire n'est pas d'établir une table d'expérience à certifier par un actuaire indépendant. Parmi l'ensemble des mémoires disponibles, certains mémoires d'actuariat (Mario Gugumus, 2009), (Camille Mosse, 2007) expliquent très bien la démarche.

a) Les variables du modèle / notions de troncature et la censure

Pour i allant de 1 à n avec n le nombre d'observation de la base de données étudiée.

Soit n le nombre d'observations et $i \in [\![1, n]\!]$ un individu quelconque. On note T_i la variable aléatoire qui représente la durée de maintien en incapacité de l'individu i.

Les variables aléatoires du modèle sont :

Quantités	Modalités de calcul	Définitions
T_i	= date de reprise – date de survenance du sinistre	Durée réelle de maintien en incapacité de l'individu i. Cela désigne le temps écoulé entre la date de survenance du sinistre et la date de sortie
C_i	= date de départ de l'observation – date de survenance du sinistre	Durée de maintien en incapacité de l'individu i écoulée avant la censure. Représente le temps écoulé entre la date de survenance et la date d'observation d'extraction lorsque l'observation est censurée (ex : les arrêts de travail en cours à la date d'extraction sont considérés comme censurés)
Y_i	$\min(C_i, T_i)$	Durée de maintien en incapacité jusqu'au départ de l'observation de l'individu i.

Lors d'une observation censurée, l'individu est observé jusqu'à C_i, et particulièrement, son maintien en incapacité jusqu'à cette date. Il s'agit d'une information à exploiter, car sans cela, les observations maintenues en incapacité jusqu'à C_i seraient supprimées, ce qui induirait une sous estimation de la durée de maintien en incapacité.

b) *Expression de l'estimateur et mode de fonctionnement*

Pour approcher la loi de maintien en incapacité, l'estimateur de Kaplan Meier est utilisé. Il s'écrit de la manière suivante : $\hat{S}_x(t) = \prod_{Y_i}^{t}(1 - \hat{q}(Y_i))$

La base de données est ordonnée par ordre croissant des Y_i.

Il se construit de proche en proche à partir d'un tableau d'évènements qui dénombre pour chaque pas de temps les quantités suivantes :

1) Le nombre d'individus sortis de la cohorte. Ce nombre correspond à la somme des quantités 2) et 3)
2) Le nombre d'individus dont la sortie de l'état d'incapacité est réellement observée,
3) Le nombre d'individus dont la sortie de la cohorte est liée à une censure,
4) Le nombre d'individus entrés dans l'état d'incapacité. Ce nombre est systématique égal à 0. En effet, pour chaque âge, une cohorte est observée.
5) Le nombre d'individus maintenus dans l'état incapacité.

c) *Mise en œuvre sous Python*

Sous Python, le package *lifelines* est utilisé avec la méthode *KaplanMeierFitter().* Cette méthode utilise principalement les paramètres de durée (colonne « Duree » de la base de données) et de censure (colonne « Censure2 »). Sous python, le paramètre de censure se lit comme suit :

- Les évènements censurées sont valorisés à « False » ou « 0 »,
- Les évènements non censurés sont à « True » ou « 1 ».

Une fois la méthode implémentée, le modèle permet d'établir les tables d'évènements pour chaque âge. La sortie python d'une telle table se présente sous la forme suivante :

event_at	removed	observed	censored	entrance	at_risk
0.0	0	0	0	417474	417474
1.0	29990	29990	0	0	417474
2.0	27656	27656	0	0	387484
3.0	32145	32145	0	0	359828

A partir de ces tables et de proche en proche, l'estimateur de Kaplan Meier se constitue de la manière suivante :

$$S(1) = \left(1 - \frac{\#\ d'individus\ dont\ la\ date\ de\ fin\ est\ observée\ (29\,990) + \#\ d'individus\ censurés\ (0)}{\#\ d'individus\ sous\ risque\ à\ la\ date\ t\ (417\,474)}\right)$$

$S(1) = \left(1 - \frac{29\,990}{417\,474}\right) = 92{,}8\%$ de la population est susceptible de dépasser une durée d'incapacité supérieure au pas de temps « 1 ».

$$S(2) = \left(1 - \frac{29\,990}{417\,474}\right) * \left(1 - \frac{27\,656}{387\,484}\right) = 86{,}2\%$$

d) *Construction de l'intervalle de confiance*

La littérature présente l'estimateur de Greenwood comme l'estimateur convergent de la variance de Kaplan Meier. Soit $0 \leq t_1 \leq \cdots \leq t_k \leq t$ une discrétisation du temps t.

$$\hat{V}_{\mathrm{n}}\left(\hat{S}_{\mathrm{n}}(t)\right) = \hat{S}_{\mathrm{n}}^2 * \sum_{t_k}^{t} \frac{\hat{q}(t_k)}{1 - \hat{q}(t_k)}$$

Il est donc possible de construire des intervalles de confiance :

$S(t) = \hat{S}_x(t) \pm u * \frac{\hat{\sigma}}{t^{\frac{1}{2}}}$ avec u est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale $N(0,1)$ et $\hat{\sigma}^2 = \sum_{t_k}^{t} \frac{\hat{q}(t_k)}{1 - \hat{q}(t_k)}$

En pratique, sous Python, la propriété *confidence_interval_* du modèle KaplanMeierFitter est utilisée pour calculer les intervalles de confiance.

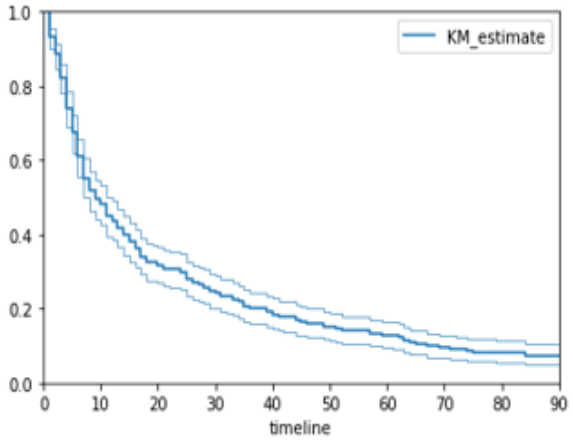

Figure 31.Graphe de la loi de maintien avec intervalle de confiance à 95%

	KM_estimate_lower_0.95	KM_estimate_upper_0.95
0.0	1.000000	1.000000
1.0	0.899953	0.954524
2.0	0.846381	0.914848
3.0	0.778678	0.860113
4.0	0.691148	0.784387
5.0	0.620885	0.720554

Figure 32. Restitution des valeurs de l'intervalle de confiance à 95% de la loi de maintien

e) Lissage de la courbe par la méthode de Whittaker-Henderson

i) Principe

Les courbes obtenues précédemment présentent des irrégularités.

Ces irrégularités n'ont pas de signification réelle et sont dues à la construction de l'échantillon : les données à disposition permettent d'approcher empiriquement la loi réelle d'incidence.

Conformément à l'absence d'a priori sur la distribution des observations, un lissage non paramétrique classique est utilisé : la méthode de Whittaker-Henderson.

ii) Utilisation de la méthode de Whittaker-Henderson

Le principe général est de trouver une courbe lissée convenable en combinant les deux critères suivants :

- « F » : la fidélité qui favorise l'aspect brut de la courbe,
- « S » : la régularité qui favorise l'aspect plus régulier de la courbe.

Cela revient à minimiser le problème exprimé sous la forme littérale suivante : $WH_h(c) = F(c) + h * S(c)$.

Avec :

- $c : k \to q_k$ qui représente la courbe lissée de la loi de maintien.
- Les deux critères s'exprimant de la manière suivante :

Critères	Expression littérale	Expression matricielle
Fidélité	$F(c) = \sum_{k=Y_i}^{t} w_k * \left(q_k - \widehat{q_k}\right)^2$ avec $w_k \geq 0$ les poids attribués à chacune des observations.	$F(c) = (c - \hat{c})^T W (c - \hat{c})$ avec W la matrice des poids w_k
Régularité	$S(c) = \sum_{k=Y_i}^{t} ((\Delta^2 q_k)^2)$ avec $\Delta q_k = q_{k+1} - q_k$	$S(c) = c^T K_2^T K_2 c$ avec $K de \dim (t - 2, t)$ *de telle sorte que* $(\Delta^2 c = K_2 c)$

- h le paramètre de lissage (si $h \to 0$ alors la courbe lissée est égale à la courbe empirique).

La solution obtenue en minimisant sous la norme 2 s'exprime sous forme de matrice de la manière suivante : $WH(h) = (W + hK^T K)^{-1} W\hat{c}$ avec : $\hat{c} : k \to \widehat{q_k}$ la courbe de la loi empirique de maintien.

f) Résultats, comparaison entre loi d'expérience et loi du BCAC

Grâce aux données d'expérience, il s'avère que l'âge est un facteur aggravant du phénomène de maintien en incapacité. L'exposé des graphes dans ce paragraphe adopte le principe suivant : pour chaque durée en nombre de jours (représentée en abscisse), la probabilité de dépasser la durée (représentée en ordonnée) est tracée.

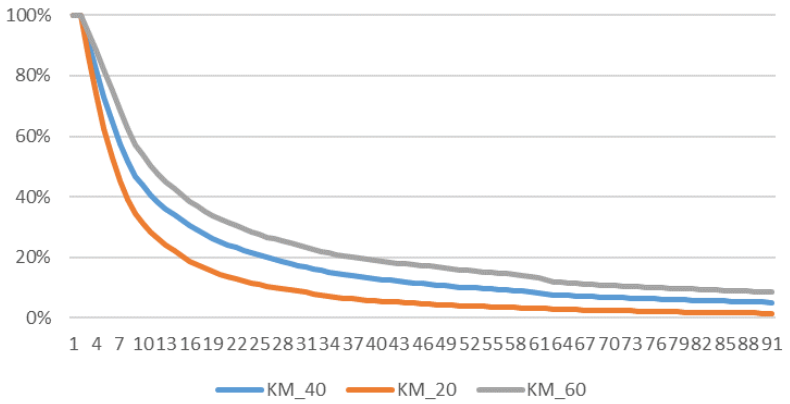

Figure 33. Graphe comparatif des lois de maintien pour trois âges différents

Par lecture graphique, la courbe de la loi de maintien pour un individu de 60 ans est au-dessus de celle pour un individu de 40 ans également au-dessus de celle de 20 ans. Pour une même durée en incapacité, la probabilité de rester en arrêt au-delà de cette durée croit avec l'âge à la survenance.

Les lois de maintien obtenues à partir des données d'expérience et celles du BCAC sont comparées.

Le graphe ci-dessus de la loi de maintien en incapacité des individus de 40 ans est exposé. Cet âge constitue la référence pour la tarification « socle ». Quant au graphe ci-dessous, les âges à la survenance sont regroupés en 4 tranches suivantes :

- tranche 1 De 16 à 30 ans inclus,
- tranche 2 De 31 à 40 ans inclus,
- tranche 3 De 41 à 50 ans inclus,
- tranche 4 De 51 à 67 ans inclus.

Sur un même graphique, les courbes d'expérience (pas journalier) et celles du BCAC 2013 (pas mensuel) sont superposées. C'est justement grâce à l'exploitation des flux DSN qu'il est désormais possible d'élaborer une courbe d'expérience par pas journalier. En effet, l'entreprise doit déclarer l'ensemble des arrêts de travail de ses salariés et ce, dès le premier jour d'arrêt. Ces informations sont désormais connues des organismes de prévoyance complémentaires.

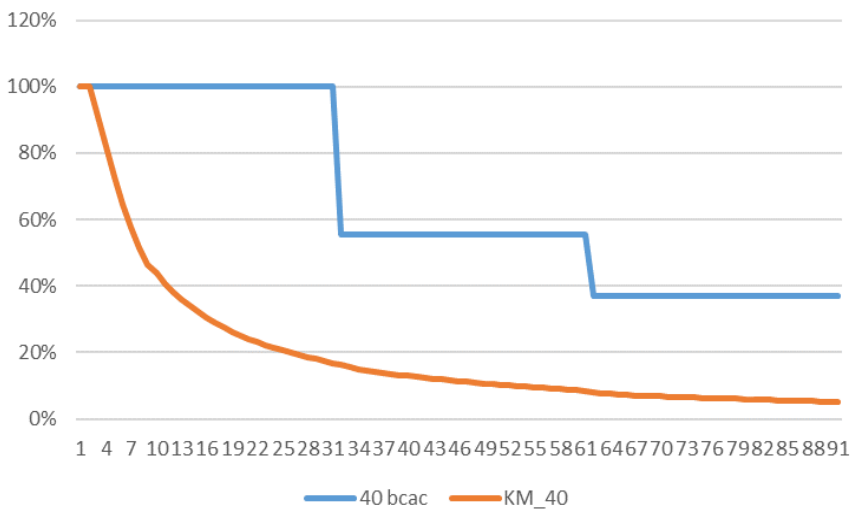

Figure 34. Graphe comparatif des lois de maintien d'expérience et BCAC pour l'âge 40 ans

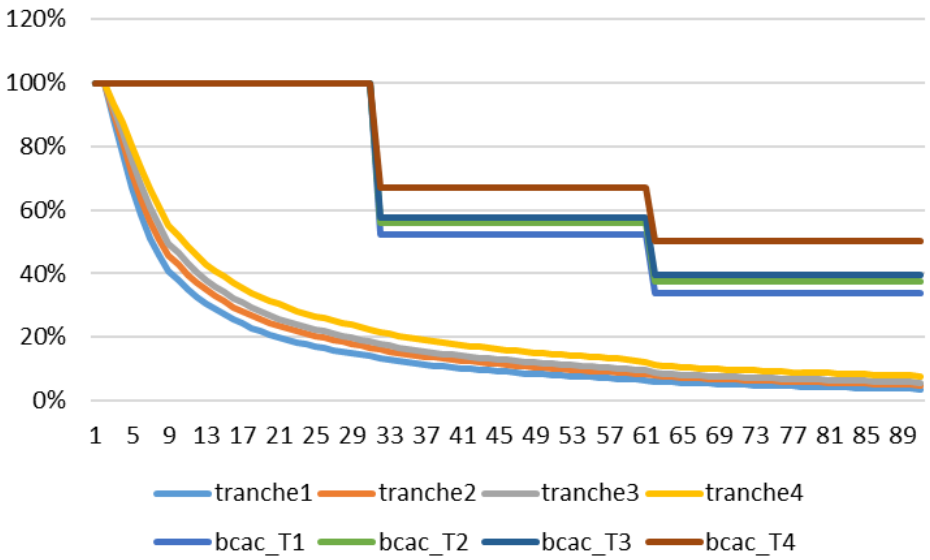

Figure 35. graphe comparatif des lois de maintien d'expérience et BCAC pour les 4 tranches d'âge

Par lecture graphique, il est aisé de constater que les courbes d'expérience sont nettement en dessous des courbes du BCAC.

Cette lecture graphique permet de conclure que les données du BCAC surestiment le phénomène de maintien en incapacité par rapport au portefeuille étudié.

Pour les grandes durées en incapacité, les probabilités de dépasser les grandes durées convergent :

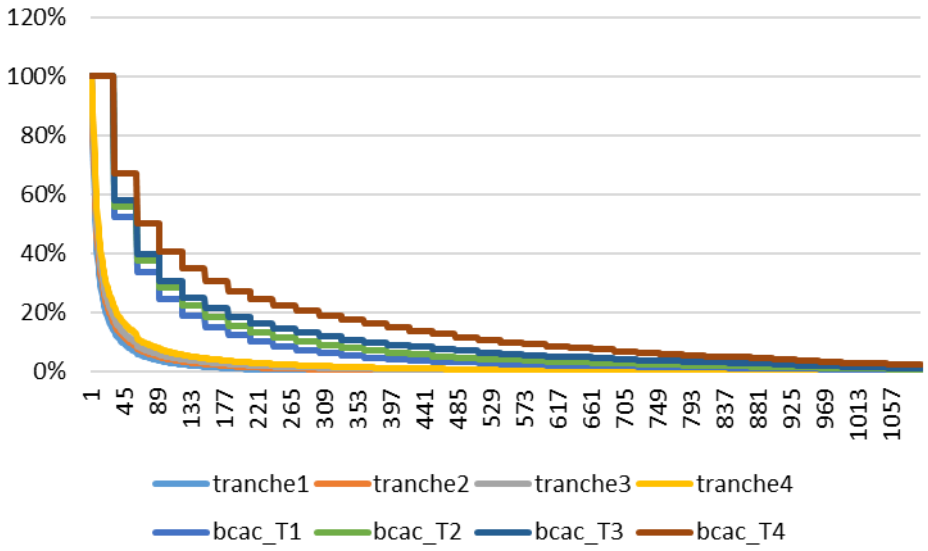

Figure 36. Graphe comparatif des lois de maintien d'expérience et BCAC par tranche et jusqu'à 1095 jours

Par lecture graphique, il s'avère qu'entre les jours 0 et 500, les courbes d'expérience restent en dessous des lois de maintien du BCAC.

Les lois BCAC sont plus prudentes que les lois obtenues par rapport aux données d'expérience.

g) Les limites

L'une des limites du modèle est d'avoir utilisé une méthode non paramétrique. Il convient donc d'être vigilant à la taille de l'échantillon puisque ce type de méthode accorde davantage d'importance à la réalité des données brutes. Ayant 10 ans d'historique de données, cela est susceptible de justifier le recours à une méthode non paramétrique : l'effet « tendance » sur la loi de maintien a pu être observé.

Enfin, la table d'expérience nécessite d'être certifiée par un actuaire certifié. Il doit accompagner l'OA à correctement maîtriser le risque de son portefeuille.

Par ailleurs, il pourrait être intéressant de comparer les résultats avec une loi du BCAC actualisée et plus récente que celle de 2013. Mais cette loi n'est pas utilisée par l'OA.

Après avoir construit la loi de maintien en incapacité, il convient de s'intéresser à la deuxième composante de la tarification : la loi d'incidence en incapacité.

4.2. CONSTRUCTION DE LA LOI D'INCIDENCE

La construction de la loi de fréquence s'effectue par reconstitution des taux d'entrée en incapacité.

Cette loi attribue le taux moyen de sinistre pour chaque groupe d'individus de la tranche d'âge $[x, x + pas\ de\ la\ tranche[$.

Pour chaque individu du groupe, il est possible d'observer entre 0 et plusieurs entrées en incapacité.

Soit N_x le nombre d'entrées en incapacité. $N_x = \sum_{i=1}^{n_x} X_i$ avec :

- n_x le nombre d'individus dans la classe d'âge x
- Les X_i sont les variables aléatoires associées à chaque observation et comportent deux issues possibles : capable ou en incapacité. De fait, les $X_i$suivent une loi de Bernoulli de paramètre p_x :

$$X_i \sim Bernoulli(p_x)$$

N_x suit une loi binomiale $B(n_x, p_x)$. Cette loi peut être approximée par une loi de Poisson de paramètre $\lambda_x > 0$ avec $n_x p_x \to \lambda_x$. (la revue de statistique appliquée d'E. Morice et P. Thionet rappelle ce résultat).

Un mémoire d'actuariat (Maxime Huttin, 2013) mentionne la démarche d'estimation de λ_x par un estimateur facile de mise en œuvre et convergent (i.e. $\lim_{n \to +\infty} P\left(\left|\widehat{\theta_n} - \theta\right| > \epsilon\right) = 0 \ \forall \theta$. La probabilité d'observer un écart est de plus en plus faible avec le nombre d'observations). Nous proposons d'estimer le paramètre λ_x, de rappeler quelques propriétés de l'estimateur choisi et de présenter les résultats obtenus pour la reconstitution de la loi d'incidence.

4.2.1) L'estimateur de λ_x

L'estimateur par maximum de vraisemblance est couramment utilisé.

Soit L la fonction de vraisemblance :

$$L = \prod_{i=1}^{n_x} P(X_i = x_i) = \prod_{i=1}^{n_x} \frac{\lambda_i^{x_i}}{x_i!} * \exp(-\lambda_i)$$

Pour maximiser cette égalité, il convient de déterminer λ tel que $\frac{\delta \ln(L)}{\delta \lambda} = 0$.

