Mémoire présenté le :

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

Par : Jimmy Barriere

Titre Recherche de nouveaux axes discriminants pour la tarification en Arrêt de Travail

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise : NATIXIS ASSURANCES

Nom : BIENAIME Nathalie

Signature :

Directeur de mémoire en entreprise :

Nom : Anna Kryshtopenko

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Natixis Assurances
Paris, 2019-2020

Recherche de nouveaux axes discriminants pour la tarification en Arrêt de Travail

Jimmy Barriere

Institut de Science Financière et d'Assurances
Mémoire d'actuariat

Tutrice en entreprise : Anna Kryshtopenko

Tuteur académique : Nabil Kazi-Tani

Avant-propos

Pour des raisons de confidentialité, les résultats chiffrés de ce mémoire ont été modifiés sans perte de généralisation.

L'ensemble des calculs de ce mémoire ont été réalisés à l'aide du langage R.

Remerciements

Je remercie le corps enseignant de l'ISFA, et plus particulièrement Mr. Kazi-Tani pour sa relecture et ses retours constructifs.

Je souhaite également remercier toute l'équipe du service Actuariat Produits Prévoyance de Natixis Assurances pour leur accueil et pour le cadre de travail qu'ils ont su me fournir. J'ai été très heureux d'en faire partie, de mon expérience cette année et je conserverai d'excellents souvenirs.

Je tiens à remercier mon père pour son soutien sans faille, pour la culture du travail et pour le goût des études qu'il a su me transmettre.

Je remercie également ma mère pour sa bienveillance, pour sa bonne humeur, notre complicité et pour tous ses encouragements.

J'aimerais remercier Patrick Gomis, responsable du service Actuariat Produits Prévoyance, pour sa présence chaleureuse, pour tous ses conseils métiers, et pour avoir su me mettre en confiance rapidement dans ce nouvel environnement.

Enfin, je souhaite remercier Anna Kryshtopenko, ma tutrice en entreprise. Merci pour ton adaptabilité, pour ces nombreuses missions intéressantes, pour tes capacités d'écoute et de compréhension, pour tes critiques constructives et pertinentes, pour ces *bi-weekly*, pour ton exigence, et plus globalement pour ton encadrement tout au long de cette aventure.

Résumé

Dans le processus de tarification en arrêt de travail, il est question d'utiliser des lois d'incidence en incapacité pertinentes afin de capter les sensibilités du portefeuille d'assurés. En 2020, Natixis Assurances distribue à travers le réseau du Groupe BPCE un nouveau produit d'assurance intitulé *Prévoyance Pro*. Dans le cadre de la mise en place de ce produit, des lois d'incidence en incapacité ont été utilisées et nous essayons d'en vérifier la pertinence dans ce mémoire.

Dans un même temps, nous tentons de mettre en lumière de nouveaux axes discriminants innovants pour la tarification comme par exemple la région d'habitation de l'assuré. Pour ce faire, nous utilisons différents algorithmes de classification et de prédiction et mettons les résultats obtenus en perspective. Dans un second temps, nous détaillons les étapes de la modélisation au cours de laquelle nous utilisons des modèles de survie, comme par exemple l'estimateur de Kaplan-Meier ou le modèle de Cox à risques proportionnels. Enfin, nous introduisons des notions de base de tarification pour calculer une nouvelle grille tarifaire. Nous la comparons avec la grille tarifaire déjà existante et discutons de sa pertinence.

Mots clés – incapacité, invalidité, arrêt de travail, tarification, taux d'incidence, apprentissage supervisé, GLM, apprentissage non-supervisé, méthode des K-Means, analyse factorielle des données mixtes, modèles de survie, estimateur de Kaplan-Meier, modèle de Cox, lissage de Whittaker-Henderson.

Abstract

When it comes to fixing a price to insurance contracts, it is often that one uses internal modelled laws to best modelize the behavior of the insurance portfolio. In 2018, *Natixis Assurances* commercializes via *Groupe BPCE* network a new insurance product named *Prévoyance Pro*. Since it is offering a protection for people having to go on sick leaves, it has to use inability entrance laws. In this thesis, we try to verify their relevance.

At the same time, we also try to identify new variables that could have an impact on likeliness to go on sick leaves, such as the location of the insuree. To do so, we are using both clustering and predictive algorithms. In a second part, we detail all the modelization steps that we are facing, using survival models, Kaplan-Meier estimator and the Cox model. Eventually, we introduce pricing notions and try to calculate new insurance prices that we compare with the already existing ones.

Keywords – inability, sick leaves, pricing, supervised learning, unsupervised learning, GLM, K-Means, survival models, Kaplan-Meier estimator, Cox model, Whittaker-Henderson.

Table des matières

1 Introduction **1**

2 Contexte général **2**
2.1 Natixis Assurances . . . 2
2.2 La prévoyance individuelle . . . 3
2.2.1 Définition du risque Arrêt de Travail . . . 3
2.2.2 Présentation des produits concernés par l'étude . . . 4

3 Préparation de la base de travail **7**
3.1 Les bases historisées et l'incidence en AT . . . 7
3.2 Pré-traitement des données . . . 8
3.2.1 Retraitement des bases et jointure . . . 9
3.2.2 Mise à la maille Client . . . 9
3.2.3 Validation avec les données comptable . . . 11
3.3 Travail sur les variables explicatives . . . 12
3.3.1 Ajout des variables géographiques . . . 13
3.3.2 Travail sur la CSP . . . 16

I Identification de variables discriminantes **19**

4 Analyse descriptive **20**
4.1 Statistiques descriptives . . . 20
4.2 Conclusion . . . 27

5 Méthodes d'apprentissage non-supervisé **28**
5.1 Méthodes d'apprentissage non-supervisé . . . 28
5.1.1 Classification Ascendante Hiérarchique . . . 28
5.1.2 Méthode des K-Means . . . 30
5.1.3 Analyse en Composantes Principales . . . 31
5.1.4 Analyse Factorielle des Données Mixtes . . . 32
5.2 Réduction des modalités des variables qualitatives . . . 33
5.2.1 Région . . . 33
5.2.2 Banque . . . 37
5.2.3 Catégorie Socio-Professionnelle . . . 40
5.2.4 Conclusion . . . 42
5.3 Application à la création de groupes d'individus similaires . . . 43
5.3.1 Analyse Factorielle des Données Mixtes . . . 43
5.3.2 Classification et analyse des groupes d'individus . . . 49
5.3.3 Conclusion . . . 60

6 Méthodes d'apprentissage supervisé **61**
6.1 Les Modèles Linéaires Généralisés . . . 61
6.2 Application du modèle . . . 64
6.2.1 Base de données, objectifs et choix du modèle . . . 64
6.2.2 Processus et retraitements . . . 65
6.2.3 Conclusion . . . 68

II Modélisation **70**

7 Introduction **71**
7.1 Définition des objectifs . . . 71
7.2 Création des groupes de risque . . . 72

8 Modélisation pour le groupe de risque neutre **75**
8.1 Calcul des taux d'incidence bruts . . . 75
8.1.1 Présentation de l'estimateur de Kaplan-Meier . . . 75
8.1.2 Création de la base de travail . . . 77
8.1.3 Application . . . 79
8.2 Lissage des taux bruts . . . 79
8.2.1 Présentation de la méthode de Whittaker-Henderson . . . 80
8.2.2 Application . . . 81
8.3 Validation de la loi . . . 82
8.3.1 Calcul des observés sur attendus . . . 82
8.3.2 Ré-apprentissage sur la base de données complète . . . 83

9 Positionnement des taux d'incidence **84**
9.1 Présentation du modèle de Cox à risques proportionnels . . . 84
9.2 Application . . . 85