En premier lieu, il est d'usage de linéariser l'expression pour faciliter les calculs. La fonction logarithme népérien est utilisée :

$$\ln(L) = -\sum_{i=1}^{n_x} \lambda_i + \sum_{i=1}^{n_x} [x_i * \ln(\lambda_i) - \ln(x_i!)] \qquad (1)$$

$$\ln(L) = -\lambda_x \sum_{i=1}^{n_x} (b_i - a_i) + \sum_{i=1}^{n_x} [x_i * \ln(\lambda_x (b_i - a_i)) - \ln(x_i!)] \qquad (2)$$

$$\ln(L) = -\lambda_x E_x + \ln(\lambda_x) N_x + \sum_{i=1}^{n_x} [x_i * \ln(b_i - a_i) - \ln(x_i!)] \qquad (3)$$

Le passage de la ligne (1) à (2) s'explique par une approximation linéaire telle que $\lambda_i \approx \lambda_x * (b - a)$ avec $[a; b]\ C [x; x + pas\ de\ la\ tranche[$. Cette approximation est issue de l'hypothèse métier suivante : entre la

date de souscription et de la résiliation, un assuré n'est pas forcément exposé au risque durant toute une année. Cette approximation correspond donc à l'exposition durant une portion de l'âge x.

Le passage de la ligne (2) à (3) s'effectue par introduction des quantités :

- E_x la somme des expositions pour les individus d'âge x
- N_x le nombre d'entrée en incapacité pour les individus d'âge x

Par dérivation, $\frac{\delta \ln(L)}{\delta \lambda} = -E_x + \frac{N_x}{\lambda_x} = 0$. L'estimateur de λ s'exprime donc $\hat{\lambda} = \frac{N_x}{E_x}$.

4.2.2) Quelques propriétés de l'estimateur

L'étude retient l'hypothèse d'une indépendance des entrées en incapacité.

$E(\widehat{\lambda_x}) = E\left(\frac{N_x}{E_x}\right) = \frac{\sum_{i=1}^{n_x} \lambda_x * (b_i - a_i)}{\sum_{i=1}^{n_x} (b_i - a_i)} = \lambda_x$

$Var(\hat{\lambda}_x) = Var\left(\frac{N_x}{E_x}\right)$

(1) $= \frac{1}{E_x^2} Var(N_x)$ s'obtient par approximation de E_x par sa partie entière $Var(\hat{\lambda}) = \frac{Var(N_x)}{E_x^2}$

(2) $= \frac{Var(\sum_{i=1}^{n_x} Poisson(\lambda_x (b_i - a_i)))}{E_x^2}$ s'obtient par approximation $\lambda_x \approx \lambda(b - a)$, $N_x \sim Poisson\left(\sum_{i=1}^{n_x} \lambda(b_i - a_i)\right) = \sum_{i=1}^{n_x} Poisson(\lambda(b_i - a_i))$.

$Var(\hat{\lambda}_x) = \frac{\lambda_x * \sum_{i=1}^{n_x} (b_i - a_i)}{E_x^2} = \frac{\lambda_x}{E_x}$

D'après le théorème central limite, l'intervalle de confiance pour λ au niveau u se construit :

$IC(\lambda_x) = [\widehat{\lambda_x} - u * \sqrt{\frac{\widehat{\lambda_x} * (1 - \widehat{\lambda_x})}{E_x}}; \widehat{\lambda_x} + u * \sqrt{\frac{\widehat{\lambda_x} * (1 - \widehat{\lambda_x})}{E_x}}]$ avec u le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale $N(0,1)$.

Pour rappel, le théorème central limite dit que $\frac{\sqrt{n}\,(\widehat{\lambda_x} - \lambda)}{\sqrt{Var(\widehat{\lambda_x})}} \overset{L}{\underset{n \to \infty}{\to}} N(0,1)$.

4.2.3) Les résultats

Cette partie est dédiée à la présentation de la démarche pour estimer λ_x $(= E(\widehat{\lambda_x}) = E\left(\frac{N_x}{E_x}\right))$ puis la construction de la loi d'incidence. Nous reconstituons les quantités suivantes :

- La somme des expositions au risque E_x,
- Le nombre d'entrées en incapacité par tranche d'âge N_x.

Ensuite, nous vérifions l'adéquation entre la loi du nombre d'entrées en incapacité et la loi de Poisson.

Enfin, les courbes représentatives des lois d'incidence font l'objet d'un lissage.

a) En pratique la reconstitution de la somme des expositions au risque

Pour chaque groupe d'assurés regroupés par âge de survenance, il s'agit de reconstituer la période d'exposition (durée écoulée entre le début d'exposition et la fin d'exposition) notée $expo(x)$.

La date de début d'exposition de l'assuré i et notée $d_d(i)$ représente la date la plus récente entre le 1er janvier de l'année N début d'observation et la date de début de contrat.

La date de fin d'exposition de l'assuré i et notée $d_f(i)$ est susceptible de représenter :

- Soit la date de survenance du premier arrêt de travail ou de la censure postérieure à la date de début d'exposition ;
- Soit en l'absence d'arrêt de travail :
    - La date de résiliation du contrat de travail, ou date de décès,
    - Ou la date de fin de surveillance : le 31 Mars 2019.

En pratique, la période d'observation est découpée en tranches annuelles :

Individu n°	Du 01/01/N au 31/12/N	Du 01/01/N+1 au 31/12/N+1	…	Du 01/01/2019 au 31/12/2019	Exposition
Individu i	$\text{expo}\ (N, i)$	$\text{expo}\ (N+1, i)$	…	$\text{expo}\ (2019, i)$	$\text{expo}(i) = \text{expo}(N, i) + \cdots + \text{expo}\ (2019, i)$
Individu $i+1$	$\text{expo}\ (N, i+1)$	$\text{expo}\ (N+1, i+1)$		$\text{expo}\ (2019, i+1)$	$\text{expo}(i+1)$

$$\text{expo}(n) = \frac{\min(31.12.n; date\ fin\ d'exposition) - \max(01.01.n; début\ de\ contrat)}{nombre\ de\ jours\ de\ l'année}$$

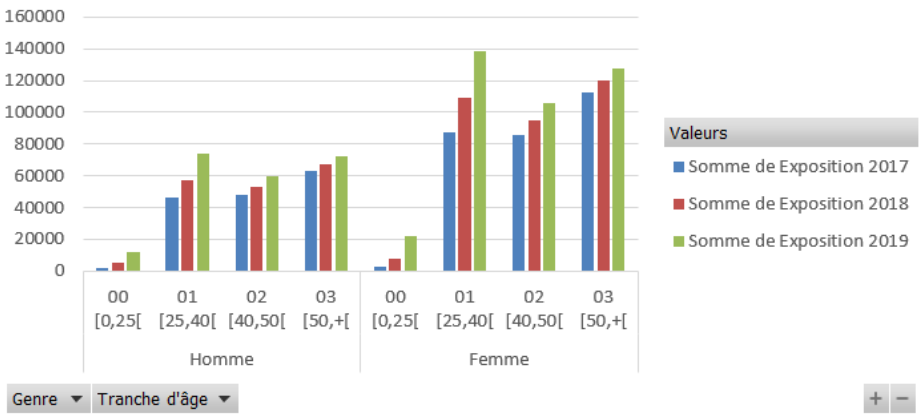

Figure 37. Répartition de l'exposition par tranche d'âge à la souscription

Sur une période de 3 ans, la tranche des 25-40 ans est la plus exposée au risque d'incapacité.

Cette étape a permis d'estimer la somme des expositions aux risques E_x de la relation $E(\widehat{\lambda_x}) = E\left(\frac{N_x}{E_x}\right) = \lambda_x$

Désormais, estimons le nombre d'entrées en incapacité N_x.

b) *En pratique la reconstitution du nombre d'entrées en incapacité*

Soit N_x le nombre d'entrées en incapacité pour les assurés d'âge [x,x+1[. Ce nombre se réécrit :

$N_x = \sum_{i=1}^{nombre\ d'individus d'âge\ x} X_i$. Les X_i représentent les variables aléatoires (incapable ou non).

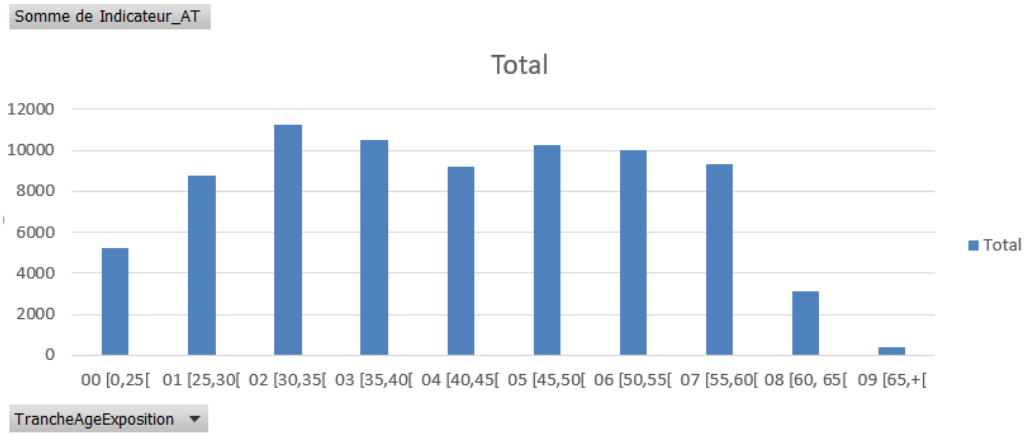

Figure 38. Répartition des nombres de sinistres par tranche d'âge à la survenance

c) *Adéquation de la loi d'incidence du 1er arrêt avec la loi de Poisson*

Il convient de vérifier $N_x \sim P(\lambda_x)$ avec x l'âge par un test d'adéquation avec la loi de Poisson (rappel de l'expression d'une loi de Poisson de paramètre λ : $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$).

Tout d'abord, les observations de l'échantillon sont classées dans deux classes distinctes :

- Classe 1 : individus sans période d'incapacité,
- Classe 2 : individus avec une période d'incapacité.

Ensuite, la seconde étape du test consiste à valider la taille n de l'échantillon :

$$n * p * (1-p) \geq 5 \Leftrightarrow n * \frac{1}{2} * \left(1 - \frac{1}{2}\right) \geq 5 \Leftrightarrow n \geq 20$$

Enfin, les statistiques de tests sont comparées ci-dessous (comparaison entre la réalité et la modélisation).

Soit T_x la statistique de test pour chaque classe d'âge x. T_x donne l'écart entre les effectifs théoriques et les effectifs observés : $T_x = \sum_i^{k-1=1} \frac{(EffectifRéel_{i,x}, -EffectifModélisé_{i,x})^2}{EffectifModélisé_{i,x}}$. k représente le nombre de classe.

Plus T_x est grand, plus l'écart est important. L'adéquation parfaite est matérialisée par $T = 0$.

La statistique $T_x$est comparée à la valeur du $\chi^2_{k-1,\alpha}$ avec :

- $k-1$, le nombre de degré de liberté,
- α la tolérance.

Soit l'hypothèse nulle suivante : $H_0 = N_x \sim Poisson(\lambda_x)$.

H_0 est validée si $P(T_x \leq \chi^2_{k-1,\alpha}) = 95\% \Leftrightarrow P(T_x > \chi^2_{k-1,\alpha}) = 5\%$ avec p_i.

$\chi^2_{k-1,\alpha}$ s'obtient par lecture de la table :

Loi de Khi-deux

Le tableau donne **x** tel que **P(K > x) = p**

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577

Figure 39. Table de la loi du χ^2

La valeur seuil est $\chi^2_{k-1,\alpha} = 3{,}8415$. H_0 est validée.

Il est possible d'observer graphiquement l'adéquation des lois pour les deux classes. Les bâtons bleus représentent les individus observés. Les bâtons rouges représentent les effectifs modélisés.

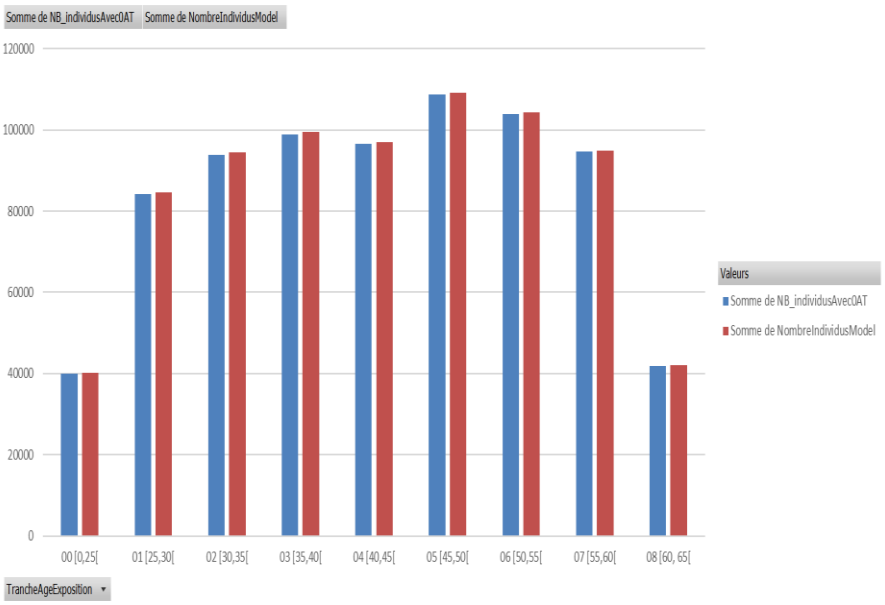

Figure 40. Distribution des individus sans période d'incapacité

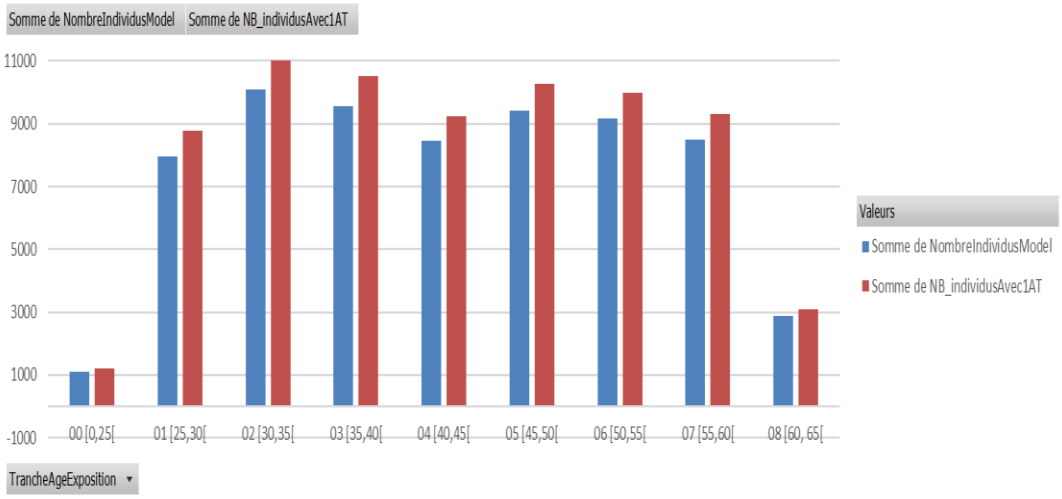

Figure 41. Répartition des individus avec un nombre de sinistre = 1

Les composantes de la loi sont reconstituées. Il est désormais possible de tracer la courbe des taux d'incidence.

d) La courbe des taux bruts d'incidence en incapacité

Les taux bruts sont affichés ci-dessous. D'après l'échantillon à disposition, les deux pics d'incidence sont observables pour les âges autour de 28 – 35 ans et 56 – 60 ans. Sur le premier pic, nous ne disposons pas suffisamment de variables pour expliquer ce pic. Le deuxième pic est susceptible de s'expliquer par la détérioration croissante avec l'âge des conditions physiques. Après ce pic, une décroissance s'observe. Les entreprises sont susceptibles d'encourager leurs salariés proches de la retraite à solder leurs congés payés,

leurs comptes épargne temps et à entrer dans un dispositif de retraite anticipée. Cela peut jouer en faveur d'une réduction du nombre d'arrêts de travail pour les personnes plus âgées.

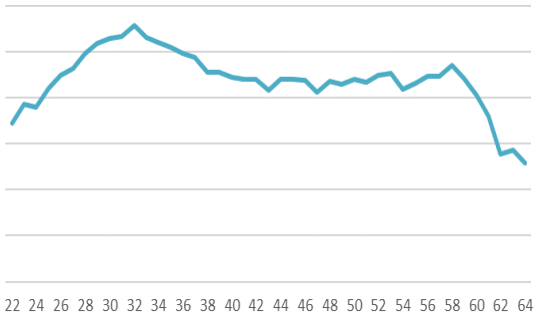

Figure 42. Taux bruts d'incidence par âge

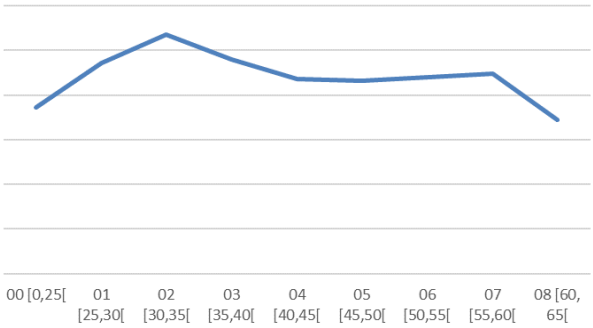

Figure 43. Taux bruts d'incidence par tranche d'âge

Ces courbes présentent des irrégularités qui ne correspondent pas nécessairement à un phénomène réel. Il est décidé de lisser les courbes.

e) Lissage de la courbe des taux d'incidence en incapacité

La méthode des moindres carrées est retenue pour lisser la courbe des taux bruts. Cette méthode implique de déterminer un polynôme tel que $E(N_x) = \sum_{i=0}^{d} c_i x^i$ avec :

- c_i le coefficient d'ordre i du polynôme
- d le degré du polynôme
- x l'âge.

Les coefficients s'obtiennent par minimisation de l'écart entre la courbe réelle et la courbe « moindre carrée ».

$$\min \sum_{i=\text{âge minimum}}^{\text{âge maximum}} \left(Effectifs_i - \sum_{j=0}^{d} c_j x^j \right)^2$$

La minimisation s'obtient par annulation des dérivées partielles par rapport aux coefficients.

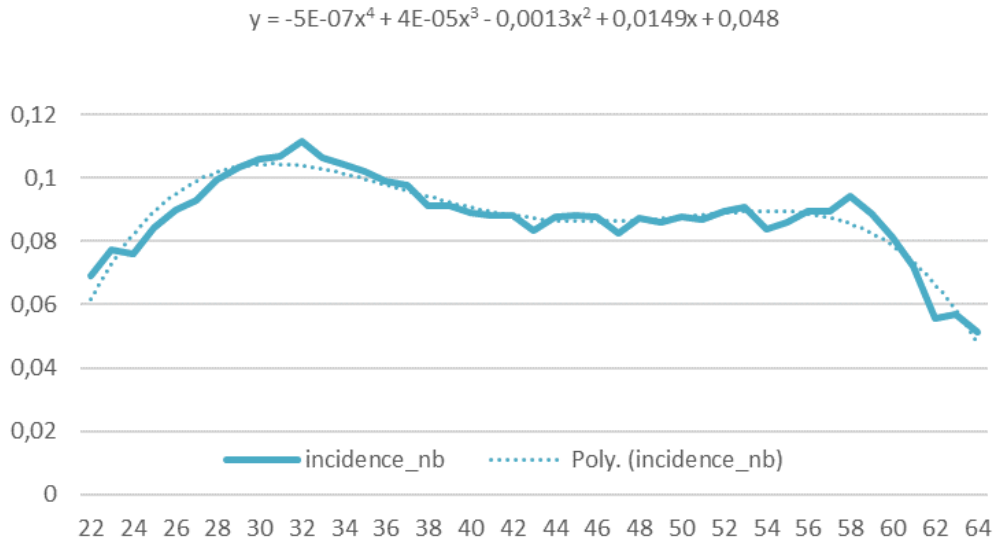

Figure 44. Courbes des taux brutes et lissées avec d=4

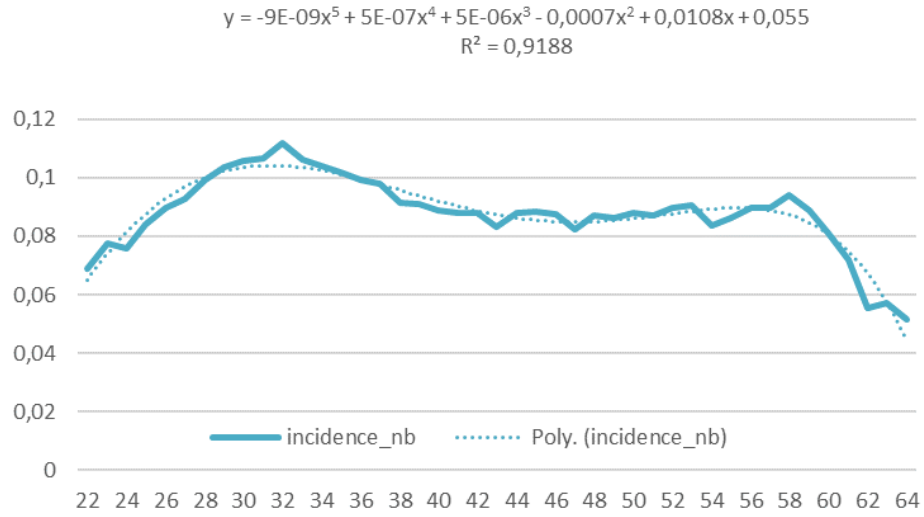

Figure 45. Courbes des taux brutes et lissées avec d = 5

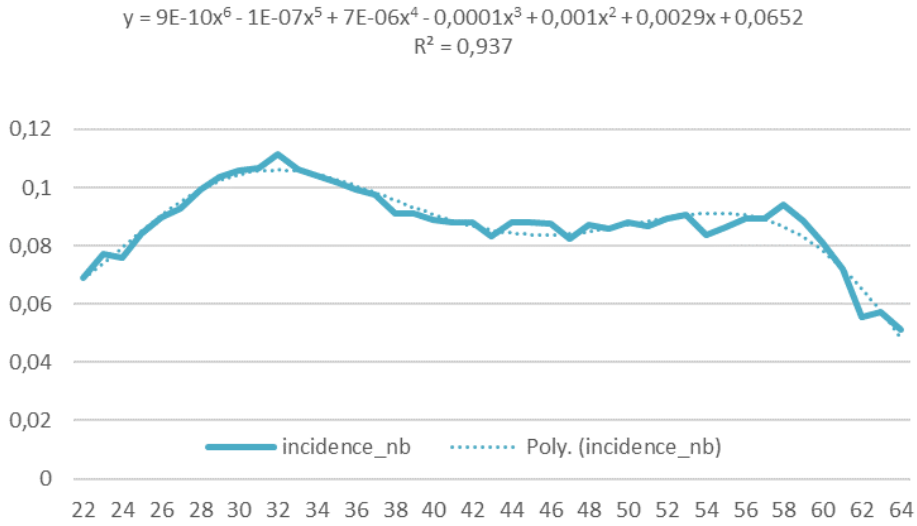

Figure 46. Courbes des taux brutes et lissées avec d = 6

6 est la valeur de d retenue.

$$y = 9E-10x6 - 1E-07x5 + 7E-06x4 - 0{,}0001x3 + 0{,}001x2 + 0{,}0029x + 0{,}0652$$

Par lecture graphique, ce polynôme semble être celui qui approche le plus fidèlement la courbe des taux bruts.

f) Limite sur le périmètre

La question du périmètre est indispensable. En effet, il est crucial d'être en mesure de reconstituer une loi d'incidence en regard des garanties offertes par les contrats d'assurance.

Or compte-tenu des données mises à disposition, le champ « motif de l'arrêt » (référence de la rubrique S21.G00.60.001) de la DSN n'est pas disponible dans les données mises à disposition de l'OA.

Toutefois, après échange avec les experts de l'organisme assureur, l'échantillon de données mise à disposition exclue bien les observations dont le motif d'arrêts auraient les valeurs suivantes :

- 01 – maladie,
- 02 – maternité,
- 03 - paternité / accueil de l'enfant,
- 07 - femme enceinte dispensée de travail.

En effet, ces natures d'arrêt ne sont pas couvertes par l'organisme assureur.

Par ailleurs et toujours sur des problématiques de périmètre, l'approche tarifaire « a priori » est privilégiée pour des contrats standards et petites entreprises. Les grands comptes souscrivent généralement des produits sur-mesure avec une approche tarifaire spécifique. Pour éviter des biais, il aurait fallu retirer les grands comptes de l'étude.

g) Autres limites de la loi d'incidence

L'irrégularité de la distribution de l'exposition des individus par exercice ne permet pas de construire une loi d'incidence représentative de l'incidence générale.

L'exposition sur l'année 2019 est sur-représentée par rapport aux années antérieures.

Les données sont tronquées à gauche : des périodes d'incapacité sont susceptibles d'être enregistrées dans les DSN des mois antérieurs à janvier 2020. Cela peut conduire à une sous-estimation du nombre d'incapacités.

Le travail serait à affiner avec une profondeur d'historique de données DSN plus importante.

Malgré les limites sur les données à disposition pour établir une loi d'incidence, nous proposons tout de même de poursuivre en nous basant sur la loi d'incidence spécifique au contexte de « restriction de données ». Elle permet tout de même d'observer des phénomènes intéressants (pics d'incidence sur les tranches d'âge 28-35 ans et 56-60 ans, décroissement des taux entre ces pics…).

Cette partie a permis de reconstituer les lois d'incidence et de maintien à partir des données d'expérience.

Nous proposons de les utiliser pour la tarification de l'incapacité. En particulier, nous nous intéresserons à l'opportunité d'introduire la notion de classe de risque pour segmenter le portefeuille d'entreprises en vue de la tarification.

5. ETUDE DE L'OPPORTUNITE D'INTRODUIRE LA NOTION DE « CLASSE DE RISQUE » POUR LA TARIFICATION DES GARANTIES INCAPACITE

5.1. LE MODELE DE TARIFICATION

5.1.1) Rappels des principes généraux

Les supports pédagogiques (C. Izart, 2018) rappellent que trois éléments principaux sont utilisés pour le calcul des tarifs :

- Un taux technique,
- Un facteur lié à l'aléa sur le maintien en incapacité,
- Des chargements.

Les deux premiers éléments permettent de calculer le montant que l'assureur doit prévoir en date de début du contrat pour honorer une garantie donnée. Le principe de base : à la souscription du contrat, les valeurs actuelles probables des engagements respectifs de l'assureur et de l'assuré sont égales.

$$Valeur\ actuelle\ probable(assuré) = Valeur\ actuelle\ probable\ (assureur)$$

La valeur actuelle probable des engagements futurs de l'assureur envers le bénéficiaire est aussi appelée prime pure unique $\pi_{pure\ unique}$.

Quant aux chargements, ils comprennent les coûts commerciaux et frais de gestion supportés par l'assureur – hors champs d'analyse du mémoire.

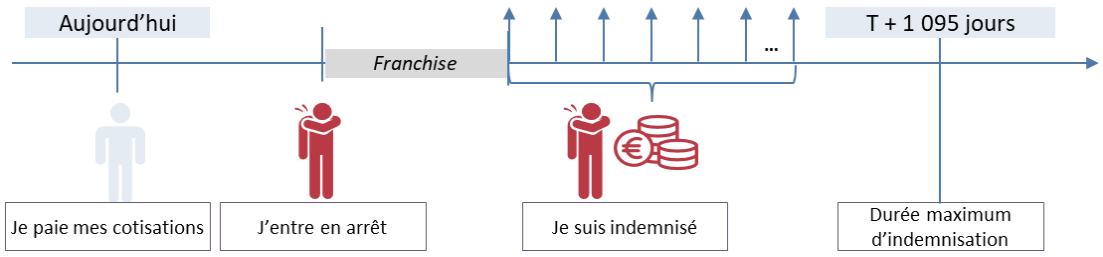

Figure 47. Schéma de l'engagement de l'assureur à honorer la garantie incapacité

5.1.2) Modèle « fréquence * coûts » pour la tarification de la garantie incapacité

Les supports pédagogiques du CEA et d'un mémoire d'actuariat (Cyprien Herbreteau, 2017) ont permis d'expliciter la formule de tarification de notre modèle.

La modélisation retenue pour la tarification de la garantie incapacité est la suivante :

$$\pi_{pure}(x, f, csp) = p_{incidence}(x, f, csp) * E(S(x))$$

Avec :

- $p_{incidence}(x, f, csp)$: la probabilité pour un individu d'âge x de catégorie socioprofessionnelle csp et ayant souscrit à une garantie de franchise f d'entrer en indemnisation au moins une fois dans l'année à venir,
- $E(S(x))$: l'espérance de la charge probable de l'arrêt de travail de cet individu.

La charge probable s'écrit de la manière suivante :

$$E\big(S(x, f, csp)\big) = \sum_{t=1}^{36-f} \frac{l_{x,t+f}}{l_{x,0}} * \frac{R_{mensuelle}}{(1+i)^{\frac{t}{12}}}$$

Avec :

- x correspond à l'âge de l'individu au moment de la survenance. Dans le mémoire, la borne x est comprise entre 22 et 60 ans,
- t correspond au temps par mois,
- i représente le taux technique annuel,
- $l_{x,f}$: le nombre d'individus d'âge x en incapacité au moment de la fin de la franchise f,
- $l_{x,t}$: le nombre d'individus d'âge x en incapacité à l'instant t,
- $R_{mensuelle}$ représente le montant mensuel de prestations pour une indemnité journalière fixée à 1€. 30,5 jours est l'hypothèse retenue pour la durée d'un mois en moyenne.

Les autres hypothèses retenues sont :

- Un démarrage des arrêts de travail en début d'âge à la survenance,
- Une indemnisation mensuelle à terme échu.

La formule générale du tarif socle est donc :

$$\pi_{pure}(x, f, csp) = p_{incidence}(x, f, csp) * \sum_{t=1}^{36-f} \frac{l_{x,t+f}}{l_{x,0}} * \frac{R_{mensuelle}}{(1+i)^{\frac{t}{12}}}$$

Cette formule contient plusieurs approximations. Sur la $VAP(Assuré)$, π_{pure} représente la prime unique pure annuelle que paie un assuré à l'assureur. Dans la pratique, l'assurée paie sa cotisation mensuellement. La formule exacte de la $VAP(Assuré)$ en tenant compte de l'aléa de mortalité sur une année $\pi_{pure} = \pi_{pure\ mensuelle} * \sum_{mois=1}^{12} \frac{l_{x+m}}{l_x} * \frac{1}{(1+i)^{\frac{m}{12}}}$.

L'approximation suivante est réalisée : $\pi_{pure} = 12 * \pi_{pure\ mensuelle} = 365{,}25 * \pi_{pure\ journalière}$.

Par ailleurs, il faudra également prendre en compte la date exacte d'incidence en incapacité. L'approximation réalisée dans ce modèle induit que l'ensemble des incidences se déroulent en début d'année. Ce qui ne correspond pas à la réalité. Il est courant de faire l'hypothèse d'une survenance en milieu de période.

5.1.3) Le modèle opérationnel de la tarification utilisé actuellement

a) La formule

La formule de tarification du paragraphe précédent nécessite de connaître la loi d'incidence et de maintien pour chaque âge à la survenance. En pratique, la démarche suivante est adoptée :

- Détermination d'un tarif central à partir des lois d'incidence et de maintien pour un individu d'âge central $x_{central}$,
- Multiplication du tarif central par des correctifs de plusieurs natures.

La formule de tarification est modifiée pour tenir compte des répartitions par genre et par CSP.

Formule Opérationnelle actuelle	$\pi_{pure}(x, f, csp, genre) = \pi_{pure}(x_{central}, f, csp, genre_{unisex}) * Correctif_{age}(x) * Correctif_{genre}(x)$
Formule Théorique (cf. § précédent)	$\pi_{pure}(x, f, csp, genre) = p_{incidence}(x, f, csp, genre) * \sum_{t=1}^{36-f} \frac{l_{x,t+f}}{l_{x,0}} * \frac{R_{mensuelle}}{(1+i)^{\frac{t}{12}}}$

b) Application opérationnelle

Pour établir le barème de tarif, la démarche se déroule en quatre étapes :

1) Reconstituer la loi d'incidence pour chaque tranche d'âge. Le résultat est formalisé sous forme de tableau :

Tranche d'âge $[x, x+k[$	CSP : Cadre			CSP : Employé			CSP : Ouvrier		
Franchise (en JOURS)	unisexe	Homme	Femme	unisexe	Homme	Femme	unisexe	Homme	Femme
0	…	…	…	…	…	…	…	…	…
1	…	…	…	…	…	…	…	…	…
f	…	p	…	…	…	…	…	…	…
1050	…	…	…	…	…	…	…	…	…

Le tableau se lit de la manière suivante : p représente la probabilité de tomber en incapacité pour un homme, cadre ayant souscrit à une garantie dont la franchise est égale à f.

2) Reconstituer la loi de maintien pour l'âge central $x_{central}$

	Ancienneté par mois				
Age à la survenance (en année)	0	1	2	…	36
20	…	…	…	…	…
22	100	80	70		1
z	…	l	…	…	…
60	…	…	…	…	…

Dans ce tableau, l représente le nombre d'individus d'âge z ayant dépassé un mois dans l'état incapacité. Pour chaque âge, le coefficient de durée est calculé $\sum_{t=1}^{36-f} \frac{l_{x,t+f}}{l_{x,0}}$

3) Reconstituer le barème de tarif pour $x = x_{central}$

	CSP de référence			CSP 1			…
Franchise (en jours)	Unisexe	Homme	Femme	Unisexe	Homme	Femme	…
0	π	$\pi * correctif_{homme}(x)$	$\pi * correctif_{femme}(x)$	$\pi * correctif_{csp1}(x)$	$\pi * correctif_{homme}(x) * correctif_{csp1}(x)$	$\pi * correctif_{femme}(x) * correctif_{csp1}(x)$	…
1	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
1050	…	…	…	…	…	…	…

4) Appliquer le correctif d'âge au tarif établi par le barème.

Le modèle opérationnel actuel de tarification tient compte uniquement des facteurs d'âge, franchise et catégorie socioprofessionnelle. Ce modèle pourrait évoluer pour prendre en compte de nouveaux facteurs.

5.1.4) Le modèle cible de la tarification

Intuitivement, de nouveaux facteurs explicatifs de la sinistralité sont susceptibles d'intervenir notamment pour tenir compte de la classe de risque de l'entreprise.

Pour cela, il est proposé d'introduire un correctif de classe de risque noté $correctif_{risque}$.

La formule de tarification cible pourrait intégrer ce correctif :

$$\pi_{pure}(x, f, csp, classe) = \pi_{pure}(x_{central}, f, csp) * Correctif_{age}(x) * Correctif_{risque}(classe)$$

A partir de maintenant, l'objectif est donc de déterminer ces classes de risque à partir de variables disponibles dans la base de données.

Pour cela, le principe est de révéler le profil type de chaque classe de risque par recours aux méthodes de Machine Learning non supervisées. Les principes sont décrits dans les paragraphes du chapitre qui suit.

5.2. DETERMINATION DES CLASSES DE RISQUES PAR MACHINE LEARNING

Pour déterminer les classes de risques homogènes, il est nécessaire de segmenter le portefeuille à disposition.

5.2.1) Exposé du besoin de segmentation

Le but est de tarifer selon le profil de risque de l'entreprise. En fonction de son profil, un correctif tarifaire sera appliqué. Pour cela, des classes de risques sont à déterminer à partir des données à disposition. Dans notre contexte, une classe de risque se définit comme un groupement de secteurs d'activité dont les caractéristiques sont homogènes.

La tentative de créer des groupes de secteurs homogènes se base sur l'analyse de la démographie de ces secteurs. En première approche, cela passe par segmenter les secteurs par les variables suivantes :

- Répartition moyenne homme / femme,
- CSP,
- Age des employés du secteur.

Pour réaliser cette segmentation, le recours au Machine Learning semble bénéfique. Il convient de définir ce que signifie le terme « Machine Learning ».

5.2.2) Définition et intérêt du recours au machine learning

Un historique de données de plus en plus profond, une capacité de stockage croissante, une puissance machine pour traiter de grands volumes de données : tels sont les prérequis pour recourir à l'intelligence artificielle et tout particulièrement au « Machine learning ». Il s'agit d'un terme anglais qui signifie « apprentissage automatique ». Il désigne un concept impliquant un processus d'analyse de données et d'implémentation d'algorithmes de résolution de problèmes nécessitant une puissance calculatoire.

Ces algorithmes mettent en exergue des corrélations, des phénomènes et permettent de prédire des situations sans avoir d'a priori sur la structure des données. Le but est donc l'implémentation d'un algorithme ou « machine » qui va apprendre à prédire une situation à partir des données à disposition.

Mathématiquement (cf. supports pédagogiques de Claire Boyer et Olivier Lopez, 2019), cela revient à trouver la meilleure machine m tel que : $Y = m(X)$ avec Y le vecteur des données à prédire et X la matrice des variables explicatives. La machine m permet d'estimer la réalité de Y. Pour chaque machine :

- La performance est mesurée par la fonction de risque R qui représente la distance mathématique entre le Yet son estimation $R(X) = \boldsymbol{E}(l(Y, m(X)))$**.** l représente la fonction de coût pertinente pour la machine,
- La complexité du modèle dépend des hyper paramètres (ex : profondeur d'un arbre de décision, nombre de plus proches voisins, …). Il est nécessaire de faire un compromis entre :
    - Modèle trop simple (biais élevé / variance faible). Cela se traduit par une mauvaise qualité de prédiction,

- Modèle trop complexe (biais faible / variance élevée). Cela se traduit par le phénomène de surapprentissage (ie. m fait correspondre quasi parfaitement l'estimation et la réalité sur la base de données mais prédit n'importe quoi sur de nouvelles observations).

Il s'agit en général donc trouver un estimateur qui :

- Relie correctement Yavec X,
- Prédit les éventuelles nouvelles observations de la base de données.

Il existe deux grands scénarios d'apprentissage :

- Apprentissage supervisé : la base de données contient pour chacune des observations de la matrice X la réponse Y.
- Apprentissage non supervisé : la base de données ne contient pas la réponse Y. Il s'agit justement de trouver Y. Il s'agit typiquement des problèmes de partitionnement des données en vue d'obtenir des catégories homogènes.

5.2.3) Principe du partitionnement des données par la méthode K-Means

Pour construire les profils de risque des entreprises, il est envisagé de partitionner le portefeuille avec des méthodes de clustering.

Le partitionnement (ou « clustering » en anglais) consiste à élaborer des groupes homogènes dans un vecteur de données d'apprentissage. Chaque groupe de ce vecteur est appelé « cluster » en anglais ou « classe » en français.

La méthode des « k-means » permet de classer chaque observation dans une seule et unique classe.

Le principe de l'algorithme est de partitionner un ensemble de points $(X_1, X_2, \dots X_N)$ en k classes (cf. support pédagogique de Maxime Sangnier).

Chaque classe répond à une contrainte d'optimisation : minimiser la distance au carré entre les points à l'intérieur de chaque classe. La contrainte de minimisation s'écrit de la manière suivante :

$$\underset{\{C_1,C_2,\dots,C_k\}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{X_j \in C_i} \left|\left|X_j - \mu_i\right|\right|^2$$

Avec :

- k le nombre total de classes à former et inférieur à N,
- i allant de 1 à k correspondant au numéro de la classe,
- $X_j \in C_i$ correspondant aux points regroupés dans la classe $C_{k\,i}$,
- μ_i est le centre de gravité de la classe C_i

L'algorithme fonctionne de la manière suivante :

Etape 1 – Saisie par l'utilisateur des données en entrée :

- k : le nombre de classes,
- T : le nombre d'itérations de l'algorithme,
- $(X_1, X_2, \dots X_N)$ les N observations à partitionner en $(C_1, \dots, C_k)$ classes.

Figure 48. Graphe du nuage d'observations à séparer

Etape 2 – Initialisation des k centres de gravité $(\hat{\mu}_1^{(0)}, \dots, \hat{\mu}_k^{(0)})$ par sélection aléatoire.