III Tarification **89**

10 Introduction à la tarification **90**
10.1 Description de l'offre . . . 90
10.2 Méthode de tarification . . . 91

11 Ventilation des taux d'incidence **93**
11.1 Distinction Accident / Maladie . . . 93
11.2 Interpolation à toutes les durées de franchise . . . 95

12 Calcul des tarifs finaux **99**
12.1 Calcul d'une nouvelle grille tarifaire . . . 99
12.2 Application au portefeuille d'assurés . . . 101

IV Conclusion **102**

Bibliographie **105**

Liste des Figures **106**

Liste des Tables **108**

1 Introduction

Comme dans une perpétuelle remise en question, une gamme de produits d'assurance est toujours vouée à évoluer. En 2020, un produit intitulé *Prévoyance Pro* est pour la première fois distribué à travers le réseau des Banques Populaires et Caisses d'Épargne. Développé par Natixis Assurances, filiale du Groupe BPCE, ce produit d'assurance propose aux travailleurs non-salariés volontaires une couverture des risques d'incapacité et d'invalidité. L'objectif est simple : proposer un produit compétitif, complet, en phase avec les objectifs du groupe.

Plusieurs étapes précèdent la mise en place d'un nouveau produit. De l'émergence d'une idée à la commercialisation, plusieurs équipes se relaient pour s'assurer de la rentabilité et de la pertinence du produit. Dans le cadre de la construction d'un produit d'arrêt de travail, il est question d'utiliser plusieurs lois d'expérience. Entre autres, l'entreprise d'assurance émet le besoin d'avoir une loi d'entrée en incapacité cohérente, sensée, et qui parvient à capter les différentes sensibilités du portefeuille d'assurés.

C'est dans ce contexte que s'inscrit notre démarche. Datant de 2018, les dernières lois d'incidence utilisées par Natixis Assurances doivent être contrôlées et suivies pour s'assurer de leur pérennité. L'objectif du mémoire est simple : nous souhaitons effectuer une nouvelle modélisation de ces lois d'incidence. Bien entendu, nous essaierons d'avoir une approche innovante.

À date, la plupart des lois d'incapacité sont construites par âge et par type de franchise. Dans le cadre de notre étude, nous rechercherons dans un premier temps de nouvelles variables discriminantes. Celles-ci pourraient être : la profession de l'assuré, son ancienneté en portefeuille, sa région d'habitation, l'établissement de souscription du contrat ou encore le fait que celui-ci ait subi une sélection médicale.

Après avoir introduit l'activité de Natixis Assurances et du Groupe BPCE dont elle fait partie, nous procéderons en trois grandes étapes :

Dans une première partie, nous détaillerons le processus de recherche et d'identification de variables discriminantes en incapacité. L'objectif est de mettre en lumière de axes innovants pour la tarification.

Dans une seconde partie, nous détaillerons le processus de modélisation des lois d'incidence. À cette étape, les variables discriminantes ont été identifiées et nous passons en revue les différents modèles utiles à la mise en place des taux.

Dans une troisième et dernière partie, nous détaillerons le processus de tarification. Ici, l'objectif est de calculer une nouvelle grille de tarifs pour les assurés du portefeuille et de discuter de sa pertinence en la comparant avec la grille tarifaire existante.

2 Contexte général

Natixis est un établissement financier français de dimension internationale spécialisé dans la gestion d'actifs et de fortune, la banque de financement et d'investissement, l'assurance et les paiements. Filiale du Groupe BPCE, deuxième acteur bancaire en France à travers ses réseaux Banque Populaire et Caisse d'Epargne, Natixis compte près de 16 000 collaborateurs dans 38 pays.

Elle accompagne et conseille sa propre clientèle d'entreprises, d'institutions financières et d'investisseurs institutionnels, ainsi que les clients des réseaux du Groupe BPCE. Pour cela, elle mobilise les expertises complémentaires de ses métiers cœurs : la Gestion d'actifs et de fortune, la Banque de Grande Clientèle, l'Assurance, les Paiements.

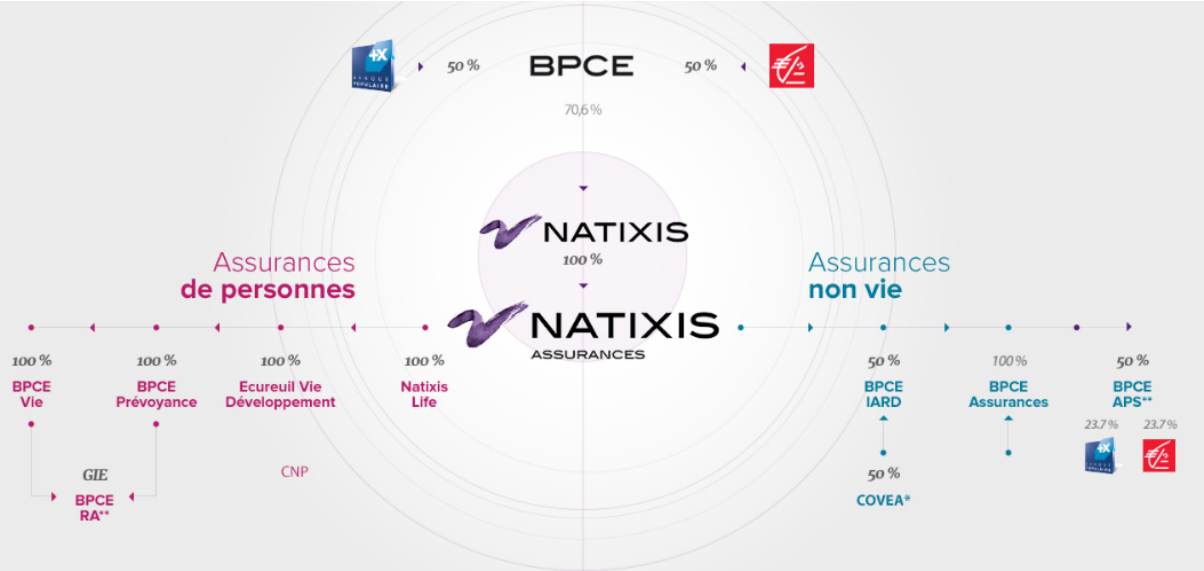

Figure 2.1: Organigramme du Groupe BPCE et de Natixis Assurances

2.1 Natixis Assurances

Natixis Assurances est le pôle Assurances de Natixis au service des réseaux du Groupe BPCE et de leurs clients. Elle conçoit et gère une offre complète de produits et services d'assurances réparties dans deux activités principales :

— Assurances de personnes : assurance vie, épargne, transmission de patrimoine, retraite, assurance décès, assurance dépendance et assurance des emprunteurs.

— Assurances non-vie : Assurance automobile, multirisque habitation, complémentaire santé, garantie des accidents de la vie, assurance des équipements multimédias, protection juridique, assurance para-bancaire, télésurveillance et assurances des professionnels.

L'année 2019 a marqué pour Natixis Assurances une étape importante avec le déploiement de projets stratégiques majeurs. En Assurances de personnes, une nouvelle offre a été lancée en Prévoyance individuelle dans le réseau des Banques Populaires : des produits simplifiés, digitalisés et plus compétitifs.

Dans un nouvel environnement à taux bas, les mesures d'adaptation engagées depuis plusieurs années par Natixis Assurances ont permis de maintenir un niveau de solvabilité et de rentabilité satisfaisant. Et, à compter de 2020, Natixis Assurances assurera les affaires nouvelles d'assurance non-vie des clients particuliers des Banques Populaires en complément de ceux des Caisses d'Epargne.