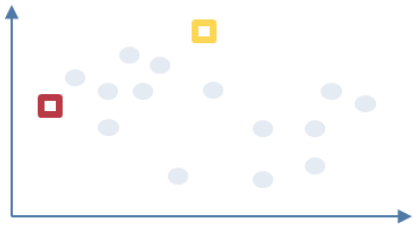

Figure 49. Initialisation de k centres de gravité (k=2)

Etape 3 – Minimisation de la distance « centre de gravité – position de l'observation » sous conditions.

Pour chaque itération t jusqu'à T, répéter les étapes suivantes :

- Affecter chaque observation à la classe C_i la plus proche en utilisant le principe de la partition de Voronoï : $C_i^{(t+1)} := \{X_1, \dots, X_N\}^{(t)} \cap C_i^{(t)}$
- Mettre à jour le centre de gravité de chaque :

$$\mu_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} * \sum_{X_j \in C_i^{(t)}} X_j$$

Illustration avec $t = 1$:

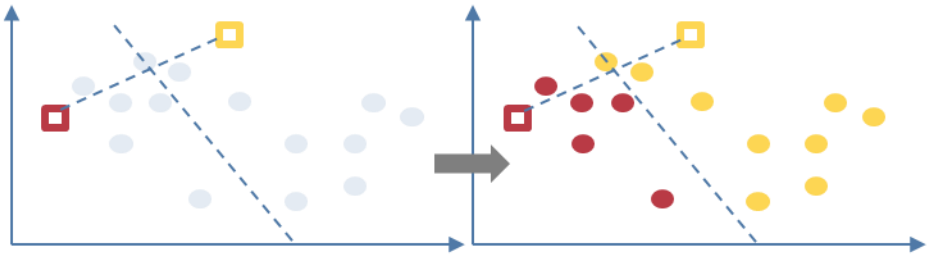

Figure 50. Illustration de l'affectation des observations à chacun des k clusters

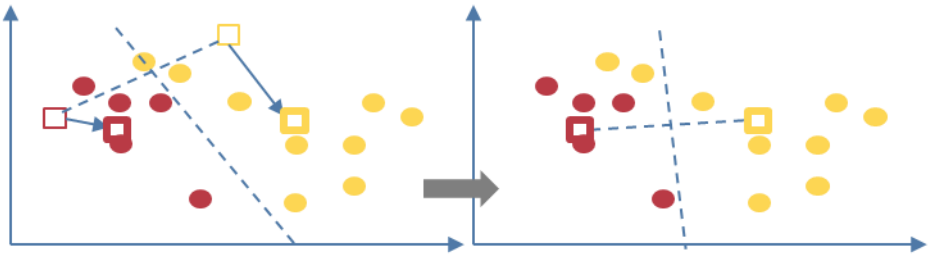

Figure 51. Illustration de la mise à jour des k centres de gravité

Illustration avec $t = 2$:

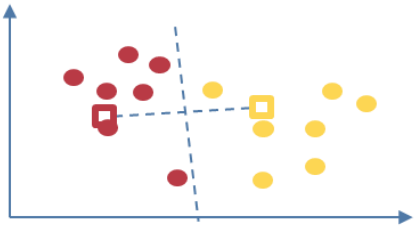

Figure 52. Illustration de l'affectation des observations à chacune des k classes (t=2)

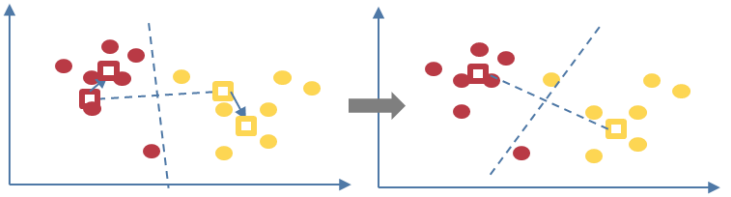

Figure 53. Illustration de la mise à jour des k centres de gravité (t=2)

Illustration pour la dernière itération $t = T$:

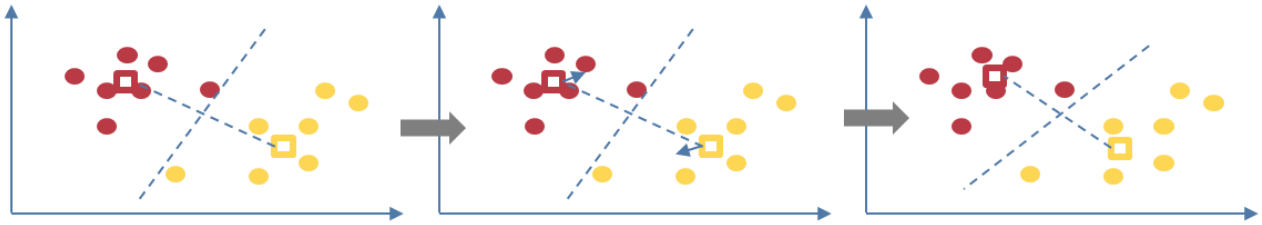

Figure 54. Illustration de l'étape 3 (t=T)

Etape 4 – l'algorithme sort les k classes $(C_1, \dots, C_k)$.

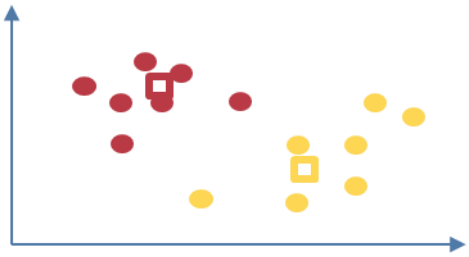

Figure 55. Illustration du résultat final du clustering

Une fois les grands principes du clustering par K-Means expliqués, nous passons à la mise en application avec les données du portefeuille mises à disposition par l'organisme assureur.

5.2.4) Mise en application : avec des données endogènes

Le processus de constitution des groupes homogènes s'effectue en cinq étapes :

- Sélectionner les variables présumées discriminantes pour la constitution des groupes,
- Partitionner les secteurs d'activité (code APET) en K classes,
- Identifier les variables les plus discriminantes,
- Analyser la distribution des taux d'incidence intra groupe et attribuer un niveau de risque à chaque classe.

a) Sélectionner les variables

Pour réaliser le clustering des secteurs d'activités (APET), les variables retenues sont :

- La proportion de femmes (variable intitulée $NombreFemme$),
- La proportion d'hommes. Cette variable est corrélée à la variable « nombre de femmes »,
- La proportion d'individus dans la tranche des âges avant 25 ans,
- La proportion d'individus dans la tranche des âges de 25 à 40 ans,
- La proportion d'individus dans la tranche des âges de 40 à 50 ans,
- La proportion d'individus dans la tranche des âges après 50 ans. Cette variable est corrélée aux autres variables « Tranche des âges… »
- La proportion d'individus dans la CSP 1 (variable intitulée NB_CSP_1),
- La proportion d'individus dans la CSP 2,
- …
- La proportion d'individus dans la CSP 10. Cette variable est corrélée avec les autres variables « Nombre d'individus dans la CSP … »

Avant de lancer l'algorithme de partitionnement, les variables redondantes avec une autre dans la base sont retirées. En effet, elles n'apportent pas d'informations supplémentaires.

Pour segmenter, nous prenons le parti de ne pas créer un nombre trop important de classes de risque. Nous allons tester le partitionnement avec des valeurs de $k = 3\ puis\ 4$ avant de conclure sur le nombre de classes de risque à éventuellement constituer.

b) *Partitionner les secteurs d'activité en K classes*

L'algorithme $kmeans$ est lancé pour $k = 3$.

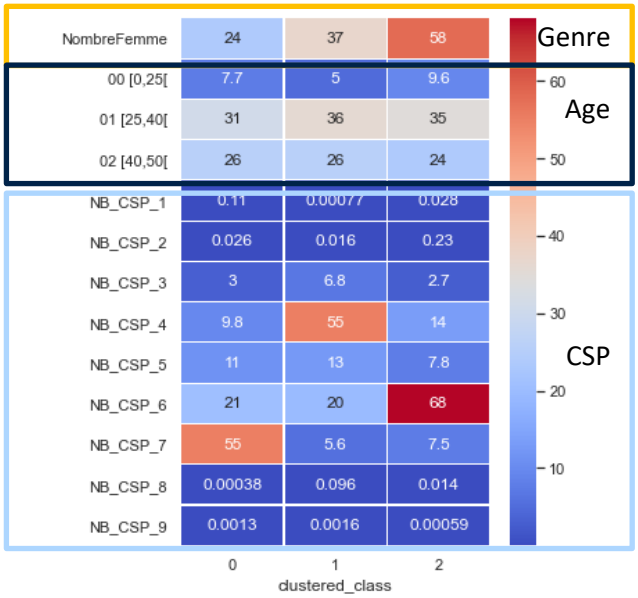

Figure 56. Partitionnement en 3 classes

Par analyse univariée des données de chaque classe, il est possible de mettre en exergue des caractéristiques de chaque classe.

La répartition homme-femme semble être un facteur clivant. La proportion de femme est plus importante dans la classe 2.

La répartition par tranche d'âge ne semble pas influer sur la constitution des classes. Chaque classe a sensiblement la même répartition.

Sur le groupe de variables des CSP, 3 CSP ressortent clairement. Il s'agit des CSP, 7, 4 et 6.

En synthèse, les critères clivant pour constituer des groupes de secteurs sont la CSP et la répartition homme-femme.

Ainsi, chaque groupe de secteurs a les caractéristiques suivantes :

- La classe 0 est constituée essentiellement d'individus de la CSP 7,
- La classe 1 est composée majoritairement d'individus de la CSP 4,
- La classe 2 est constituée essentiellement de femmes et d'individus de la CSP 6.

Pour étudier l'effet du nombre de cluster, l'algorithme est lancé de nouveau avec un nombre de clusters $k = 4$

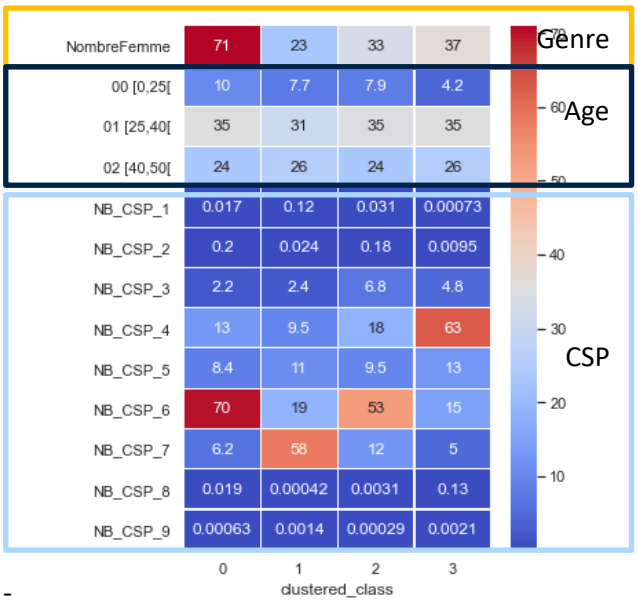

Figure 57. Partitionnement en 4 clusters

Par rapport au partitionnement en 3 clusters, celui en $k = 4$ ne semble pas apporter d'information supplémentaire :

- La classe $0_{k=4}$ a le même profil que la classe $2_{k=3}$: classe constituée essentiellement de femmes et de la CSP 6,
- La classe $1_{k=4}$a le même profil que la classe $0_{k=3}$: classe constituée majoritairement d'individus de la CSP 7,
- La classe $2_{k=4}$ semble être un sous-ensemble de la classe $2_{k=3}$ (classe constituée de la CSP 6).
- La classe $3_{k=4}$ a le même profil que la classe $1_{k=3}$: classe constituée d'individus appartenant à la CSP 4.

L'ajout d'un 4ème cluster n'a pas permis d'affiner la classification des secteurs d'activité. De ce fait, nous retenons la classification proposée avec $k = 3$.

c) Identifier les variables les plus discriminantes

Pour confirmer que le facteur d'âge n'intervient pas dans le partitionnement, l'algorithme est relancé sans ce facteur.

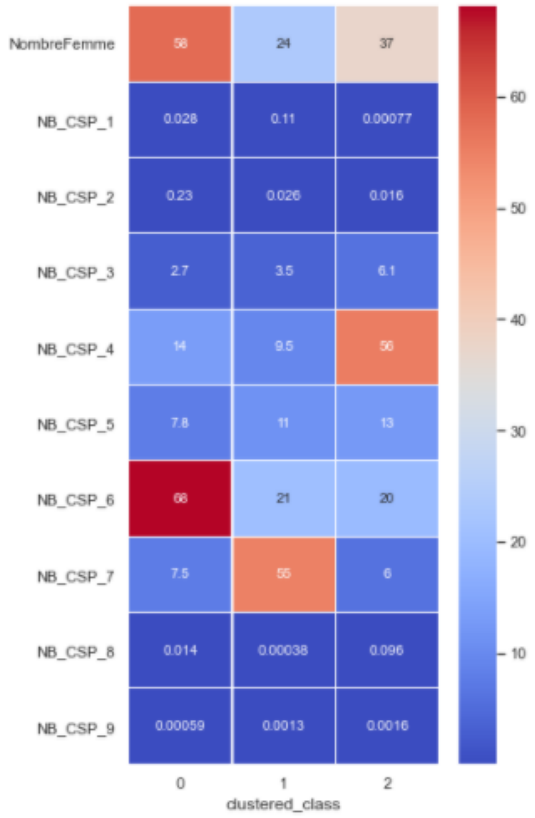

Figure 58. Clustering sans le facteur âge

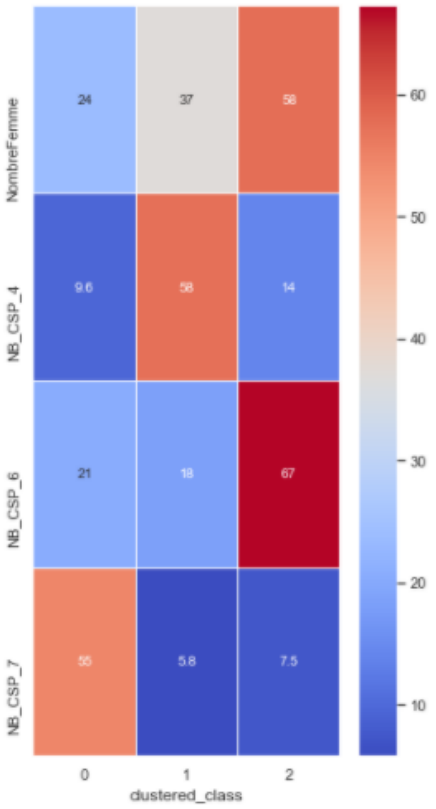

Figure 59. Clustering sans le facteur d'âge et avec uniquement les CSP pertinentes

Les cartes de chaleur mettent en exergue les variables les plus utilisées pour catégoriser. Il s'agit des variables représentant la proportion en termes de :

- Nombre de femmes,
- Nombre d'individus dans la CSP_6,
- Nombre d'individus dans la CSP_4,
- Nombre d'individus dans la CSP_7.

Le facteur d'âge ne semble pas être un facteur discriminant dans le partitionnement des groupes de secteurs d'activité.

En cohérence avec l'étude du portefeuille à disposition, il s'avère que la catégorie CSP_6 est sur-représentée dans l'échantillon.

Il convient toutefois de vérifier les liaisons entre les variables qui déterminent lesdites classes. Pour cela, le coefficient de Pearson permet de matérialiser le degré de corrélation. Son expression est la suivante :

$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$ avec :

$$Cov(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x = x_i \cap y = y_j) - E(x)E(y)$$

$$\sigma_x = \sqrt{Var(x)} = \sqrt{Cov(x,x)} = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \overline{x_i})^2}$$

Si $\rho = 0$ alors les variables x et y ne sont pas corrélées.

Si $\rho = 1$ alors les variables x et y sont corrélées.

Si $\rho = -1$ alors les variables x et y sont anti corrélées. C'est-à-dire, lorsqu'une des variables évolue dans un sens l'autre évolue en sens opposé.

Pour visualiser les corrélations entre les variables, la matrice $Matcorr$ de corrélation est construite.

$$Matcorr_{i,j} = \rho(x_i, y_j)$$

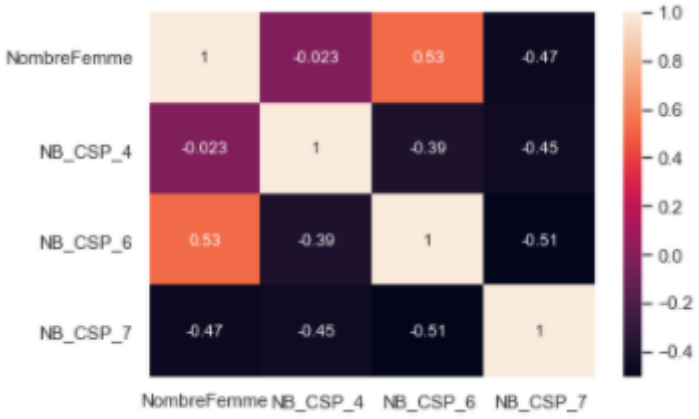

Figure 60. Matrice de corrélation des variables explicatives du 3-partitionnement

Une corrélation existe bien entre les facteurs « répartition homme-femme » et csp_6 ». Ce qui est susceptible de limiter le pouvoir discriminant de ces deux facteurs dans la détermination de classes de risque.

A partir des données endogènes, trois classes de risques sont retenues.

Plus fonctionnellement, les classes de secteurs d'activité sont composées de la manière suivantes :

- Un premier groupe constitué majoritairement de femmes, employés administratif d'entreprise, de commerce ou agent de service,
- Un second groupe regroupant les cadres,
- Un troisième représenté par des ouvriers qualifiés et non qualifiés.

Suite à l'obtention de cette proposition de segmentation, nous souhaitons étudier l'opportunité d'affiner le partitionnement avec d'autres données.

5.2.5) Mise en application : ajout de variable exogène

La question s'est posée d'affiner le partitionnement par ajout d'autres valeurs.

L'accès aux données de l'OA étant limité, l'échantillon à disposition n'a pas pu être enrichi d'autres données de la DSN. Les données susceptibles d'être explorées sont : l'ancienneté moyenne de l'entreprise, l'effectif dans l'entreprise, etc.

Toutefois, il existe des sources de données externes en libre-accès.

a) Les Open Data

Les open data ou « données ouvertes » sont des données consultables librement sur internet. Elles sont diffusées par des organisations publiques ou privées. Généralement, ces open data sont structurées et sont diffusées dans le respect des réglementations en vigueur en matière de protection des données personnelles, etc.

Tous les secteurs sont concernés par le mouvement de mise en commun des données.

En France, le gouvernement met à disposition de nombreuses données : l'agriculture et l'alimentation, la culture, l'économie et l'emploi, l'éducation et la recherche, le logement, le développement durable et l'énergie, la santé, le social, la société, les transports, le tourisme et territoires.

Figure 61. Logo du site data.gouv.fr

Dans le cadre de l'analyse du portefeuille, les données pertinentes à obtenir tournent autour des entreprises. Il s'agit de trouver des données permettant d'enrichir l'information sur les entreprises.

b) Cas 1 – Mise en application avec des données intuitivement intéressantes mais difficiles d'exploitation

Toujours dans l'objectif de segmenter les entreprises en plusieurs groupes, l'un des axes d'étude a été de s'interroger sur l'effet des conditions de travail. Pour cela, le site dares.travail-emploi.gouv.fr recense de nombreuses enquêtes.

Dans le cadre du mémoire, l'exploration des données « conditions de travail » s'est limitée à l'une des enquêtes réalisées par la DARES. Les résultats de cette enquête concernent des sujets autour de la reconnaissance, de l'évaluation du travail et du sentiment d'insécurité de l'emploi et des changements d'organisation. De nombreuses thématiques sont abordées :

Figure 62. Liste des thématiques de l'enquête du document

« Reconnaissance, insécurité et changements dans le travail » de la DARES

Pour permettre d'exploiter les données, celles-ci sont consignées dans des fichiers au format csv.

Chacune des 19 questions fait l'objet d'une restitution avec 11 axes d'analyse. Le document mentionne ces axes :

Les tableaux présentent les résultats :

1) pour l'ensemble des salariés (et par sexe)
2) par catégorie socioprofessionnelle
3) par catégorie socioprofessionnelle pour les hommes
4) par catégorie socioprofessionnelle pour les femmes
5) par âge
6) par âge pour les hommes
7) par âge pour les femmes
8) par secteur d'activité
9) par type d'employeur
10) par type d'employeur pour les hommes
11) par type d'employeur pour les femmes

Figure 63. Liste des axes de répartition des réponses

Tableau 2.8 • **Reconnaissance du travail (2), rémunération en 2016 selon le secteur d'activité (Naf rèv.2) de l'ENSEMBLE DES SALARIÉS**

En %

Secteur d'activité économique (Naf rév.2)	Effectifs en milliers	Proportion de salariés qui estiment que compte tenu du travail qu'ils réalisent, ils sont				
		Très bien payé(e)s	Bien payé(e)s	Normalement payé(e)s	Plutôt mal payé(e)s	Très mal payé(e)s
Agriculture, sylviculture et pêche	240	0,9	22,6	45,1	28,0	3,4
Fabrication aliments, boissons, tabac	603	4,3	17,5	42,0	33,3	2,9
Cokéfaction et raffinage	17	-	-	-	-	-
Fabrication d'équipements et de machines	478	4,5	23,1	58,7	12,8	0,9
Fabrication de matériels de transport	416	0,5	18,5	44,2	33,0	3,8
Fabrication d'autres produits industriels	1 574	0,5	22,5	44,7	27,1	5,2
Ind. extractives, énergie, eau, déchets, dépollution	397	2,4	17,0	50,5	25,0	5,1
Construction	1 332	3,0	23,7	45,0	25,2	3,1
Transports et entreposage	1 527	3,6	19,7	45,9	25,8	5,0
Hébergement et restauration	815	2,0	18,3	35,8	30,6	13,3
Information et communication	684	2,6	24,7	44,8	25,5	2,4
Activités financières et d'assurance	856	4,0	20,8	50,8	21,1	3,2
Activités immobilières	314	0,4	11,6	59,5	23,4	5,1
Act. scientifiques, techniques, services adm., soutien	2 212	1,2	18,2	47,5	27,1	5,9
Autres activités de services	1 332	2,1	15,2	45,2	30,3	7,1
Non renseigné	28	-	-	-	-	-
Adm. publique, enseignemnt, santé et social	7 528	1,2	12,2	41,4	37,4	7,7
Commerce et réparation d'automobiles et de motocycles	2 881	1,4	18,4	42,9	29,7	7,6
Ensemble	23 236	1,8	17,2	44,3	30,5	6,3

- Effectifs insuffisants.
Lecture : en 2016, 44,3 % des salariés estiment être normalement payés compte tenu du travail qu'ils réalisent.
Champ : ensemble des salariés ; France métropolitaine.
Source : Dares-Drees-DGAFP-Insee, enquêtes Conditions de travail.

Figure 64. Exemple de restitution de résultats de l'enquête

Pour faciliter leur manipulation, une base « concaténation des réponses » est construite à partir des résultats des 19 enquêtes restitués par secteur d'activité.

Secteur d'activité économique (Naf rèv.2)	Effectifs en milliers	Crainte_emploi	Chgt_emploi	Crainte_mutation_oui	Crainte_mutation_non	Retrouver_emploi_facile	Retrouver_emploi_difficile	Retrouver_emploi_sa	Sentiment_Abri_financier PDTD	Sentiment_Abri_financier PD	Sentiment_Abri_financier D	Sentiment_Abri_financier TAFD	Sentiment_Abri_financier NC	Jusqu_a_retraite_oui	Jusqu_a_retraite_non	Modif_environnement_travail_poste	Modif_travail_iqu
Agriculture, sylviculture et pêche	240,00	11,42	30,59	4,63	95,37	51,78	31,75	16,47	50,50	38,70	7,53	2,96	0,30	37,07	42,25	4,78	4,0
Fabrication aliments, boissons, tabac	603,00	24,29	33,67	14,93	85,07	39,39	48,62	11,99	57,51	26,33	12,53	2,51	1,12	43,14	44,81	18,03	21,!
Cokéfaction et raffinage	17,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,0
Fabrication d'équipements et de machines	478,00	28,20	29,89	27,65	72,35	32,41	53,75	13,84	58,63	27,55	12,26	1,30	0,26	36,13	43,34	22,71	21,!
Fabrication de matériels de transport	416,00	26,22	40,67	30,07	69,93	29,72	63,03	7,18	56,96	33,46	6,29	2,15	1,15	37,31	48,86	30,43	23,(
Fabrication d'autres produits industriels	1 574,00	30,26	35,35	18,13	81,87	36,25	52,43	11,32	60,98	27,47	8,84	2,07	0,63	33,11	39,63	17,82	18,(
Ind. extractives, énergie, eau, déchets, dépollution	397,00	22,23	35,05	23,80	76,20	37,14	52,46	10,40	65,81	23,68	8,82	0,67	1,00	34,63	42,77	18,44	17,(
Construction	1 332,00	31,80	33,06	10,03	89,97	49,48	42,15	8,37	61,78	21,15	11,06	4,61	1,41	50,85	44,14	8,76	7,8
Transports et entreposage	1 527,00	29,53	35,22	23,92	76,08	37,05	52,17	10,78	66,41	21,74	8,41	2,42	1,02	37,45	41,25	22,41	17,!
Hébergement et restauration	815,00	23,75	44,59	13,53	86,47	54,47	33,58	11,95	61,88	25,57	8,70	3,30	0,54	44,81	49,62	16,05	13,:
Information et communication	684,00	26,04	33,68	14,76	85,24	45,25	41,43	13,31	56,16	26,39	15,00	1,52	0,93	36,81	57,00	13,41	15,:
Activités financières et d'assurance	856,00	20,02	36,47	29,26	70,74	44,89	44,55	10,55	49,16	37,06	9,75	4,00	0,03	37,61	55,97	23,70	22,:
Activités immobilières	314,00	31,18	36,49	11,45	88,55	30,00	56,63	13,37	43,95	24,27	19,31	11,67	0,80	38,62	45,40	14,53	11,(
Act. scientifiques, techniques, services adm., soutien	2 212,00	26,76	36,25	16,88	83,12	40,97	46,57	12,47	56,48	29,22	8,49	2,56	3,25	38,46	47,74	14,94	10,(
Autres activités de services	1 332,00	31,69	31,48	7,96	92,04	41,03	45,52	13,45	59,05	21,40	11,46	3,38	4,72	35,80	35,99	8,79	5,4
Non renseigné	28,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,0
Adm. publique, enseignemnt, santé et social	7 528,00	19,98	29,22	22,75	77,25	38,20	50,67	11,11	58,23	26,07	10,77	3,13	1,79	40,76	38,51	15,45	11,:
Commerce et réparation d'automobiles et de motocycles	2 881,00	23,93	36,20	14,82	85,18	41,76	45,07	13,17	56,31	27,90	11,13	2,20	2,47	46,55	51,35	13,17	14,:
Ensemble	23 236,00	24,55	33,36	18,77	81,23	40,26	48,12	11,61	58,38	26,49	10,42	2,92	1,79	40,45	43,71	15,63	13,:

Figure 65. Concaténation des variables de réponse

Pour rappel, la base de données des individus du portefeuille est nominative (provenance DSN). Elle contient l'information du code de l'activité principale de l'entreprise dans laquelle l'individu est embauché.

La base de données « individus » est ensuite complétée de la base « concaténation des réponses » par jointure sur le secteur d'activité. Cette jointure fait intervenir une matrice de correspondance entre le code APET et le secteur d'activité.

APET	Secteur_activité_OPENDATA
0111Z	Agriculture, sylviculture et pêche
0112Z	Agriculture, sylviculture et pêche
0113Z	Agriculture, sylviculture et pêche
0114Z	Agriculture, sylviculture et pêche
0115Z	Agriculture, sylviculture et pêche
0116Z	Agriculture, sylviculture et pêche
0119Z	Agriculture, sylviculture et pêche
0121Z	Agriculture, sylviculture et pêche
0122Z	Agriculture, sylviculture et pêche
0123Z	Agriculture, sylviculture et pêche
0124Z	Agriculture, sylviculture et pêche
0125Z	Agriculture, sylviculture et pêche
0126Z	Agriculture, sylviculture et pêche
0127Z	Agriculture, sylviculture et pêche
0128Z	Agriculture, sylviculture et pêche

Figure 66. Extrait de la matrice de correspondance "CODE APET" - Secteur d'activité

A l'instar de ce qui est réalisé à l'étape « Sélectionner les variables » du chapitre « mise en application avec des données endogènes », les variables clairement corrélées entre elles sont retirées.

L'algorithme de clustering est relancé et les résultats ci-dessous sont obtenus :

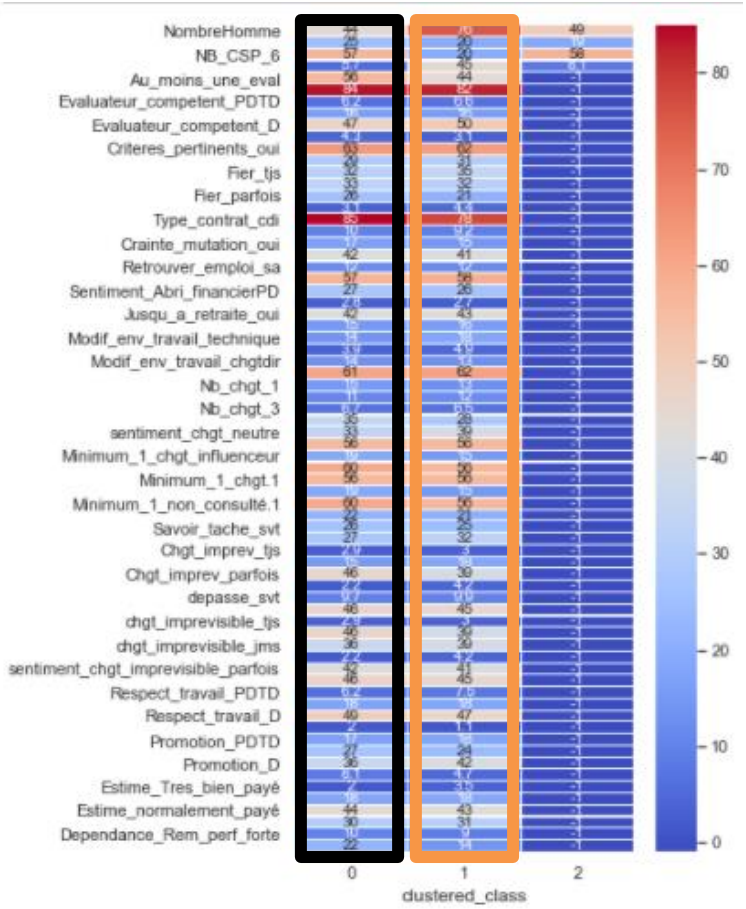

Figure 67. Partitionnement des codes APET avec les données endogènes et OPEN data

Par analyse des variables issues de l'open data, les classes 0 et 1 semblent avoir le même profil.

L'ajout des données open avec la granularité « secteur d'activité » ne nous a pas permis d'affiner la classification.

Pour aller au-delà :

- Il pourrait être envisagé de recourir à des combinaisons de variables pour mieux faire ressortir un éventuel phénomène,
- D'autres données open (auxquelles nous n'avons pas pensé) auraient pu être utilisées.

Nous ne tenons pas compte de ces variables pour le modèle.

c) Cas 2 – Mise en application des données « entreprise »

Pour enrichir l'échantillon de données entreprise à disposition, il semble intéressant d'explorer les données en lien avec le site des données des Greffes des Tribunaux de Commerce. Le site internet Datainfogreffe est justement dédié à l'open data des entreprises.

Figure 68. Logo du site DataInfoGreffe

Sur ce site, des données au niveau de granularité « entreprise » sont disponibles. *Il s'agit des chiffres clés des sociétés commerciales ayant déposé leurs comptes annuels pour l' exercice 2019, enrichis des années 2017 et 2018.* Pour chaque entreprise, la base de données « chiffres clés 2019 » restitue plusieurs typologies de données :

- Dénomination,
- Libellé APE,
- Localisation,
- Données spécifiques au rattachement GREFFE.
- Tranche de chiffre d'affaires par millésime.

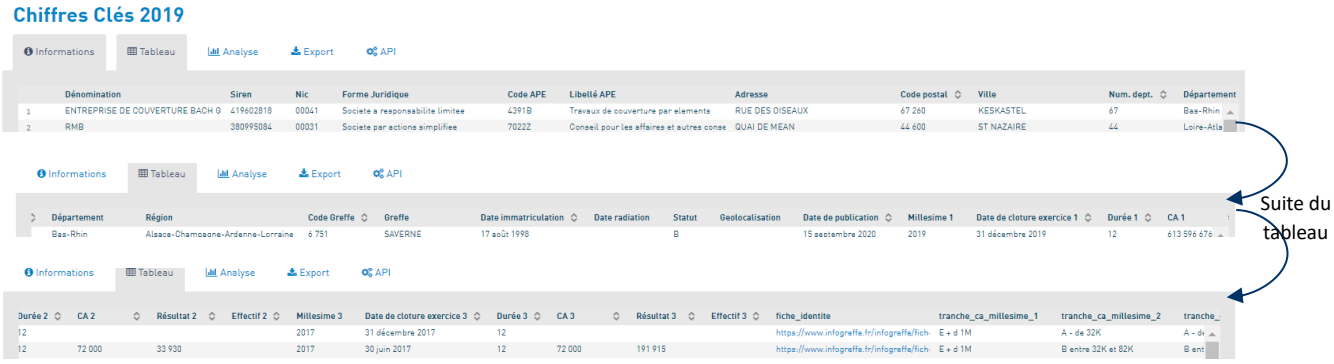

Figure 69. Extrait des variables du tableau chiffres clés 2019 des entreprises

La base de données individus / SIRET à disposition est enrichie des données exogènes issues de la base « Chiffres clés 2018 ». Sur un périmètre initial de 45 871 SIRET (comprenant 572 APET), le croisement des données a pu être réalisé sur 10 % des SIRET (3 632) correspondant à 68% des codes APET (388).

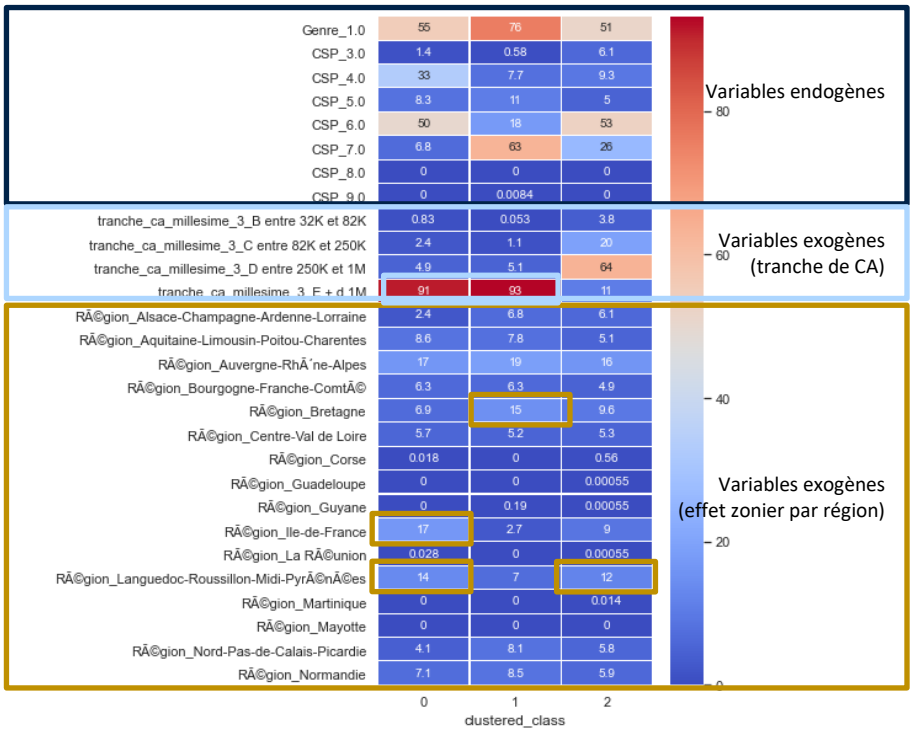

Figure 70. Partitionnement avec k=3 avec les variables endogènes et exogènes

Sur les variables endogènes, les facteurs CSP 4, 6 et 7 restent discriminants.

Sur les variables exogènes, le constat est le suivant :

- Le facteur « chiffre d'affaires » semble intervenir dans la segmentation des secteurs d'activité,
- Le facteur zonier (granularité régionale) semble intervenir également : les régions « Île de France », « Languedoc-Roussillon-Midi-Pyrénées » et « Bretagne » semblent être discriminantes. Toutefois, la localisation géographique (effet zonier) et les secteurs sont susceptibles d'être corrélés. A titre d'exemple, il est bien connu que le bassin d'activité de l'aéronautique en France est Toulouse. Un retraitement est éventuellement à faire pour exploiter complètement ce facteur.

Il est possible de réaliser partiellement la correspondance entre ces nouveaux clusters 0, 1 et 2 (déterminés par des variables endogènes et exogènes) et les classes déterminées par les variables endogènes.

Pour approfondir l'analyse, il aurait pu être intéressant d'identifier les facteurs les plus pertinents pour la segmentation. Pour cela, les forêts aléatoires constituent un recours possible pour évaluer « l'importance des variables ». Avec python, il s'agit de mettre en œuvre un modèle $Y = m(X)$ et d'utiliser la méthode $m.feature\_importances\_$. Y correspondrait au vecteur réponse des clusters, X la matrice avec les variables en colonne et m le modèle.

Pour simplifier la lecture, les classes déterminées par les variables endogènes sont dites classes « endo » et celles déterminées par les variables endogènes et exogènes sont dites classes « exo ».

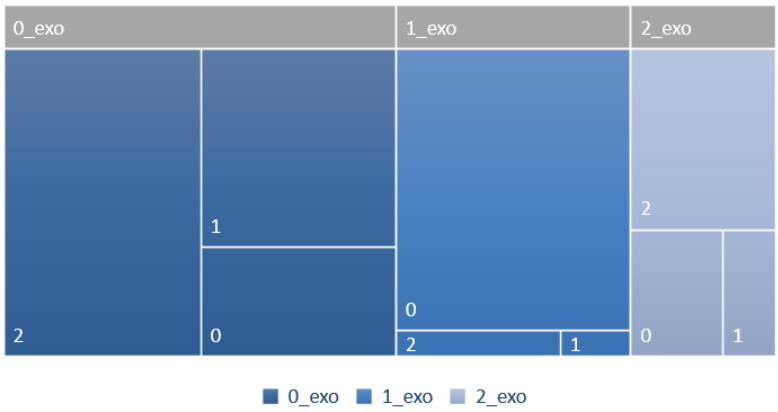

Figure 71. Correspondance clusters "exo" - clusters "endo"

Pour 45% des cas, il n'était pas possible de réaliser de correspondance. En effet, pour rappel, 68% des codes APET ont pu être complétés par les variables issues de l'open data.

d) Analyser la distribution des taux d'incidence intra groupe pour l'échantillon aux variables strictement endogènes et attribuer un niveau de risque à chaque cluster

A partir du tableau des APET regroupés en 3 clusters selon des variables endogènes, le taux d'incidence par APET est reconstitué :

APET	clustered_class	Nombre d_AT	valuecount	exposition&quotite	taux
0113Z	0	7,38	669	151135,9	4,88302E-05
0121Z	0	0	2	440,13	0
0124Z	0	0	1	368,99	0
0126Z	0	0	2	790	0
0129Z	2	15,38	23	5195,83	0,002960066

Figure 72. Illustration taux d'arrêt de travail par APET

Il convient d'analyser la répartition des taux d'incidence par cluster pour statuer sur le comportement moyen de chaque cluster vis-à-vis de la sinistralité.

Pour un nombre de cluster $k = 3$, les boîtes à moustaches ci-dessous représentent la répartition du taux de sinistres par classe pour l'échantillon des APE sans variable issue de l'open data.

La répartition des taux de sinistre par cluster est relativement homogène pour la classe 2.

Quant aux clusters 0 et 1, leur composition intra groupe semble plus dispersée.