2.2 La prévoyance individuelle

Natixis Assurances compte deux filiales : BPCE Vie et BPCE Prévoyance qui sont également des sociétés d'assurance. L'activité principale de ces deux entités consiste en la couverture des risques de prévoyance comme le décès, l'arrêt de travail, les accidents de la vie, les obsèques, l'assurance des emprunteurs et la dépendance en Prévoyance Individuelle.

2.2.1 Définition du risque Arrêt de Travail

Dans le cadre de ce mémoire, nous concentrerons notre étude plus particulièrement sur le risque Arrêt de Travail, que nous noterons parfois simplement *AT*. On retrouve 3 types d'arrêts différents que nous définissons de la manière suivante :

— Incapacité Temporaire Totale (ITT) : un assuré entre en ITT lorsqu'à la suite d'un accident ou d'une maladie, il se trouve dans l'impossibilité absolue d'exercer son activité professionnelle. Si ce-dernier peut exercer son activité professionnelle à temps partiel, alors cela ne constitue pas un état d'ITT. Cet état ne peut se poursuivre au-delà de trois ans, temps à partir duquel l'assuré entre dans un état d'Incapacité Permanente et Totale de Travail (IPT). Les conséquences de l'accident ou de la maladie deviennent ainsi permanentes et sont présumées définitives.

— Invalidité Permanente et Totale de Travail (IPT) : un assuré entre en IPT si, après une période en ITT à la suite d'un accident ou d'une maladie, son état se consolide et le maintient dans l'impossibilité permanente et totale d'exercer toute activité professionnelle (y compris de direction ou de surveillance) susceptible de lui procurer gain ou profit. Son taux d'incapacité permanente, qui est fixé en fonction de son taux d'incapacité fonctionnelle et de son taux d'incapacité professionnelle, est supérieur ou égal à 66%.

— Invalidité Permanente Partielle (IPP) : un assuré est considéré en état d'IPP si, à la date de consolidation de son état, il reste dans l'impossibilité totale d'exercer toute activité professionnelle. Son taux d'incapacité permanente, qui est fixé en fonction de son taux d'incapacité fonctionnelle et de son taux d'incapacité professionnelle, est supérieur ou égal à 33% et inférieur à 66%.

2.2.2 Présentation des produits concernés par l'étude

Afin de parvenir à modéliser correctement une loi d'incidence en incapacité, il est important de bien connaître les différentes modalités définies par les produits qui constituent notre portefeuille et qui régissent le comportement des assurés. C'est pourquoi nous vous présenterons dans cette partie les différents produits dont il est question.

En 2020, Natixis Assurances a lancé un nouveau produit sur le marché intitulé : *Prévoyance Pro*. C'est une nouvelle offre qui propose, entre autres, une couverture du risque *AT*. Pour pouvoir mener à bien notre étude, ce produit n'a pas été considéré. Il a été proposé aux assurés seulement très récemment, l'exposition observable est trop faible et l'ancienneté sur le marché trop courte pour nous permettre de calibrer des modèles fiables sur l'incidence en Arrêt de Travail. Il sera néanmoins intéressant de comparer nos résultats issus de la phase de tarification avec les études menées en vue de l'introduction sur le marché du nouveau produit.

Nous concentrerons donc notre étude sur le reste des produits du portefeuille, qui permettent d'avoir une bonne vision et une excellente compréhension de la sinistralité en AT. Enfin, nous focaliserons notre étude sur le réseau des Banques Populaires dont l'historique de données est le plus large et le plus fiable.

Les axes de recherches du mémoire concernent des caractéristiques propres au client comme son âge, sa catégorie socio-professionnelle, sa région d'habitation, la banque à laquelle il a souscrit son contrat sans prendre en compte caractéristiques liées aux produits. Les résultats présentés sur le risque d'incidence seront ainsi généralisables à tous les individus du portefeuille indépendamment du type de contrat auquel ils ont souscrit ou du réseau dont ils font partie.

Nous présentons maintenant les différents produits porteurs du risque *AT* au sein du portefeuille des Banques Populaires.

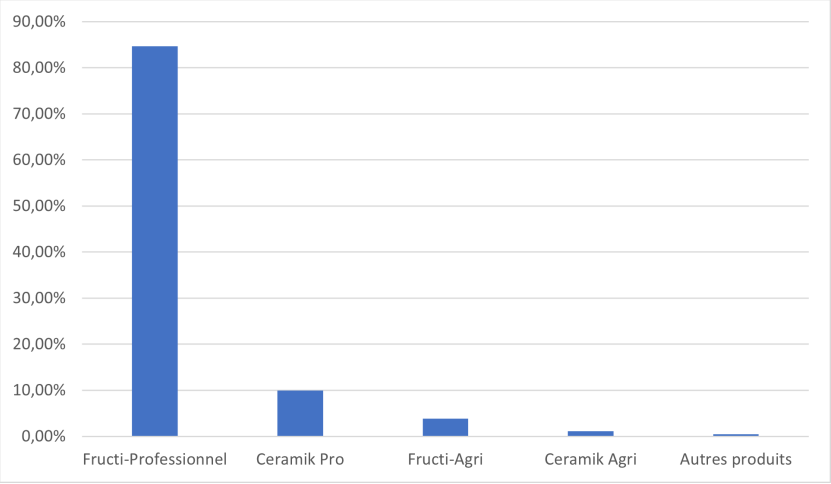

Figure 2.2: Distribution des produits au sein du portefeuille des Banques Populaires

Fructi-Professionnel

C'était jusqu'à très récemment le produit principal proposé par l'entreprise pour la couverture du risque *AT*. Aujourd'hui, le produit est en « run-off » [1], bien qu'il représente toujours environ 85% des clients en portefeuille. Il a récemment été remplacé par son homologue *Prévoyance Pro*.

Fructi-Professionnel est un contrat d'assurance de groupe à adhésion facultative ayant pour objet le versement d'une indemnité journalière payable mensuellement en cas d'arrêt de travail temporaire et/ou définitif (option ITT et option IPT). Le contrat est étbli pour un an, renouvelable par tacite reconduction au 1er janvier. Les risques assurés sont l'accident ou l'accident et la maladie. Les limites d'âge minimale et maximale à l'adhésion sont respectivement de 18 ans et 54 ans (version accident et maladie) ou 59 ans (version accident seul).

En fonction de l'offre de couverture choisie par le client, trois types de franchises sont proposés :

— Franchise absolue [2] 7 jours pour la couverture Accident seul.
— Franchise absolue 7 jours pour l'Accident et 15 jours pour la Maladie.
— Franchise relative [3] 30 jours pour l'Accident et la Maladie.

Outre les caractéristiques propres aux individus, l'étude devra tenir compte de la durée de franchise. En effet, il semble intuitivement que, plus la durée de franchise est élevée, moins le risque d'arrêt est grand pour une même couverture de risque.

Fructi-Agri

Bien que le produit ne soit aujourd'hui plus commercialisé et que l'exposition générale soit assez faible, celui-ci représente une part non-négligeable du portefeuille d'assurés.

Fructi-Agri est un contrat d'assurance de groupe à adhésion facultative ayant pour objet le versement d'une indemnité journalière en cas d'incapacité temporaire totale (ITT). À l'instar de son homologue *Fructi-Professionnel*, ce produit s'adresse aux chefs d'exploitations et d'entreprises agricoles. La durée du contrat est d'un an, renouvelable par tacite reconduction au 1er janvier.