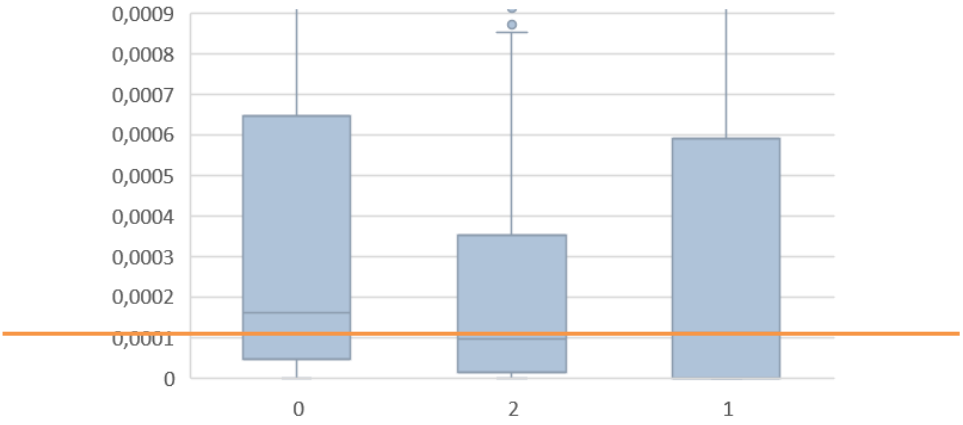

Figure 73. Répartition des taux d'incidence par classe (données endogènes)

La représentation en boîte à moustache ci-dessus permet d'identifier la classe 2 comme ayant le niveau de risque moyen le plus faible par rapport aux clusters 0 et 1. Cela corrobore l'idée d'éventuellement rajouter des variables discriminantes pour partitionner de nouveau ce cluster.

e) Analyser la distribution des taux d'incidence intra groupe pour l'échantillon aux variables endogènes et exogènes et attribuer un niveau de risque à chaque cluster

A l'instar de ce qui est réalisé pour les clusters « endo », les taux d'incidence sont répartis par cluster.

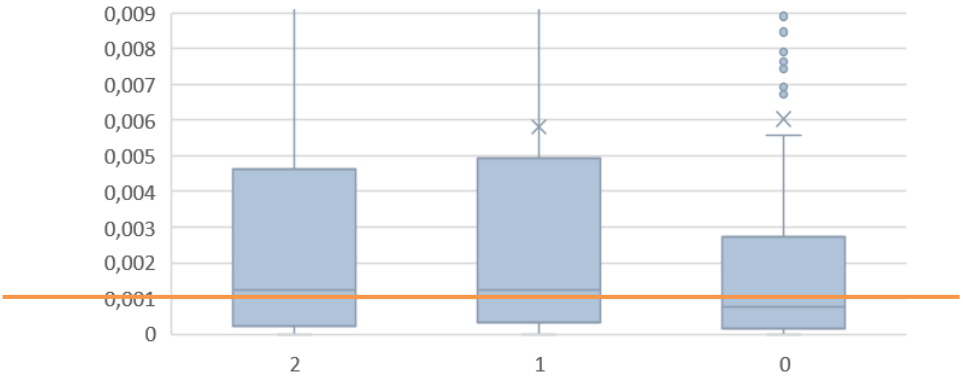

Figure 74. Répartition des taux d'incidence par classe (données exogènes)

La représentation en boîte à moustache ci-dessus est équivalente à celle du paragraphe précédent. La classe 0 est celle qui a le niveau de risque moyen le plus faible par rapport aux deux autres.

La segmentation des codes APET permet de dégager une légère tendance pour la segmentation. Mais celle-ci n'est pas suffisamment significative.

Malgré la faiblesse du résultat et afin de dérouler le raisonnement jusqu'au bout, nous proposons de nous baser sur le clustering avec les données exogènes pour déterminer les différentes courbes de tarifs.

5.3. APPLICATION A LA TARIFICATION

En appliquant les lois de maintien et d'incidence, il est possible de tracer la courbe des tarifs « bruts » par âge x.

$$\pi_{pure}(x, f, csp, genre, i) = \pi_{pure}(x, 0, csp, genre_{unisex}) * Correctif_{classe\ de\ risque_i}(x)$$

L'application est réalisée pour les 3 situations (Classes 0, 1 et 2).

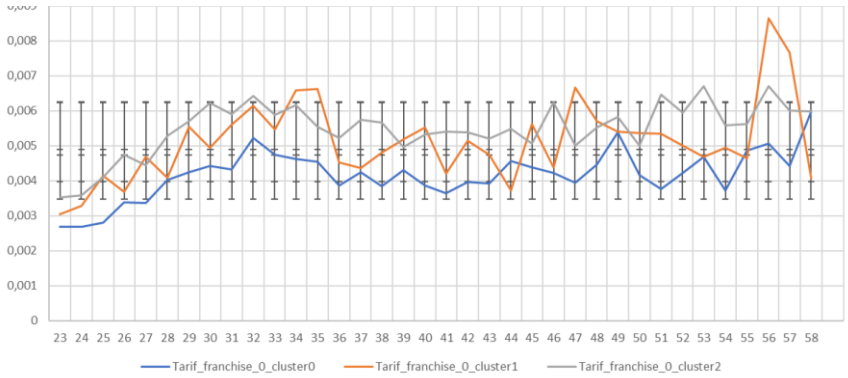

Figure 75. Tarifs bruts journaliers en fonction de l'âge pour une franchise à zéro

Ces courbes présentent des irrégularités (qui n'ont pas de signification réelle). Il conviendrait de les lisser.

Visuellement, les courbes des classes de codes APET 0 et 2 sont séparées. La courbe des tarifs de la classe 0 est en dessous des courbes des deux autres. Le risque semble plus élevé pour les individus des codes APET de la classe 2. La courbe du groupe 1 est en tendance entre les deux courbes mais deux tranches d'âge (30-35ans et 45-49 ans) sont proches des tarifs du groupe 2.

Le point faible de la technique est matérialisé par les intervalles de confiance : cela ne permet pas de statuer complètement sur la significativité de la segmentation proposée.

La courbe des tarifs avec l'intégralité des codes APET est ajoutée pour visualiser la tarification non segmentée (courbe verte) :

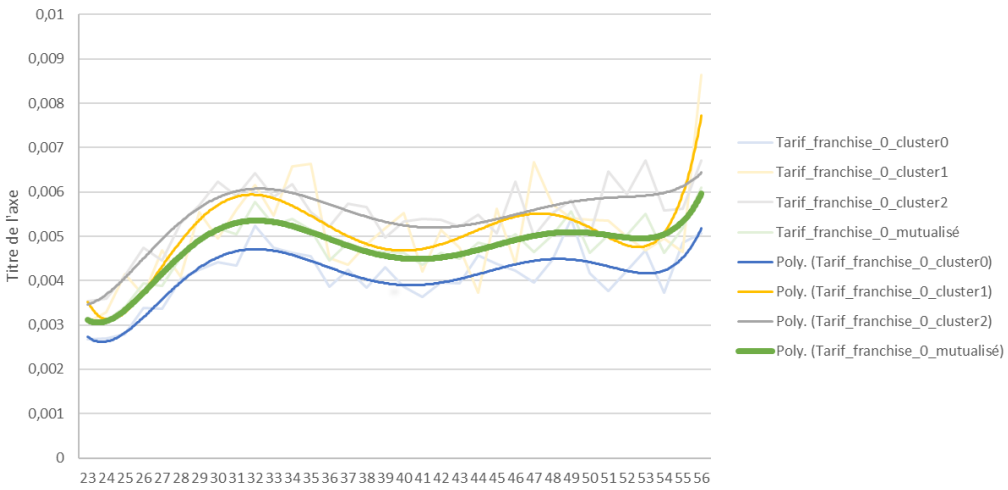

Figure 76. Tarifs bruts journaliers en fonction de l'âge

En regroupant par tranche d'âge et en considérant le tarif mutualisé avec la franchise 0 comme référence, cela donne le tableau des coefficients suivant :

Tranche d'âge	Tarif_cluster0	Tarif_cluster1	Tarif_cluster2	Tarif_mutualisé
00 [0,25[	87%	103%	115%	100%
01 [25,30[	86%	107%	118%	100%
02 [30,35[	88%	108%	115%	100%
03 [35,40[	88%	107%	114%	100%
04 [40,45[	87%	102%	117%	100%
05 [45,50[	89%	111%	111%	100%
06 [50,55[	83%	102%	120%	100%
07 [55,60[	90%	111%	108%	100%

Figure 77. Grille de coefficient de majoration – minoration

La classe 0 est composée de secteurs comportant majoritairement des employés (CSP 6) représentés majoritairement par des femmes présentes en région Île de France, Rhône-Alpes, Languedoc Roussillon-Midi-Pyrénées avec une tranche de CA des entreprises supérieure à 1M d'euros.

La classe 1 est composée de secteurs employant majoritairement des ouvriers (CSP 7) et dont les entreprises ont un chiffre d'affaires supérieur à 1M€, implantées en Rhône-Alpes, Bretagne.

La classe 2 est composée de secteurs dont la répartition est mixte en termes de genre. Ce cluster comporte également de la CSP 6 avec des entreprises dont le chiffre d'affaires est dans la tranche inférieure à celle de la classe 0.

Pour aller jusqu'au bout de l'exploration, la tarification pour toutes les franchises souscrites pourrait être analysée.

Il conviendrait également de tester l'adéquation de la segmentation proposée avec des données historiques réelles.

Les résultats obtenus semblent corroborer l'intuition initiale : une segmentation par secteur d'activité est susceptible d'être intéressante. Toutefois, il convient d'être vigilant. En effet, une segmentation imparfaite pour l'assureur peut l'exposer à un risque d'antisélection important (cf. Planchet F. (2019).) : conserver les mauvais risques et rebuter les bons risques. Il convient donc d'être vigilant au compromis mutualisation versus segmentation.

Pour finaliser notre démarche, nous proposons d'étudier la robustesse de la modélisation. Pour cela, il est courant de réaliser des études de sensibilité, des analyses d'impacts.

5.4. ETUDE DE SENSIBILITE SUR LA PROPOSITION DE SEGMENTATION TARIFAIRE BRUTE

L'analyse de sensibilité consiste à étudier l'effet de perturbations des facteurs d'entrée (notre proposition de segmentation) sur les variables de sortie (notre tarification segmentée).

Nous choisissons de fusionner quelques classes et d'étudier l'effet sur la tarification. La démarche consiste à :

- Regrouper des modalités d'un des facteurs,
- Relancer l'algorithme de partitionnement,
- Etablir la tarification avec les nouvelles classes de risques.

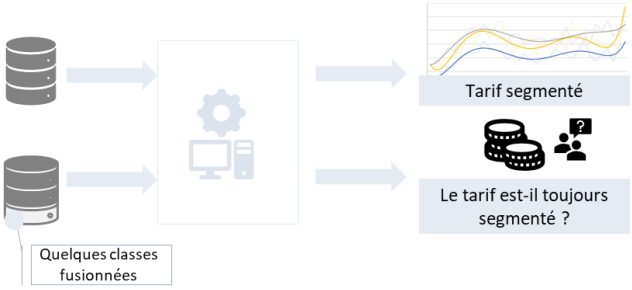

Figure 78. Démarche de l'étude de sensibilité par fusion de classe

Comme le facteur CSP semble discriminant, nous décidons de tester la sensibilité du modèle au regroupement des effectifs par CSP : les CSP 3, 5, 8 et 9 identifiées comme a priori peu influentes sur la segmentation des classes de secteurs d'activité sont regroupées.

Après avoir lancé l'algorithme de clustering et déterminé de nouveau la tarification, les courbes de tarifs bruts et lissés en fonction de l'âge restent quasiment inchangées :

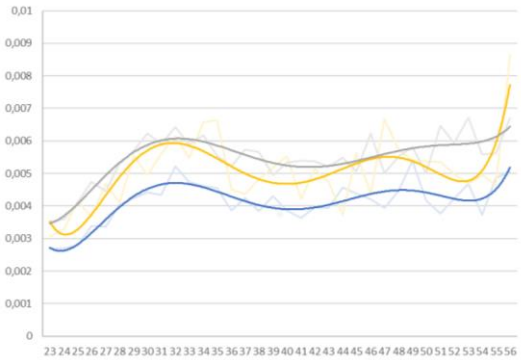

Figure 79. Tarification brute et lissée initiale

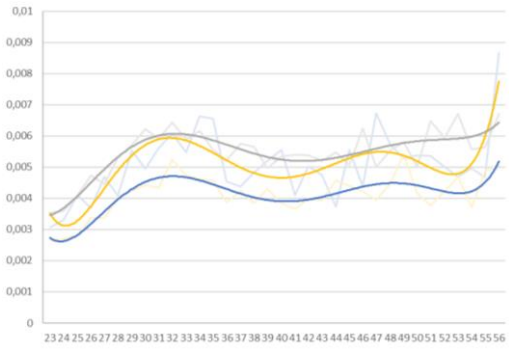

Figure 80. Tarification modifiée suite au regroupement de CSP (3, 5, 8 et 9)

Pour confirmer le constat, les écarts relatifs entre les courbes sont tracés dans le graphe ci-dessous. Chaque écart est calculé de la manière suivante : $écart = \frac{valeur_{modèle\ initial} - valeur_{modèle\ csp\ regroupé}}{valeur_{modèle\ initial}}$.

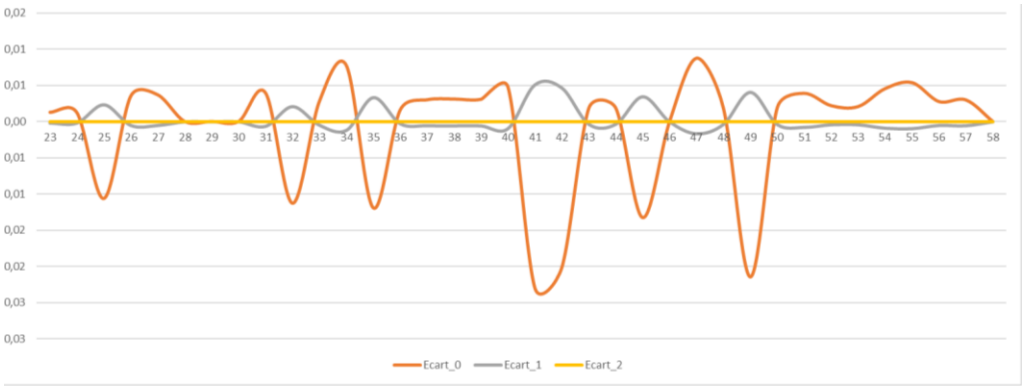

Figure 81. Fluctuations entre le modèle initial et le modèle "CSP regroupés"

Les variations relatives sont faibles et inférieures à 3%.

Nous renouvelons l'expérience en regroupant les CSP 4 et 6. Les courbes ci-dessous montrent un écart significatif.

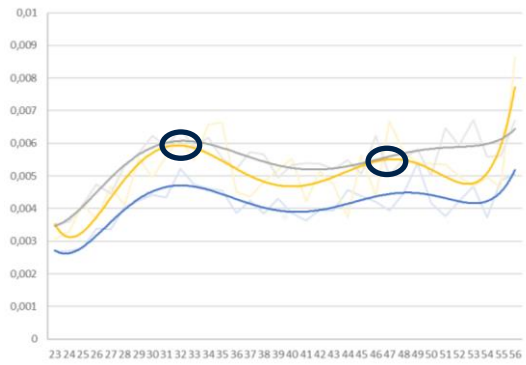

Figure 82. Tarification brute et lissée initiale

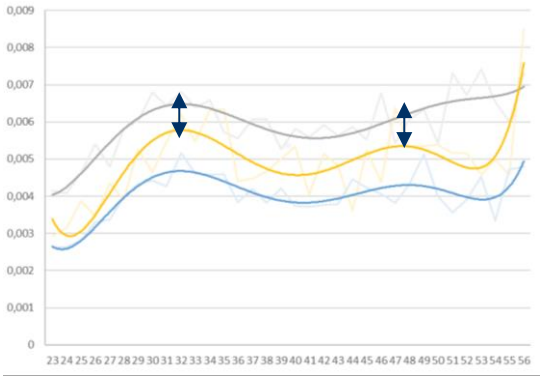

Figure 83. Tarification modifiée suite au regroupement de CSP (4 et 6)

L'expérience semble montrer les points suivants :

- Les tarifs restent segmentés,
- Le modèle est sensible au regroupement de CSP. Dans la deuxième expérience, la courbe de tarif jaune est visuellement davantage « séparée » de la courbe grise que dans le modèle initial.

Suite à ce dernier constat, nous nous sommes interrogés sur l'impact du paramètre de distance dans la proposition de segmentation tarifaire.

5.5. INFLUENCE DU PARAMETRE DE DISTANCE SUR LA PROPOSITION DE SEGMENTATION TARIFAIRE BRUTE

En particulier, nous souhaitions vérifier l'impact du choix du paramètre de distance sur le pouvoir de correctement partitionner et in fine sur la segmentation tarifaire.

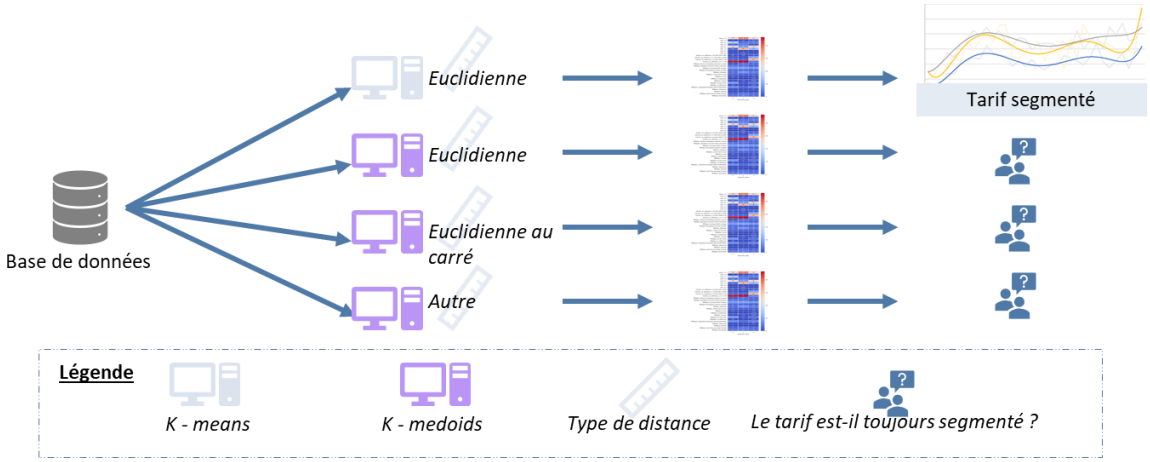

Figure 84. Démarche de l'étude de l'influence du paramètre de distance

Pour rappel, l'algorithme des K-means minimise la quantité : $\underset{\{C_1,C_2,\dots,C_k\}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{X_j \in C_i} \left|\left|X_j - \mu_i\right|\right|_{l2}^{2}$

Avec :

- k le nombre total de classes à former et inférieur à N,
- i allant de 1 à k correspondant au numéro de la classe,
- $X_j \in C_i$ correspondant aux points regroupés dans la classe C_{k_i},
- μ_i est le centre de gravité de la classe C_i.

Cela revient à minimiser une variance intra-classe, c'est-à-dire minimiser la somme des distances euclidiennes au carré de chaque point de la classe par rapport au centre. Pour mieux visualiser l'analogie, nous rappelons :

- L'expression de la variance mathématique : $V(X) = \frac{1}{n} * \sum_{i=1}^{n}(x_i - \bar{x})^2 \ \ avec\ \bar{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i)$,
- L'expression du carré de la distance euclidienne entre deux points $X\ et\ Y$: $\sum_{i=1}^{n}(x_i - y_i)^2)$.

Si nous souhaitons modifier le critère de distance, il faut donc utiliser les algorithmes similaires à K - means (également présentés dans le cours de Maxime Sangnier). Il s'agit de l'algorithme des K - medoids avec un critère de minimisation modifié : pour chaque j appartenant à l'ensemble des k , $\hat{\mu}(C_j) = X_t$ avec $t \in argmin_{i \in [n]} \sum_{X \in C_j} d(X, X_i)^2$ avec :

- $\hat{\mu}(C_j)$ étant le centroïde de la classe. A la différence majeure avec K - means, le centroïde fait partie de l'échantillon des observations $(X_1, X_2, \dots X_n)$,
- t correspond à l'indice de l'observation correspondant au critère de minimisation.

Dans un premier temps, nous proposons d'analyser les éventuels écarts entre les propositions de segmentation tarifaire fournies par k-means ou k-medoids.

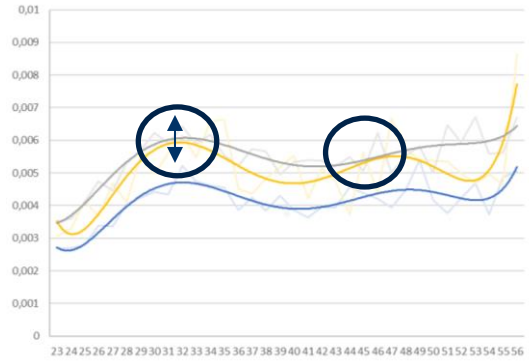

Figure 85. Tarification (segmentation k-means)

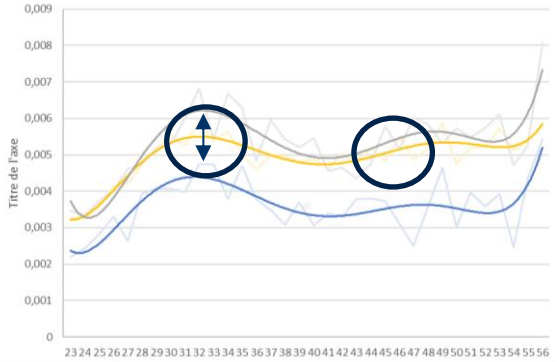

Figure 86. Tarification (segmentation k-medoids)

Les courbes de tarifs ci-dessus montrent que la proposition de segmentation tarifaire est sensible au critère : centroïde appartenant au jeu de données / centroïde n'appartenant pas au jeu de données.

Nous avons lancé l'algorithme des k-medoids avec 3 distances différentes :

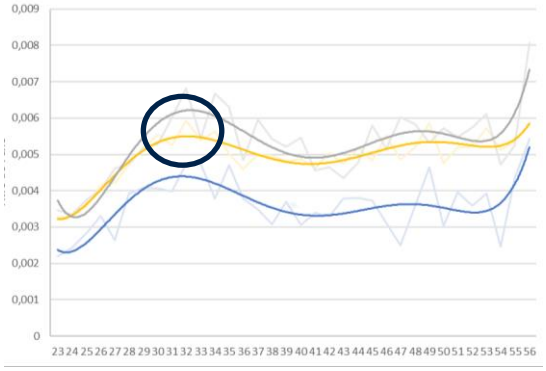

Figure 87. Tarification (d=euclidienne au carré)

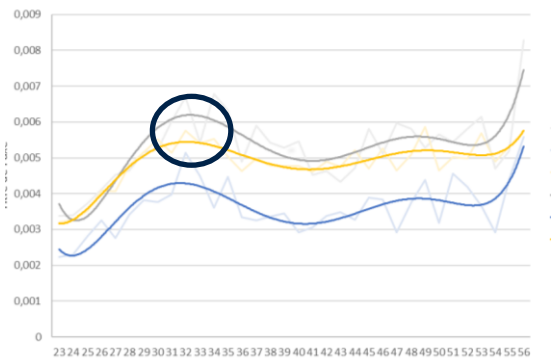

Figure 88. Tarification (d = euclidienne)

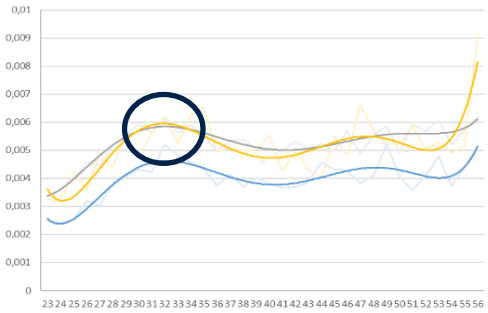

Figure 89. Tarification (d= de Manhattan)

Distance euclidienne au carré	$\sum_{i=1}^{n}(x_i - y_i)^2$
Distance euclidienne	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Distance de Manhattan	$\sum_{i=1}^{n}\lvert x_i - y_i\rvert$

D'après les graphes ci-dessus, la proposition de segmentation tarifaire est sensible au choix du paramètre de distance.

5.6. LES LIMITES DU CHOIX D'UN ALGORITHME DE TYPE K-MEANS

Pour rappel, le but était de consister des classes de risque pour la tarification. Pour des raisons exploratoires, nous avons sélectionné l'algorithme K-means. L'approche consistait à constituer k classes homogènes avec toutes les variables descriptives d'un échantillon à disposition. Pour segmenter les secteurs d'activité, nous avons considéré les facteurs CSP, région, répartition hommes – femmes... Mais la limite principale de cette approche réside justement dans la sélection des facteurs. Si nous avions considéré la variable « couleur des yeux », l'algorithme s'en serait également servi pour départager les secteurs d'activité en plusieurs classes homogènes.

Comme l'objectif est de construire des classes de risque homogènes et non simplement des classes homogènes, il est donc nécessaire de disposer de variables susceptibles d'influer sur le risque. C'est donc par exemple pour cette raison que la couleur des yeux n'est pas choisie (*au-delà du fait que l'assureur ne dispose pas de la donnée*).

Notre approche revient donc à :

1) Suspecter l'influence d'une variable sur le risque,
2) Déterminer une segmentation par les K-means,
3) Infirmer ou confirmer la suspicion de l'influence de la variable sur le risque.

Cette démarche est susceptible d'être pertinente lorsque nous avons une intuition sur le degré d'influence de la variable sur le risque. Mais elle ne semble pas pouvoir garantir une constitution des classes de risques optimales.

Traditionnellement, les arbres de régression ou de classification (algorithme CART) sont privilégiés pour ce type de problème. En effet, ils ont une visée explicative : dès lors que la variable n'a pas d'influence sur l'objectif recherché (pour nous, c'est le risque) alors la variable est éliminée. Dans ce cas, les variables sont hiérarchisées par des algorithmes prévus à cet effet.

5.7. LES CONTROLES A METTRE EN ŒUVRE POUR VALIDER LA FIABILITE DE L'APPROCHE TARIFAIRE

La mise en œuvre d'une démarche de contrôle de cohérence est un prérequis avant d'utiliser en production de nouvelles lois d'expérience. Parmi eux, il convient a minima de contrôler :

- Le niveau de prudence,
- L'équilibre technique,
- La rentabilité globale.

NB : compte-tenu du durcissement des conditions d'accès aux données entre la phase de démarrage du mémoire et la présente phase de « backtesting » (contraintes émises par le DPO + fin d'intervention chez l'organisme assureur)

5.7.1) 1er contrôle : impact sur le niveau de prudence

Dès lors que les lois d'expérience conduisent à un niveau de prudence plus bas que celui de la réglementation, l'organisme assureur est exposé à un risque de perte accrue.

5.7.2) 2ème contrôle : analyse de la proximité des tarifs avec les tarifs d'équilibre technique

Le ratio $Sinistres/Cotisation$ dit S/C est un indicateur qui mesure l'équilibre technique des contrats d'assurance par année de survenance. Il se calcule de la manière suivante :

$$\frac{S}{C} = \frac{Sinistres\ versés\ par\ année\ N_s + Provisions\ Mathématiques\ constituée\ par\ année\ N_s}{Cotisations\ perçues\ dans\ l'année}$$

Avec $N_s = année$ à $la\ survenance$

Le contrat est à l'équilibre technique lorsque $\frac{S}{C} = 100\%$.

Pour recalculer les tarifs d'équilibre, les ratios S/C sont calculés par année de survenance. Pour ce faire, la démarche consiste à reconstituer un triangle de liquidation à partir des sources suivantes :

- Le SI de gestion des sinistres : 10 ans d'historique de données disponibles,
- Le SI de gestion du recouvrement (ou autres données de type « infocentre ») : données non fournies par l'organisme assureur.

Le triangle de liquidation se présente par année de survenance sous la forme du tableau suivant :

	N	$N+1$	$N+2$	$N+3$	…	$N+10$	Somme des sinistres	PM	Somme des cotisations
N	$s_{N,N}$	$s_{N,N+1}$	$s_{N,N+2}$	0	0	0	S_N	PM_N	C_N
$N+1$	-	$s_{N+1,N+1}$	$s_{N+1,N+2}$	$s_{N+1,N+3}$	0	0	S_{N+1}	PM_{N+1}	C_{N+1}
N+2	-	-	$s_{N+2,N+2}$	$s_{N+2,N+3}$	$s_{N+2,N+\cdots}$	0	S_{N+2}	PM_{N+2}	C_{N+2}
$N+3$	-	-	-	$s_{N+3,N+3}$	$s_{N+3,N+\cdots}$	0	S_{N+3}	PM_{N+3}	C_{N+3}
…	-	-	-	-	$s_{N+\cdots,N+\cdots}$	0	$S_{N+\cdots}$	$PM_{N+\cdots}$	$C_{N+\cdots}$
N+10	-	-	-	-	-	$s_{N+10,N+10}$	S_{N+10}	PM_{N+10}	C_{N+10}
Total									

Pour chaque ligne du tableau, il convient de calculer le tarif d'équilibre pour un échantillon d'entreprises présélectionnées.

En effet, ce contrôle est à réaliser sur un panel d'entreprises représentatif du portefeuille. Pour notre client, la segmentation s'effectue selon la taille de l'entreprise (grands comptes, petites et moyennes entreprises, bas de segment). Les entreprises sans sinistralité observée sont à exclure du panel.

Par chaque entreprise du panel, les tarifs d'équilibre sont à comparer avec les tarifs fournis par les modèles.

Le tableau ci-dessous est une illustration (dans une version confidentielle) des résultats du back-testing :

Identifiant			Caractéristiques du contrat			Démographie							Barèmes (par tranche de rémunération)			
Raison sociale	Siren	APE (2 caractères)	Collège	Franchise	Niveau de prestations IJ	effectif cadres	effectif maitrise	effectif ouvrier	effectif employé	effectif total	%homme	âge moyen	Tarif existant	Tarif d'équilibre	Classe de risque	Barème modélisé
Entreprise A	.....	..	Cadre	30 jrs continues	100% T12	19	0	0	0	19	68%	47	0,72%	0,74%	2	0,53%
Entreprise B	.....	..	Non cadre	60 jrs continues	90% T12	0	0	0	309	309	30%	40	0,47%	0,85%	2	0,48%
…	… …	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

NB : dans le tableau, les barèmes sont exprimés en pourcentage de tranche de rémunération.

En fonction des résultats, le tarif modélisé nécessite d'être ajusté. En effet :

- Si le tarif modélisé est proche du tarif d'équilibre alors il est retenu,
- Si non, il est ajusté pour que l'équilibre technique soit atteint.

L'inconvénient de cette démarche réside dans le fait de se limiter à l'analyse de l'historique sans prendre en compte les phénomènes futurs de dérive. Pour intégrer la part d'incertitude liée à la dérive de sinistralité, il pourrait être envisagé d'intégrer des éléments de simulation.

5.7.3) 3ème contrôle : analyse de la rentabilité globale

Pour vérifier l'impact sur la rentabilité globale, le ratio combiné est un des indicateurs susceptibles de la mesurer. Il s'exprime de la manière suivante :

$$Ratio\ combiné = \frac{Sinistres + commissions\ (apporteurs) + frais\ (gestion, acquisition, \dots)}{Primes}$$

	Avec table du BCAC	Avec table d'expérience	Règles de comparaison
Ratio combiné	R_{bcac}	R_{exp}	Si $R_{bcac} > R_{exp}$, alors l'utilisation des tables d'expérience représente un gain de rentabilité. Si non, il convient de maintenir l'utilisation des lois du BCAC.

5.7.4) Les acteurs impliqués dans les contrôles de cohérence

En pratique, le processus de validation est transverse. Il implique plusieurs directions :

- La souscription,
- La surveillance de portefeuille,
- La fonction actuariat.

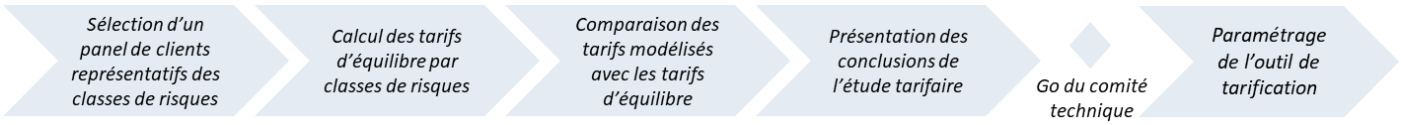

Figure 90. Macro-process de validation d'un nouveau barème de tarification

6. CONCLUSION

L'objet du mémoire était d'étudier l'opportunité de mettre à jour le barème de tarification en vigueur et d'éventuellement envisager une segmentation par la création d'une notion de classe de risque.

La première étape consistait à identifier les principales poches de sinistralité par l'étude descriptive des données à disposition.

Un premier travail a consisté à reconstituer les premiers éléments constitutifs de la formule de tarification : loi de maintien en incapacité et loi d'incidence.

La reconstitution de la loi de maintien a permis de constater que la loi du BCAC surestimait les probabilités de maintien en incapacité des sinistres du portefeuille. Toutefois, une convergence s'observe sur les grandes durées d'arrêt de travail.

La reconstitution de la loi d'incidence est biaisée par la surreprésentation des données 2019 dans l'échantillon à disposition. Elle reste possible avec une sous-estimation du nombre de sinistres observés : faute de disposer des DSN des mois antérieurs, l'échantillon fait l'objet d'une importante troncature à gauche. Toutefois, il est décidé de poursuivre les travaux pour étudier l'opportunité de créer des classes de risque pertinente pour la tarification.

Pour étudier l'opportunité de segmenter le portefeuille d'entreprises, la démographie des secteurs d'activité est étudiée. L'utilisation de l'algorithme K-Means a permis d'établir des groupes homogènes de secteurs. En réalité, cela a été possible par « répercutions » suite au choix de variables pressenties influentes sur le risque. La détermination des classes de secteurs d'activité s'est limitée à l'exploration de variables endogènes (âge, CSP, répartition homme femme) ainsi que sur les variables exogènes liés au contexte géographique et économique. Des variables supplémentaires disponibles dans les DSN auraient pu être utilisées : l'effectif, le turnover (sur l'année combien d'entrée et de sortie), ancienneté dans l'entreprise, les franchises souscrites par secteur d'activité, situation familiale.

Les résultats obtenus pour la tarification de chaque classe semblent montrer un intérêt à segmenter la tarification. Toutefois, cette proposition de segmentation reste à tester avec les données utilisées pour tarifer la grille actuellement en vigueur.

Pour continuer les travaux, il conviendrait de tester la segmentation sur des données réelles et analyser la répartition des risques entre assuré et assureur. Par ailleurs, la constitution de classes de risques pourrait être enrichie par d'autres critères (seul le critère « secteur d'activité » a été exploré dans ce mémoire).

7. BIBLIOGRAPHIE

Les Mémoires d'actuariat

[1.] Camille Mosse. (2007). Construction d'un indicateur de maintien en arrêt de travail - Apport des simulations à la tarification et à la décision – Mémoire d'actuariat ISFA
[2.] Mario GUGUMUS. (2009). Modélisation de l'incapacité temporaire et de l'invalidité en prévoyance collective – mémoire de l'université d'actuariat de Strasbourg
[3.] Maxime Huttin. (2013). Refonte tarifaire du produit Prévoyance Evolution, accompagnée du modèle de rentabilité - Mémoire d'actuariat EURIA
[4.] Cyprien Herbreteau. (2017). Impacts des nouvelles tables du BCAC en termes de provisionnement, construction d'une nouvelle loi de maintien et d'un outil de tarification pour l'incapacité

D'autres mémoires pour compléter la compréhension des méthodes d'estimation de loi d'incidence et de maintien :
[5.] Alexandre de La Morinerie. (2016). Conceptualisation d'un modèle multi-états en arrêt de travail et application à une loi d'incidence en incapacité – Mémoire ENSAE
[6.] Aurélie GAUMET. (2001). Construction de tables d'expérience pour l'entrée et le maintien en Incapacité

Les supports pédagogiques
[7.] Aymric Kamega, Frédéric Planchet, Roberto Wolfrum. (2013). Présentation et comparaison des nouvelles tables BCAC
[8.] Olivier Lopez. (2019). Big Data and Actuarial Sciences - cours dispensé au CEA
[9.] Christophe Izart. (2018). Prime pure / prime commerciale – cours dispensé au CEA
[10.] Claire Boyer. (2019). Machine Learning - cours dispensé au CEA
[11.] Maxime Sangnier. (March 21, 2019). Unsupervised learning - Cours dispensé au CEA
[12.] Ludovic Macaire. (11 mars 2014). Clustering par kmeans - Cf. cours de l'Université Lille 1
[13.] V. Jourdan. (2008/2009). Statistiques bivariées - Université Marc Bloch – Strasbourg 2
[14.] Ricco Rakotomalala - Analyse de corrélation (Étude des dépendances - Variables quantitatives) - Cours de l'Université Lyon 2
[15.] Planchet F. (2019). Quelques réflexions sur la segmentation en assurance – ISFA : https://documentcloud.adobe.com/link/review?uri=urn:aaid:scds:US:14b83a3a-a132-4555-bf1f-221f17cc102e#pageNum=1

Sources open data
[16.] Marilyne Beque, Aimée Kingsada, Amélie Mauroux. (2019). Synthèse Stat' - Reconnaissance, insécurité et changements dans le travail (Numéro 29 avril 2019) : https://dares.travail-emploi.gouv.fr/sites/default/files/pdf/dares_synthese_stat___reconnaissance_insecurite_changements.pdf
[17.] Site internet de l'InfoGreffe : https://opendata.datainfogreffe.fr

Les autres références
[18.] Formation interne aVB : Panorama de l'assurance collective, présentation donnée en interne aVB.
[19.] Site internet du Centre Technique des Institutions de Prévoyance : https://ctip.asso.fr/la-prevoyance-collective/quest-ce-que-la-prevoyance-collective/
[20.] Site web de la Fédération Française de l'assurance : https://www.ffa-assurance.fr/
[21.] Morice, E.; Thionet, P. (1969). Revue de Statistique Appliquée, Tome 17 no. 3, pp. 75-89 : http://www.numdam.org/article/RSA_1969__17_3_75_0.pdf
[22.] Caisse nationale d'assurance vieillesse. (2020). Cahier technique DSN : http://www.dsn-info.fr/documentation/dsn-cahier-technique-2020.pdf

[23.] BCAC. (2013). Arrêt de travail : lois de provisionnement du BCAC: http://www.ressources-actuarielles.net/bcac
[24.] Cam Davidson-Pilon. (2014-2021). LIFELINES - https://lifelines.readthedocs.io/en/latest/ - Librairie pour l'analyse de survie avec le langage Python

8. ANNEXES

8.1. TABLES DES ILLUSTRATIONS

8.2. LES CSP

01 - agriculteur salarié de son exploitation
02 - artisan ou commerçant salarié de son entreprise
03 - cadre dirigeant
04 - autres cadres au sens de la CCN
05 - profession intermédiaire
06 - employé administratif d'entreprise, de commerce, agent de service
07 - ouvriers qualifié et non qualifié y compris ouvriers agricoles
08 - agent de la fonction publique d Etat
09 - agent de la fonction publique hospitalière
10 - agent de la fonction publique territoriale

9. NOTE DE SYNTHESE

Le marché de la prévoyance collective est ultra concurrentiel. Ces acteurs lancent des initiatives pour trouver de nouvelles sources de rentabilité. Par exemple, divers projets de transformation sont lancés pour réduire les coûts, optimiser les processus de gestion en intégrant de nouveaux flux de gestion (ex : la Déclaration Sociale Nominative). Sur la gestion de l'absentéisme, certains acteurs s'interrogent sur la compétitivité de leurs garanties. C'est justement le cas de l'organisme assureur partenaire. L'objectif de ce mémoire est d'étudier l'opportunité de proposer un tarif adapté en fonction du profil de l'entreprise.

Cela permettrait à l'OA de mieux organiser ses actions commerciales, de proposer des tarifs plus adaptés aux profils des souscripteurs. Cela passe par proposer un tarif qui reste prudent pour l'assureur mais plus attractif pour l'assuré.