Les risques assurés sont l'accident ou l'accident avec ou sans la maladie. Les limites d'âge minimale et maximale à l'adhésion sont respectivement de 18 ans et 54 ans (version accident et maladie) ou de 59 ans (version accident seul). Les prestations ainsi que les franchises varient également en fonction du profil de l'assuré et de la couverture choisie.

1. C'est à dire qu'il n'est plus commercialisé
2. Franchise absolue : l'assureur ne paye que les jours de dépassement de la franchise.
3. Franchise relative : l'assureur paye pour tous les jours d'arrêt dès lors que la franchise est dépassée.

Ceramik Pro

Le produit a été lancé en 2017 pour remplacer *Fructi-Professionnel*. Après trois ans, le produit représente finalement environ 10% des individus en portefeuille, ce qui caractérise une bonne commercialisation.

Ceramik Pro est un contrat d'assurance de groupe à adhésion facultative, éligible à la loi Madelin, qui prévoit le versement d'une indemnité journalière en cas d'Arrêt de Travail. Il existe trois garanties : une garantie ITT, une garantie IPT, une garantie IPP ou IPT. Comme pour les autres produits, les risques assurés sont l'accident ou l'accident et la maladie. Les prestations et les franchises varient en fonction de ces variables ainsi que du profil de l'assuré, qui remplit un questionnaire de sante simplifié à l'adhésion.

Ceramik Agri

Comme la gamme de produit *Fructi*, les produits *Ceramik* proposent aussi une alternative pour les travailleurs agricoles qui représentent population visée.

De la même manière que son homologue, le produit propose un contrat d'assurance de groupe à adhésion facultative, qui prévoit le versement d'une indemnité journalière en cas d'Arrêt de Travail pour 3 garanties différentes : ITT, IPT, IPP ou IPT. Les caractéristiques terminales du produit portant sur les prestations ou les franchises sont globalement les mêmes que celles du produit *Ceramik Pro*.

3 Préparation de la base de travail

Dans ce mémoire, nous allons tenter de comprendre et de mettre en évidence des caractéristiques qui seraient discriminantes pour l'incidence en Arrêt de Travail. Pour ce faire, nous avons à notre disposition des données sur les clients qui constituent notre portefeuille. Néanmoins, il faut tenir compte de plusieurs difficultés :

— Les données dont nous disposons sont à la mauvaise échelle. Nous aimerions avoir une base de donnée en vision client, puisque nous cherchons après tout à déterminer des caractéristiques propres aux clients.
— Ces données présentent aussi parfois des erreurs et des valeurs aberrantes. Il faudra les corriger et valider la base de données.
— Enfin, ces données ne sont pas complètes. Nous utiliserons parfois plusieurs bases pour avoir accès à toutes les informations souhaitées.

Dans cette première partie, nous vous présenterons le schéma de construction de notre base de travail finale et les différentes étapes qui ont permis de surmonter les problèmes présentés ci-dessus.

3.1 Les bases historisées et l'incidence en AT

Au sein du service en prévoyance Individuelle, les bases historisées « Solvency » sont souvent utilisées lorsqu'il s'agit de modéliser ou d'effectuer le suivi des différentes lois d'expérience au sein du service. Ces bases de données sont extraites trimestriellement par le service informatique et offrent une vision du stock à la date d'extraction du portefeuille d'assurés en prévoyance individuelle.

Ces bases historisées contiennent plusieurs variables qui comportent des informations sur les caractéristiques des assurés, leurs contrats et les garanties souscrites. En particulier, la variable *date de survenance* retiendra toute notre attention puisqu'elle nous de prendre connaissance des sinistres survenus au cours d'une période définie.

L'apparition d'un sinistre dans les bases historisées est transcrite à plusieurs étapes. Premièrement, l'assuré doit, à la suite d'un accident ou d'une maladie (en fonction de ses garanties), rester en arrêt de travail pour une durée supérieure à la durée de franchise de son contrat. Lorsque cela se produit et seulement à partir de ce moment, celui-ci déclare son sinistre ; la variable date de survenance est renseignée. Il est possible que le sinistre soit refusé par l'organisme d'assurance si certains évènements sont contestables ou ne correspondent pas aux termes du contrat.

Dans un contexte de tarification, nous prendrons en compte tous les sinistres, refusés ou non. Cette façon de procéder pour calculer une nouvelle prime d'assurance permet par définition de bénéficier d'une marge de prudence implicite ; elle est couramment employée.

Maintenant que nous avons présenté les données dont nous disposons, passons à l'étape suivante étape : celle du traitement des données.

3.2 Pré-traitement des données

Les bases historisées présentent une vision de stock à une date d'extraction précise. Or, il peut arriver que pour un client, un contrat et une garantie précise donnés, faisant l'objet de plusieurs sinistres en arrêt de travail, seul le dernier soit renseigné sur la ligne correspondante. En découle un problème de comptage puisqu'un nombre important de sinistres peut de ce fait ne pas être comptabilisé.

La première étape du pré-traitement des données consistera donc en une jointure de toutes les bases « Solvency » extraites à différentes dates d'extraction entre 2016 et 2020. À raison d'une base par trimestre il faut effectuer une jointure entre seize bases historisées ayant chacune quatre millions de lignes environ et cinquante-sept colonnes. Comme expliqué précédemment, les bases historisées extraites trimestriellement que nous avons à notre disposition présentent une vision par garantie de notre portefeuille d'assurés.

client <dbl>	contrat <dbl>	garantie <dbl>
5273	6668	21
5273	6962	21
12489	23768	21
12489	23768	30
12489	23768	42
12489	23768	49

Figure 3.1: Les trois premières colonnes de la base solvency

Ainsi, un client apparaît autant de fois que le nombre de garanties qui se trouvent dans chaque contrat auxquels il a souscrit. Ce n'est pas la vision qui convient à notre étude puisque l'objectif est de mettre en avant les caractéristiques propres à chaque client.

Dés lors, l'une des étapes du pré-traitement des données consistera en un dédoublonnage pour mettre la base sous une forme qui réponde à nos attentes. Enfin, il conviendra de valider cette base de travail finale en utilisant une base de données des flux comptables.

3.2.1 Retraitement des bases et jointure

Comme indiqué précédemment, la première étape de préparation de notre base de travail consiste en une jointure des bases Solvency à différentes dates d'extraction dans le temps. Toutefois, avant d'effectuer ces jointures, nous procédons au retraitement de chacune des bases historisées. C'est à dire que :

- Le format des fichiers d'origine est modifié pour permettre une lecture et une sauvegarde plus rapide des données.
- Nous ajoutons des variables non-déjà présentes dans les bases comme le libellé produit (*Fructi-Pro, etc ...*), le réseau (*BP ou CE*), la franchise (*Absolue 30 jours, etc ...*).
- Nous retraitons chacune des variables de la base en redéfinissant leur type et en effectuant des modifications pour les rendre plus lisibles et compréhensibles.

Après avoir effectué ces retraitements, nous filtrerons les données sur celles pour lesquelles le risque couvert est le risque *AT*. Enfin, nous sélectionnerons chacune des bases « Solvency » retraitées et effectuerons la jointure. Il convient également de dédoublonner la base finale obtenue selon la clé client, contrat, garantie, date de survenance pour s'assurer de l'unicité de chacune des lignes.

Nous obtenons au final une base de 91 311 lignes, contre quatre millions dans une base « Solvency » d'origine. Cette réduction s'explique simplement par le fait que l'Arrêt de Travail n'est pas le principal risque couvert au sein du portefeuille global. Ainsi, nous diminuons significativement notre population. L'exposition reste néanmoins tout à fait suffisante pour la réalisation de notre étude.