Cela suppose dans un premier temps de réaliser une étude descriptive de la sinistralité à partir des données d'expérience. Cette étude permet de mieux connaitre le portefeuille et en particulier d'identifier les principales poches de sinistralité.

A partir des données du système de gestion, une loi de maintien est reconstituée par l'estimateur non paramétrique de Kaplan-Meier. Celle-ci est comparée à la loi de maintien du BCAC. La loi donnée par les données d'expérience montre clairement que la loi BCAC surestime la probabilité de maintien. Toutefois, la loi d'expérience aurait pu être utilisée telle quelle uniquement sous réserve d'une estimation plus prudente que la loi réglementaire. Par ailleurs, la limite de l'analyse réside dans le fait que seuls les sinistres ayant dépassé la franchise sont pris en compte.

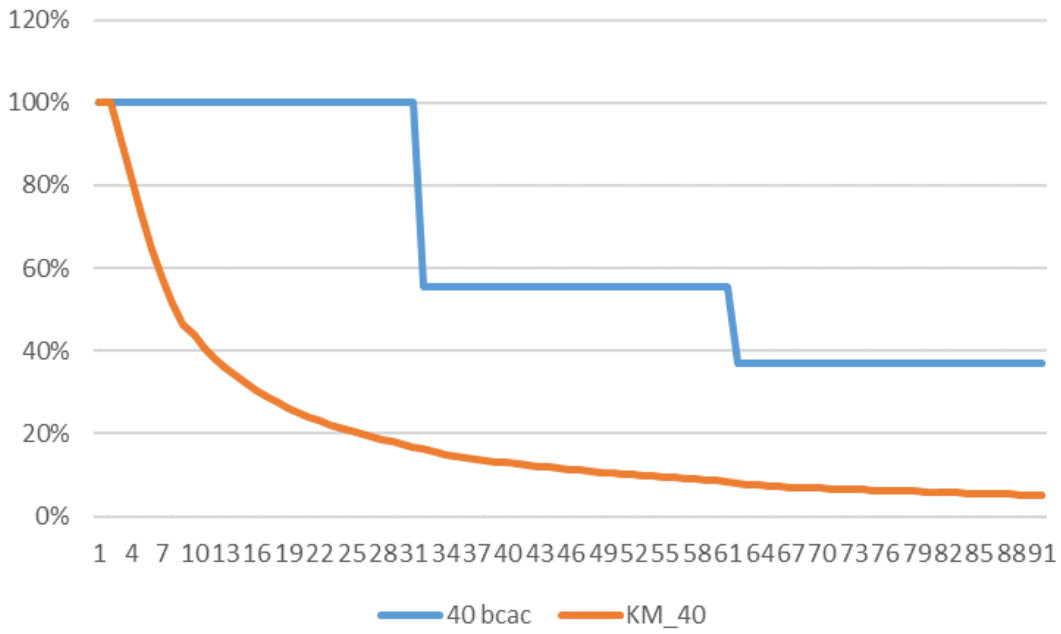

Figure 91. Loi de maintien BCAC versus données d'expérience

Une loi d'incidence est reconstituée à partir de la volumétrie restreinte des données à disposition. Une étude ultérieure est susceptible d'affiner le résultat en fonction de l'enrichissement de données. Dans le cadre du mémoire, l'intérêt est de montrer qu'il est possible de reconstituer cette loi grâce aux données issues de la DSN.

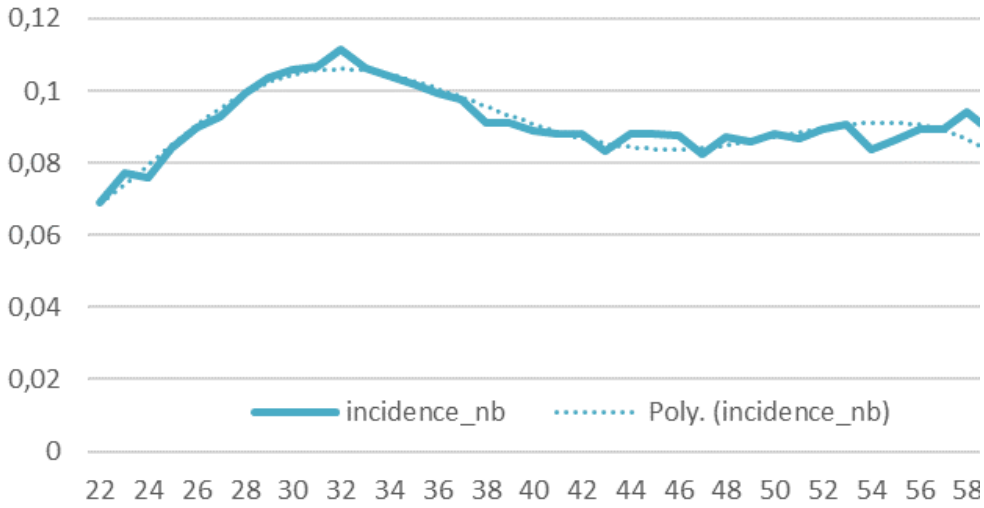

Figure 92. Loi d'incidence

Le mémoire propose d'introduire le secteur d'activité comme facteur à intégrer dans le modèle de tarification. Pour cela, un regroupement en groupe homogène est réalisé sur les codes APET en se basant sur les caractéristiques démographiques de chaque APET de l'échantillon à disposition. L'algorithme K-means est utilisé pour réaliser ce partitionnement.

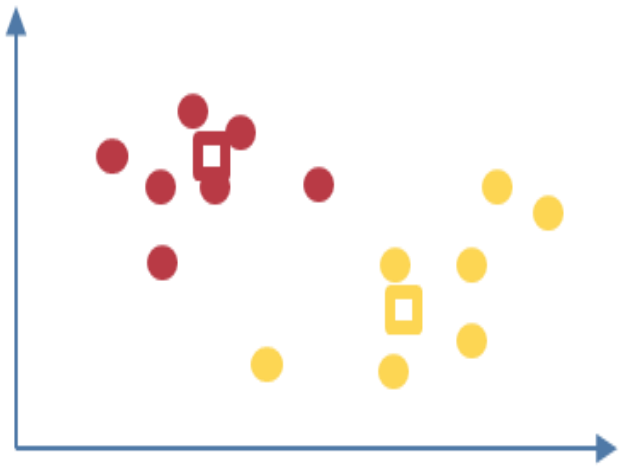

Figure 93. Illustration du principe du clustering par K-Means

Les variables démographiques sont de deux natures :

- endogène (genre, CSP) l'âge n'intervient pas,
- exogène : l'implantation régionale et la tranche de chiffre d'affaires annuel.

Il aurait été intéressant d'étudier les effets de variables de type « effectif dans l'entreprise », turnover, consommation en santé....

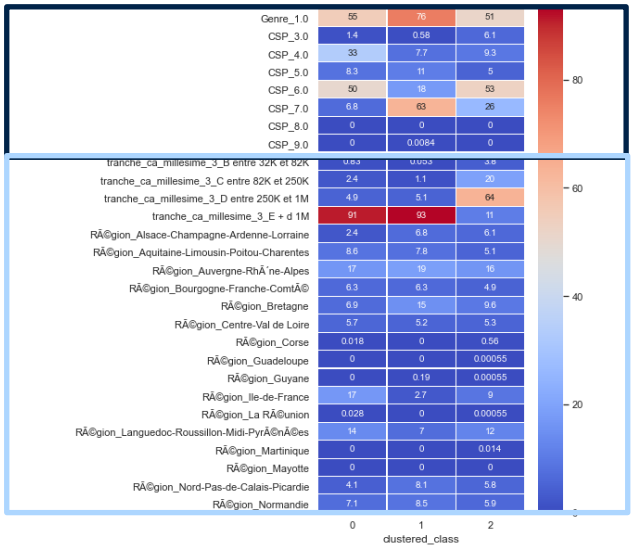

Figure 94. Partitionnement avec k=3 avec les variables endogènes (encadrées en noir) et exogènes (encadrées en bleu)

L'application de la formule de tarification permet de reconstituer les courbes de tarification pour chacune des classes déterminées et le tarif toutes classes confondues. Visuellement, il semble pertinent de distinguer des classes de risque.

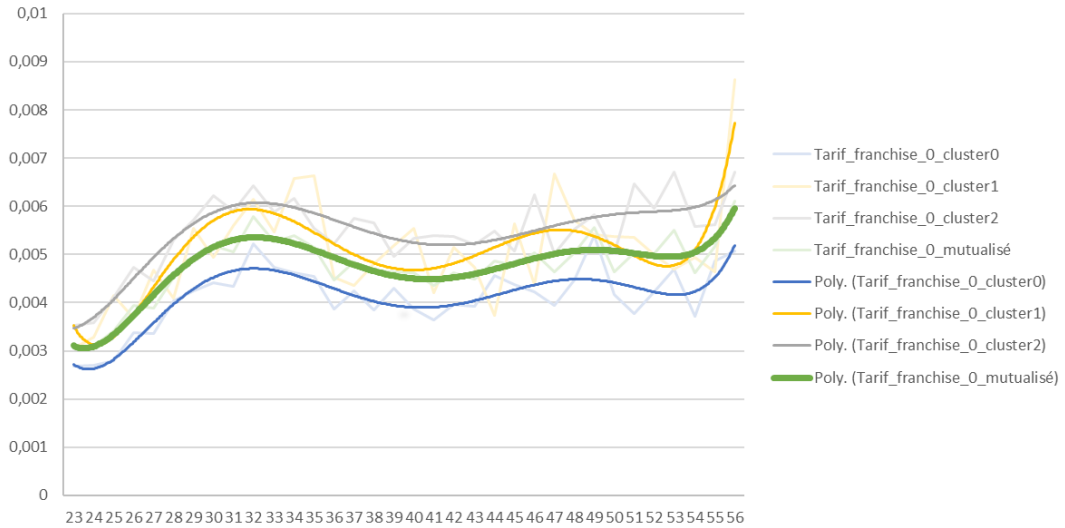

Figure 95. Tarification

10. SYNTHESIS NOTE

The group insurance market is highly competitive. These players are launching initiatives to find new sources of profitability. For example, various transformation projects are launched to reduce costs and optimize management processes by integrating new flows. On the management of absenteeism, some players wonder about the competitiveness of their guarantees. This is precisely the case with the partner insurer. The objective of this thesis is to study the advisability of offering an adapted rate according to the profile of the company.

This would allow the insurance organism to better organize its commercial actions, to offer prices more suited to the profiles of the subscribing companies. This involves offering a price that remains prudent for the insurer but more attractive for the insured.

This first involves carrying out a descriptive study of the loss experience based on experience data. This study allows us to better understand the portfolio and in particular to identify the main pockets of claims.

From the management system data, a probability distribution of maintenance is reconstructed by the Kaplan-Meier nonparametric estimator. This is compared to the BCAC probability distribution of maintenance. The probability distribution given by the experimental data clearly shows that the BCAC probability distribution overestimates the probability of retention. The limitation of the study lies in the fact that only claims exceeding the deductible are taken into account.

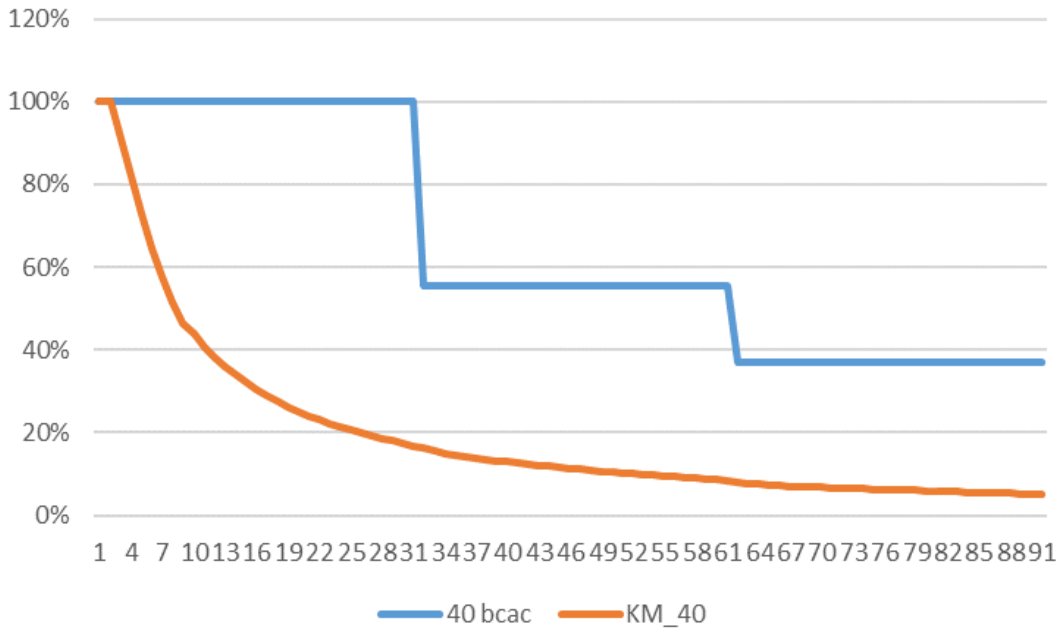

Figure 96. BCAC distribution of maintenance versus experimental data

A probability distribution of incidence is reconstructed from the restricted volume of data available. Further study is likely to refine the result based on data enrichment. As part of the study, the interest is to show that it is possible to reconstruct this probability distribution using data from the DSN.

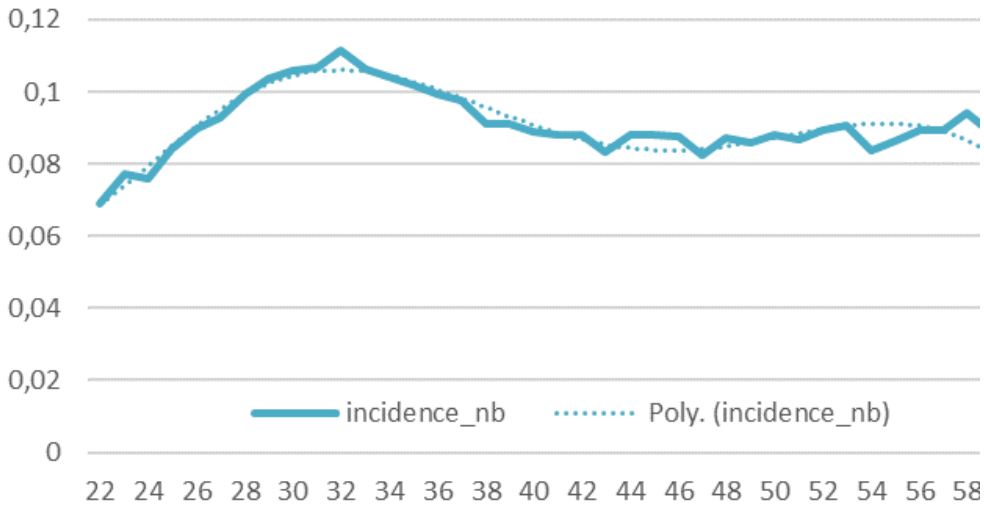

Figure 97. Probability distribution of incidence

The study proposes to introduce the industry as a factor to be integrated into the pricing model. For this, a homogeneous grouping is carried out on the main activity of the company codes (APET in french) based on the demographic characteristics of each APET in the sample available. The K-means algorithm is used to achieve this partitioning.

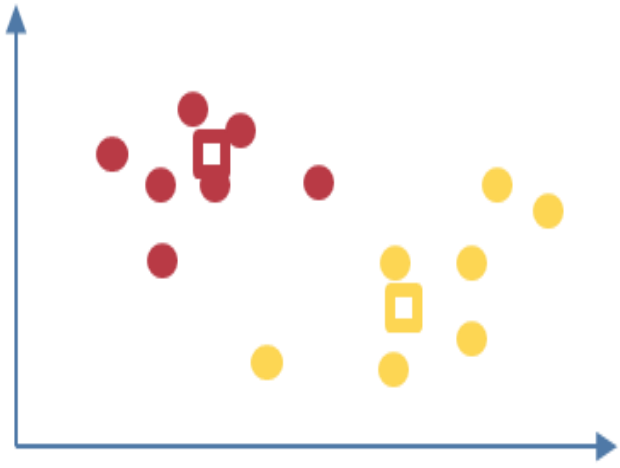

Figure 98. Illustration of the principle of clustering by K-Means

The demographic variables are of two kinds:

- endogenous (gender, socio-professional category) age does not intervene,
- exogenous: the regional location and the annual turnover segment.

It would have been interesting to study the effects of variables such as "workforce in the company", turnover, health consumption, etc.

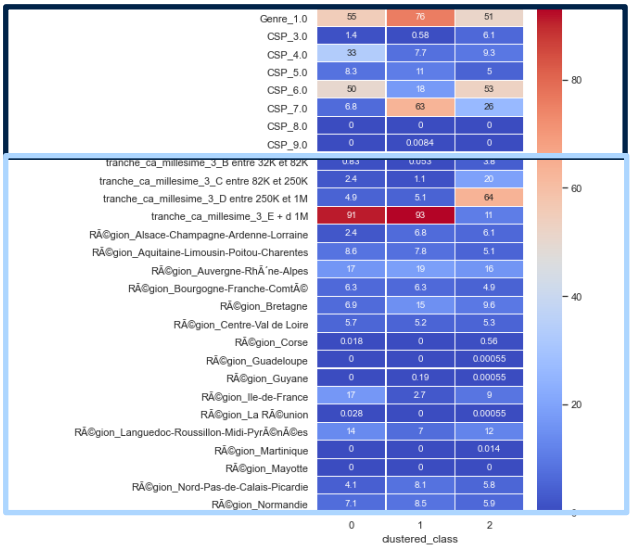

Figure 99. Partitioning with k = 3 with endogenous (framed in black) and exogenous (framed in blue) variables

The application of the pricing formula makes it possible to reconstitute the pricing curves for each of the determined clusters and the tariff for any cluster combined. Visually, it seems relevant to distinguish risk classes.

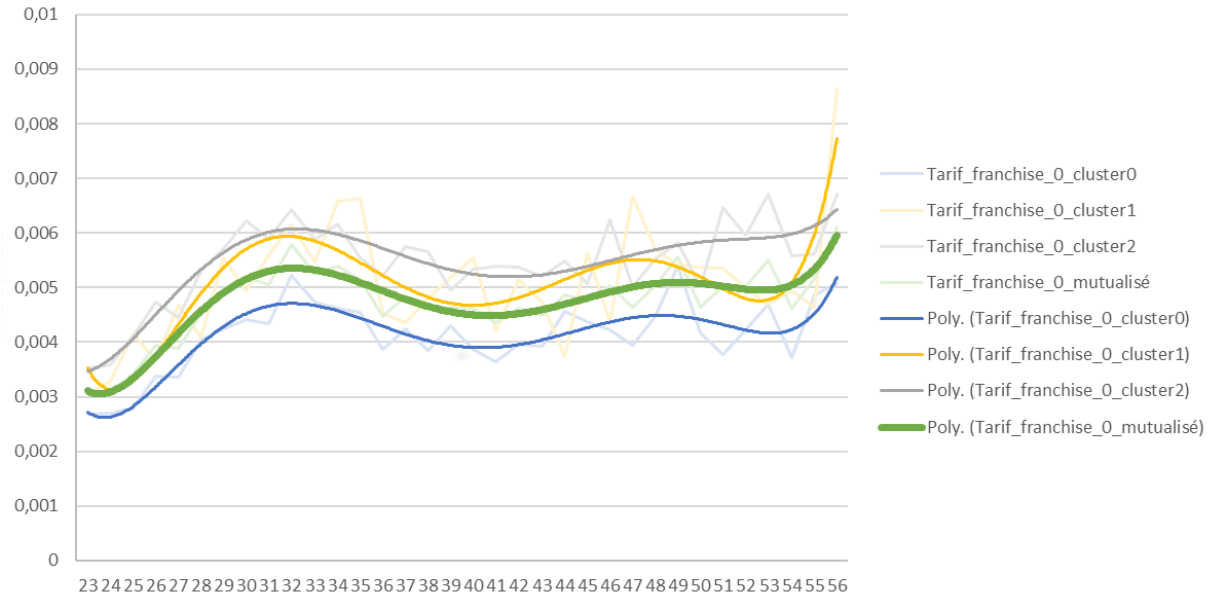

Figure 100. *Pricing*