3.2.2 Mise à la maille Client

La base obtenue dans la précédente sous-partie est une base dédoublonnée à la maille client, contrat, garantie, date de survenance. L'objectif final de cette partie est d'obtenir une base à l'échelle année d'observation, client.

Cela signifie que si un client entre en portefeuille le 15 mars 2014, fait l'objet d'une ITT en 2015 et résilie son contrat en avril 2016, nous aimerions avoir les informations sous cette forme :

annee_obs <int>	client <dbl>	coeff_expo <dbl>	nb_sin <dbl>
2014	150820	0.8	0
2015	150820	1.0	1
2016	150820	0.3	0

Figure 3.2: Base à la maille année d'observation / client

Le client apparaît autant de fois qu'il a d'années de présence en portefeuille. Pour chaque année d'observation, en plus des variables qui caractérisent le client, nous en comptons aussi deux nouvelles :

- Le coefficient d'exposition, qui correspond au temps de présence de l'individu pour une année d'observation donnée. Si l'individu était présent du 1er janvier au 31 décembre de l'année N alors son coefficient d'exposition vaut 1.
- Le nombre de sinistres, qui correspond au nombre de sinistres en Arrêt de Travail pour l'individu pour une année d'observation donnée.

Pour mettre notre base de travail sous la forme voulue, nous allons passer plusieurs étapes. Dans un premier temps, il s'agit de rajouter la variable *année d'observation*. Pour ce faire, nous allons d'abord calculer le coefficient d'exposition de chaque client dans la base entre les années 2014 et 2019 ; puis rajouter la variable année d'observation ; enfin ne conserver pour chaque année que les clients ayant un coefficient d'exposition supérieur à 0.

Nous avons maintenant une base à la maille : année d'observation, client, contrat, garantie, date de survenance. Notre objectif final est d'avoir une base à la maille année d'observation, client.

Pour ce faire, nous allons effectuer un dédoublonnage en trois temps :

- Nous dédoublonnerons d'abord à la maille année d'observation, client, contrat, date de survenance.
- Nous dédoublonnerons ensuite à la maille année d'observation, client, contrat.
- Nous dédoublonnerons enfin à la maille année d'observation, client.

Les raisons de ce dédoublonnage en 3 temps sont simples. Si nous considérons le montant des primes par exemple : pour un couple unique *client, contrat* fixé, il est renseigné sur chaque ligne (soit pour chaque garantie) un même montant de prime unique. C'est normal puisque c'est le montant de prime qui a été fixé pour le contrat total. Ainsi, quand nous effectuerons la première étape du dédoublonnage, nous conserverons l'unique montant de prime renseigné.

En revanche, à la dernière étape de dédoublonnage, nous sommerons les montants de prime puisque la prime totale payée par un individu correspond bien à la somme des montants des primes de chacun de ses contrats.

Après retraitements, la base finale a l'allure suivante :

	annee_obs	client	reseau	sexe	csp	mnt_gar	banque	code_postal	mnt_prest	mnt_prime_ann	mnt_prime_per	sel_med	per_pcht	coeff_expo
1:	2014	482860	BP	H	30	1378.28	39	1340	0	436.76	109.19	N	4	1
2:	2014	516368	BP	H	0	2033.12	8	21130	0	696.04	174.01	N	4	1
3:	2014	649448	BP	H	0	677.68	78	46800	0	232.00	116.00	N	2	1
4:	2014	740339	BP	F	0	1400.18	19	63420	0	479.36	239.68	O	2	1
5:	2014	746575	BP	H	0	1328.74	38	50260	0	454.92	37.91	N	12	1

113971:	2019	5023636	BP	H	NA	1079.24	46	26790	0	194.28	16.19	N	12	1
113972:	2019	5078032	BP	H	NA	539.61	46	13570	0	97.08	8.09	N	12	1
113973:	2019	5086760	BP	H	NA	1079.24	552	73210	0	336.72	28.06	O	12	1
113974:	2019	5340324	BP	F	NA	10226.86	46	13210	0	155.16	12.93	N	12	1
113975:	2019	5677736	BP	H	NA	1049.39	46	84400	0	188.88	15.74	N	12	1

	nb_sin	nb_contrats	soc	entite	autres_prod	ceramik_agri	ceramik_pro	fructi_agri	fructi_pro	duree_franchise	anciennete	age
1:	0	1	124	BPCE Prev	0	0	0	1	0	715	13	44
2:	0	1	124	BPCE Prev	0	0	0	1	0	715	14	55
3:	0	1	124	BPCE Prev	0	0	0	1	0	715	15	55
4:	0	1	124	BPCE Prev	0	0	0	1	0	715	14	57
5:	0	1	124	BPCE Prev	0	0	0	1	0	715	12	60

113971:	0	1	124	BPCE Prev	1	0	0	0	0	7	4	62
113972:	0	1	124	BPCE Prev	1	0	0	0	0	7	4	44
113973:	0	1	124	BPCE Prev	1	0	0	0	0	7	4	51
113974:	0	1	124	BPCE Prev	1	0	0	0	0	7	3	61
113975:	0	1	124	BPCE Prev	1	0	0	0	0	7	2	50

Figure 3.3: Base après retraitement

Dans la prochaine sous-partie, nous validerons cette base finale de travail obtenue en utilisant les données comptables que nous avons à disposition. Par ailleurs, nous préciserons le rôle de chacune des variables observables ci-dessus dans une prochaine partie du mémoire.

3.2.3 Validation avec les données comptable

Afin de s'assurer de la significativité de notre étude et de la véracité de notre base de travail, il est important d'en vérifier les données en utilisant les flux comptables. Pour ce faire, nous allons comparer les montants de prestations versés pour chaque année entre 2014 et 2019. Nous n'accepterons que des écarts relatifs inférieurs à 5%.

Année	Prestations observées	Prestations comptables	Ecart (en %)
2014	3 233 094	3 269 006	1.1
2015	2 893 549	2 896 536	< 1
2016	2 526 159	2 522 855	< 1
2017	2 664 108	2 648 241	< 1
2018	2 409 656	2 468 499	2.3
2019	1 265 652	801 557	58

Table 3.1: Tableau comparatif avec les flux comptables (en euros)

Nous constatons que pour toutes les années inférieures à 2019, les flux de prestations observés correspondent bien aux flux comptables pour les années correspondantes. En 2019 néanmoins, nous observons un écart important entre les deux, et des valeurs de prestations plus faibles que les autres années. Néanmoins, cela est justifiable, en effet :

— Au moment de la validation des données, les flux comptables sont arrêtés au premier trimestre 2020. Or, à cause de l'étalement des versements de prestation, ainsi qu'à cause du temps de procédure, tous les flux comptables pour tous les sinistres de 2019 ne sont pas encore comptabilisés dans les flux comptables.

— Le point précédent est confirmé par le fait que les prestations observées sont bien supérieures aux prestations comptables.
— Globalement, les prestations sont plus faibles que les autres années. Néanmoins, cela vient encore du fait que les bases historisées comme les flux comptables sont arrêtés au premier trimestre 2020. Ainsi, nous n'avons pas encore connaissance de toutes les prestations futures.

Nous procédons aussi à la validation par l'étude des montants de prime :

Année	Prime observée	Prime comptable	Ecart (en %)
2016	6 095 290€	6 455 834€	5.58%
2017	6 531 603€	6 355 827€	2.8
2018	6 342 336€	6 132 193€	3.4
2019	6 259 372€	6 088 865€	2.8

Table 3.2: Tableau comparatif avec les flux comptables (en euros)

Nous constatons que les écarts sont bien tous inférieurs à 5% excepté pour 2016. Néanmoins, comme l'écart reste seulement très légèrement supérieur à 5% et que tous les autres montants (y compris les prestations) ont été validés, nous validerons tout de même notre base de travail.

Nous pouvons donc commencer à utiliser notre base de données et vous présenterons dans la partie suivante certains des derniers retraitements effectués.

3.3 Travail sur les variables explicatives

En plus des travaux précédemment effectués sur les variables, nous allons présenter dans cette partie deux phases importantes dans la création de notre base de données.

Premièrement, nous avons ajouté des variables géographiques selon une procédure que nous vous détaillerons. Ces variables sont utiles à notre étude puisqu'elles permettent de mettre en relation la sinistralité de l'assuré avec sa région d'habitation par exemple.

Ensuite, nous avons travaillé sur la complétude de la variable *catégorie socio-professionnelle* puisque celle-ci présentait un certain nombre de valeurs manquantes. Nous définirons cette variable et présenterons notre étude dans un deuxième temps.

3.3.1 Ajout des variables géographiques

Dans le cadre de la recherche de nouvelles variables discriminantes sur l'incidence en arrêt de travail, nous cherchons des caractéristiques qui sont propres aux clients. A priori, l'une de ces caractéristiques pourrait être la région d'habitation de l'assuré. En effet, il semble intéressant de voir si la zone de résidence pourrait avoir un impact sur la sinistralité. Nous disposons dans notre base de données deux variables qui permettent de localiser les assurés :

— La variable *code postal*, à laquelle nous pouvons rattacher un département ou une région.
— La variable *établissement*, qui transmet l'information sur l'entité bancaire auprès de laquelle l'assuré a souscrit son contrat.

Dans un premier temps, nous choisissons de représenter sur une carte la distribution de l'exposition par région au sein du portefeuille.

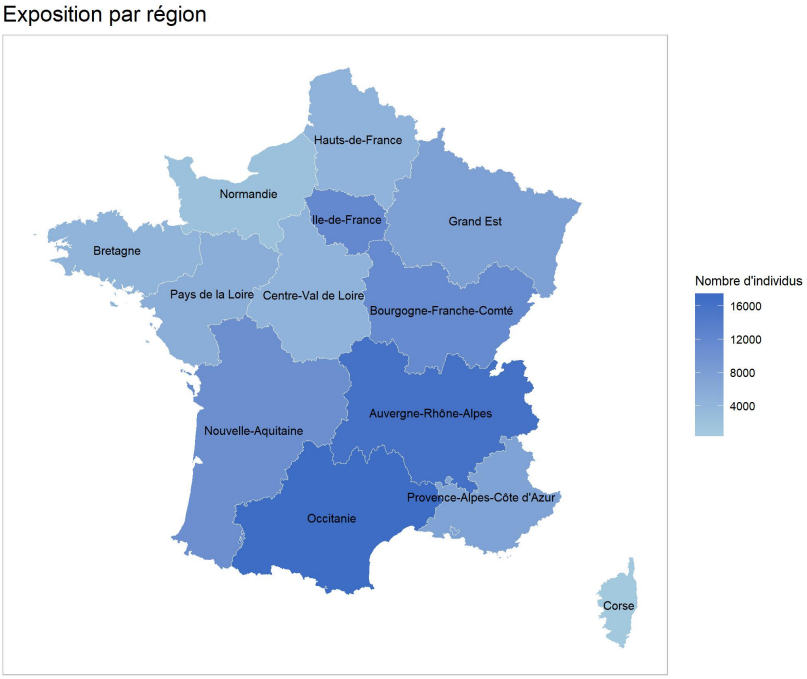

Figure 3.4: Exposition par région

Nous observons une exposition forte pour les régions Auvergne-Rhône-Alpes et Occitanie qui présentent environ 16 000 individus dans les deux cas. Des effectifs plus faibles sont observables, en particulier pour les régions du Nord et du Nord-Ouest de la France. Le territoire de la Corse reste toutefois le territoire le moins représenté avec légèrement moins de 1000 individus observés.

En ce qui concerne l'exposition par banque, une carte des Banques Populaires a été construite à partir des codes postaux des agences commerciales que nous avons regroupés par grande entité bancaire sous la forme de différents polygones géographiques. Nous traçons ainsi la carte de l'exposition par établissement :

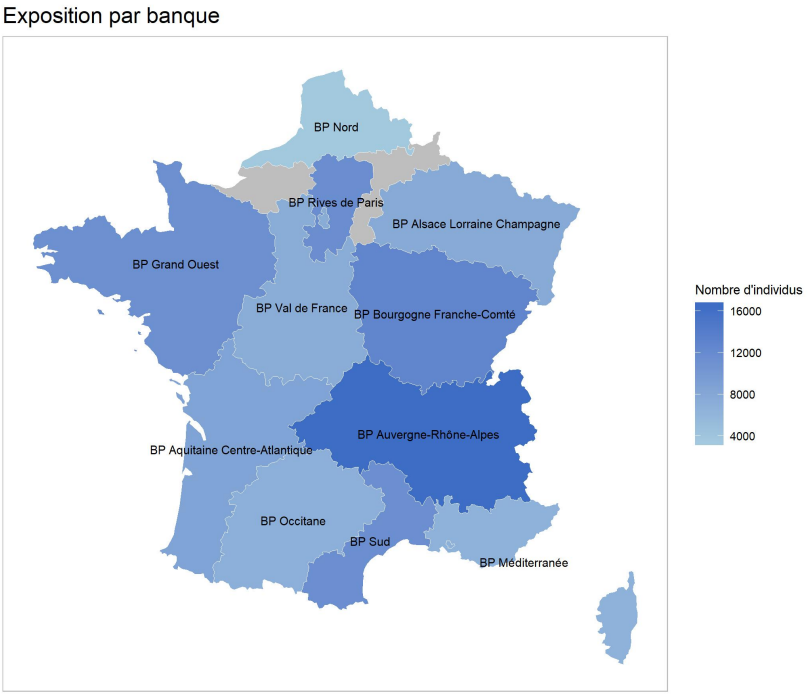

Figure 3.5: Exposition par banque

Nous trouvons dans notre base de travail les Banques Populaires suivantes : BP Nord, BP Grand Ouest, BRED [4], BP Rives de Paris, BP Alsace-Lorraine-Champagne, BP Val de France, BP Bourgogne-Franche-Comté, BP Aquitaine-Centre-Atlantique, BP Auvergne-Rhône-Alpes, BP Occitane, BP Sud, BP Méditerranée.

La zone grisée de la carte correspond à la banque BRED dont nous avons décidé de ne pas représenter l'exposition puisque celle-ci est très faible (de l'ordre de 300 observations). De plus, le tracé de son polygone est légèrement faussé puisqu'il est difficile d'associer à cette banque une zone géographique précise.

Autrement, nous observons, conformément à la distribution de l'exposition par région, que la Banque Populaire Auvergne-Rhône-Alpes est la plus représentée avec environ 16 000 individus. Le résultat sur la région Occitanie n'est quant à lui pas tout à fait retrouvé dans l'exposition de la BP Occitane puisque la région comprend en fait aussi la Banque Populaire Sud.

4. BRED : Banque régionale d'escompte et de dépôts

Par le même raisonnement, nous constatons que la Corse n'apparaît plus aussi peu représentée puisqu'elle est maintenant liée à la BP Méditerranée. Enfin, conformément aux résultats par région précédents, nous observons que la Banque Populaire du Nord présente l'exposition la plus faible avec seulement environ 4000 individus observés en portefeuille.

Pour une meilleure précision de nos données, nous étudions aussi le taux d'expatriation par banque. C'est à dire que nous pouvons représenter sur une carte le pourcentage de population qui a souscrit un contrat dans une certaine entité bancaire mais ne vit pas ou plus dans le polygone défini par cette entité bancaire. Voici le résultat obtenu :

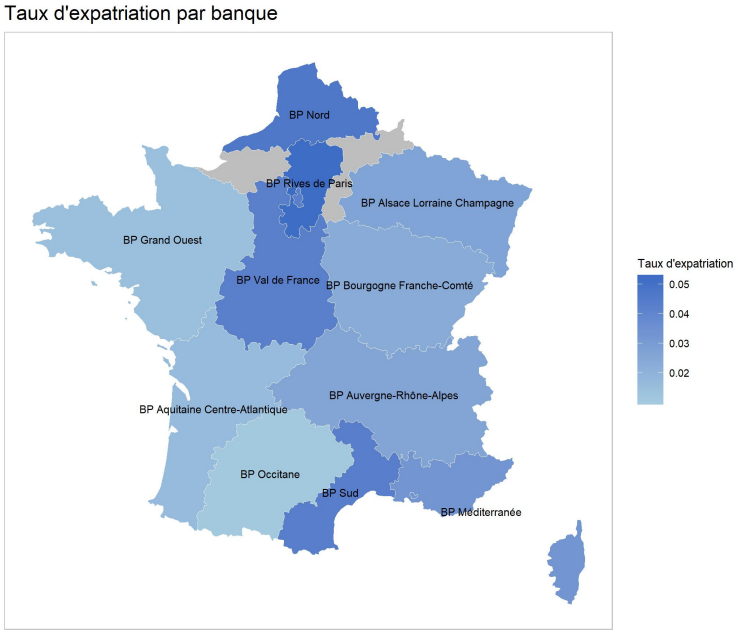

Figure 3.6: Taux d'expatriation par banque

Nous constatons que les taux d'expatriation sont relativement faibles (de l'ordre de 2 à 3% en moyenne). De plus, les Banques Populaires qui présentent un taux d'expatriation le plus élevé sont celles du Nord de la France.

Dans la prochaine sous-partie, nous parlerons de la Catégorie Socio-Professionnelle des individus présents en portefeuille.

3.3.2 Travail sur la CSP

La notion de Catégorie Socio-Professionnelle (CSP) a été introduite par l'INSEE [6] en 1954. L'objectif était de classer les individus selon leur situation professionnelle en tenant compte de plusieurs critères : métier, activité économique, qualification, position hiérarchique et statut. Elle comprenait neuf grands groupes subdivisés en 30 catégories socioprofessionnelles.

Cette classification a toutefois été abandonnée en 1982 et remplacée par la nomenclature des professions et catégories socioprofessionnelles (PCS), utilisée notamment pour les recensements de 1982, 1990, 1999.

Dans notre base de travail, nous disposons de la variable CSP sous sa forme originelle antérieure à 1982. Voici les différentes nomenclatures que nous observons :

	INDIVIDUS		MÉNAGES classés selon la CS du chef	
CSP	*Nombre (en milliers)*	*‰ de la population active* [1]	*Nombre (en milliers)*	*‰ du total des ménages* [1]
0. AGRICULTEURS EXPLOITANTS	1 448	62	800	41
00. Agriculteurs exploitants				
1. SALARIÉS AGRICOLES	304	13	175	9
10. Salariés agricoles				
2. PATRONS DE L'INDUSTRIE ET DU COMMERCE	1 737	74	1 156	59
21. Industriels	71	3	60	3
22. Artisans	573	24	464	24
23. Patrons pêcheurs	13	1	11	1
26. Gros commerçants	210	9	146	7
27. Petits commerçants	870	37	474	24
3. PROFESSIONS LIBÉRALES ET CADRES SUPÉRIEURS	1 810	77	1 409	72
30. Professions libérales	220	9	171	9
32. Professeurs, professions littéraires et scientifiques	479	20	292	15
33. Ingénieurs	347	15	319	16
34. Cadres administratifs supérieurs	764	32	627	32
4. CADRES MOYENS	3 254	138	1 872	95
41. Instituteurs, professions intellectuelles diverses	828	35	382	19
42. Services médicaux et sociaux	432	18	170	9
43. Techniciens	923	39	728	37
44. Cadres administratifs moyens	1 071	46	592	30
5. EMPLOYÉS	4 677	199	1 882	96
51. Employés de bureau	3 746	159	1 498	76
53. Employés de commerce	931	40	383	20
6. OUVRIERS	8 266	351	5 102	261
60. Contremaîtres	464	20	426	22
61. Ouvriers qualifiés	3 292	140	2 336	119
63. Ouvriers spécialisés	2 605	111	1 603	82
65. Mineurs	49	2	40	2
66. Marins et pêcheurs	33	1	25	1
67. Apprentis ouvriers	124	5	4	e
68. Manœuvres	1 699	72	668	34
7. PERSONNELS DE SERVICE	1 531	65	509	26
70. Gens de maison	214	9	43	2
71. Femmes de ménage	112	5	22	1
72. Autres personnels de service	1 205	51	444	23
8. AUTRES CATÉGORIES	498	21	370	19
80. Artistes	74	3	49	3
81. Clergé	61	3	22	1
82. Armée et police	363	15	299	15
POPULATION ACTIVE	23 525	1 000	13 273	678
POPULATION INACTIVE	30 748		6 317	322
d'âge ⩾15 ans	19 515		6 317	322
d'âge <15 ans	11 233		0	0
POPULATION TOTALE	*54 273*		*19 590*	*1 000*

FIGURE 3.7: Les catégories socio-professionnelles avant 1982 (source : INSEE)

Après 1982, les nomenclatures changent pour devenir les suivantes :

	INDIVIDUS		MÉNAGES classés selon la CS du chef	
PCS	*Nombre (en milliers)*	*‰ de la population active* [1]	*Nombre (en milliers)*	*‰ du total des ménages* [1]
1. AGRICULTEURS EXPLOITANTS	1 475	63	817	42
11. Agriculteurs sur petite exploitation	690	29	390	20
12. Agriculteurs sur moyenne exploitation	450	19	247	14
13. Agriculteurs sur grande exploitation	335	14	179	9
2. ARTISANS, COMMERÇANTS ET CHEFS D'ENTREPRISE	1 835	78	1 243	64
21. Artisans	904	38	669	34
22. Commerçants	797	34	460	23
23. Chefs d'entreprise de 10 salariés ou plus	134	6	113	6
3. CADRES ET PROFESSIONS INTELLECTUELLES SUPÉRIEURES	1 895	81	1 513	77
31. Professions libérales	239	10	182	20
33. Cadres de la fonction publique	244	10	197	10
34. Professeurs, professions scientifiques	353	15	227	12
35. Professions de l'information, des arts et des spectacles	117	5	78	4
37. Cadres administratifs et commerciaux d'entreprise	559	24	473	24
38. Ingénieurs et cadres techniques d'entreprise	383	16	356	18
4. PROFESSIONS INTERMÉDIAIRES	3 971	169	2 534	130
42. Instituteurs et assimilés	777	33	347	18
43. Professions intermédiaires de la santé et du travail social	613	26	256	13
44. Clergé, religieux	59	3	21	1
45. Professions intermédiaires administratives de la fonction publique	278	12	172	9
46. Professions intermédiaires administratives et commerciales des entreprises	995	42	646	33
47. Techniciens	678	29	564	29
48. Contremaîtres, agents de maîtrise	571	24	528	27
5. EMPLOYÉS	6 247	265	2 219	113
52. Employés civils et agents de service de la fonction publique	1 703	72	642	33
53. Policiers et militaires	380	16	305	16
54. Employés administratifs d'entreprise	2 532	108	819	42
55. Employés de commerce	742	32	208	11
56. Personnels des services directs aux particuliers	890	38	245	13
6. OUVRIERS	7 749	329	4 921	251
62. Ouvriers qualifiés de type industriel	1 602	68	1 232	63
63. Ouvriers qualifiés de type artisanal	1 509	64	1 059	54
64. Chauffeurs	567	24	491	25
65. Ouvriers qualifiés de la manutention, du magasinage et du transport	417	18	327	17
67. Ouvriers non qualifiés de type industriel	2 353	100	1 172	60
68. Ouvriers non qualifiés de type artisanal	1 007	43	473	24
69. Ouvriers agricoles	294	12	168	9
81. CHÔMEURS N'AYANT JAMAIS TRAVAILLÉ	353	15	26	1
POPULATION ACTIVE	23 525	1 000	13 273	678
POPULATION INACTIVE	30 748		6 317	322
71. Anciens agriculteurs	1 346		726	37
72. Anciens artisans, commerçants, chefs d'entreprises	755		491	25
74. Anciens cadres	344		282	14
75. Anciennes professions intermédiaires	722		502	26
77. Anciens employés	1 976		1 273	64
78. Anciens ouvriers	2 294		1 651	84
83. Militaires du contingent	251		13	1
84. Élèves, étudiants de plus de 15 ans	4 278		195	10
85. Autres inactifs entre 15 et 60 ans	5 572		584	30
86. Autres inactifs de plus de 60 ans	1 978		600	31
Enfants de moins de 15 ans	11 233		0	0
POPULATION TOTALE	*54 273*		*19 590*	*1 000*

Figure 3.8: Les catégories socio-professionnelles après 1982 (source : INSEE)

Dans le cadre de la recherche de nouvelles variables discriminantes pour l'incidence en arrêt de travail, la variable CSP semble être une candidate idéale. En effet, c'est une variable directement lié aux caractéristiques de l'individu et, avant-même de procéder à des mesures de sinistralité, il serait cohérent que la propension d'un individu à subir des arrêts de travail dépende de sa profession.

Néanmoins, un des problèmes de notre base de données est que la variable Catégorie Socio-Professionnelle est manquante à 50%. Ce qui signifie que l'on connait la profession de seulement la moitié de notre portefeuille d'assurés. Ce taux de valeurs manquantes élevé peut avoir plusieurs explications :

— Les individus qui souscrivent un contrat ont une profession qui pourrait être difficile à lier avec une CSP unique. Ainsi, lors du renseignement informatique des données de l'assuré, la variable CSP serait laissée vide.

— Certains groupes de tarifications existent déjà et ont été déterminés sur la base des métiers précis des individus. Ainsi, lors du renseignement informatique des données de l'assuré, il n'est pas nécessaire de renseigner la variable CSP.
— Enfin, on pourrait l'expliquer par de simples erreurs de gestion ou un manque de rigueur lors des renseignements des bases.

Une première solution pour palier à ce taux de données manquantes s'offre à nous : il s'agit de récupérer les données bancaires des assurés qui ont non seulement des contrats d'assurances mais aussi des comptes bancaires parmi l'une des différentes Banques Populaires du groupe. A partir de là, nous pourrions effectuer une jointure des données et récupérer les Catégories Socio-Professionnelles supplémentaires renseignées dans les données bancaires.

C'est donc de cette manière que nous avons d'abord procédé. Toutefois, les CSP renseignées côté banque suivent la nomenclature postérieure à 1982. Il a donc fallu d'abord définir des liaisons entre la nomenclature antérieure à 1982 et celle-ci afin de s'assurer de la significativité de la variable Catégorie Socio-Professionnelle finale.

Au final, nous parvenons à renseigner 15% de données supplémentaires et obtenons donc une variable CSP avec 65% de données connues. Voici la répartition des clients de notre base de travail entre les différentes CSP de la nomenclature antérieure à 1982 :

lib_csp <chr>	Exposition <dbl>
NA	32887.80
Artisans	20674.51
Industriels	19753.86
Commerçants	8759.12
Cadres et professions intellectuelles supérieures	6350.05
Ouvriers	3720.61
Agriculteurs et salariés agricoles	3608.14
Cadres moyens	3571.74
Employés	3404.49
Personnel de service	534.36
Autres catégories	57.22

Figure 3.9: La répartition des clients par CSP dans notre base de travail

Nous constatons que les Artisans et les Industriels sont les catégories les plus représentées. Autrement, la distribution est assez uniforme mis à part pour les Personnels de service et les Autres catégories. Plus loin dans ce mémoire, nous vous présenterons différentes méthodes que nous avons essayé d'appliquer afin de parvenir à compléter les données manquantes de cette variable CSP.

Première partie

Identification de variables discriminantes

4 Analyse descriptive

Dans cette partie, nous procéderons à une analyse descriptive simple de notre base de travail avec pour objectif d'effectuer *a priori* des premiers constats sur les variables qui ont une influence sur l'incidence en arrêt de travail. Ces constats serviront de base de départ pour les futures études approfondies présentées plus loin dans ce mémoire.

4.1 Statistiques descriptives

Pour isoler des variables discriminantes, nous allons dans cette partie passer en revue un à un les taux d'incidence exprimés en fonction des variables de la base.

Influence de la région d'habitation

Nous commençons par tracer le taux d'incidence en arrêt de travail en fonction des différentes régions françaises.

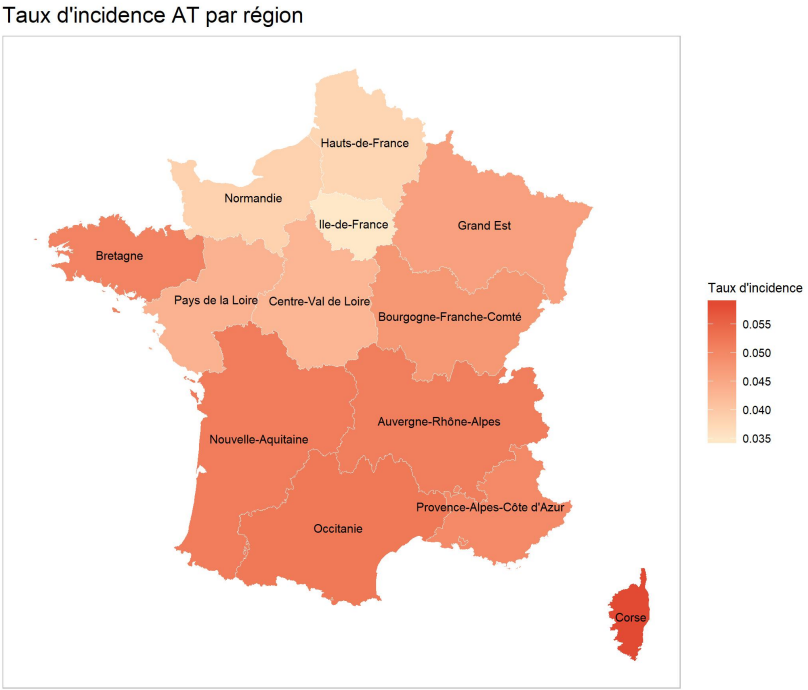

Figure 4.1: Taux d'incidence AT par région

Il apparaît que les régions du nord de la France comportent un niveau de risque moins élevé que leurs homologues du sud [5]. La Corse quant à elle présente le taux d'incidence le plus élevé avec un taux de 5,8%.

5. Hormis pour la Bretagne.

Influence de l'entité bancaire

Nous poursuivons notre étude avec l'observation du taux d'incidence en arrêt de travail par grande entité bancaire.

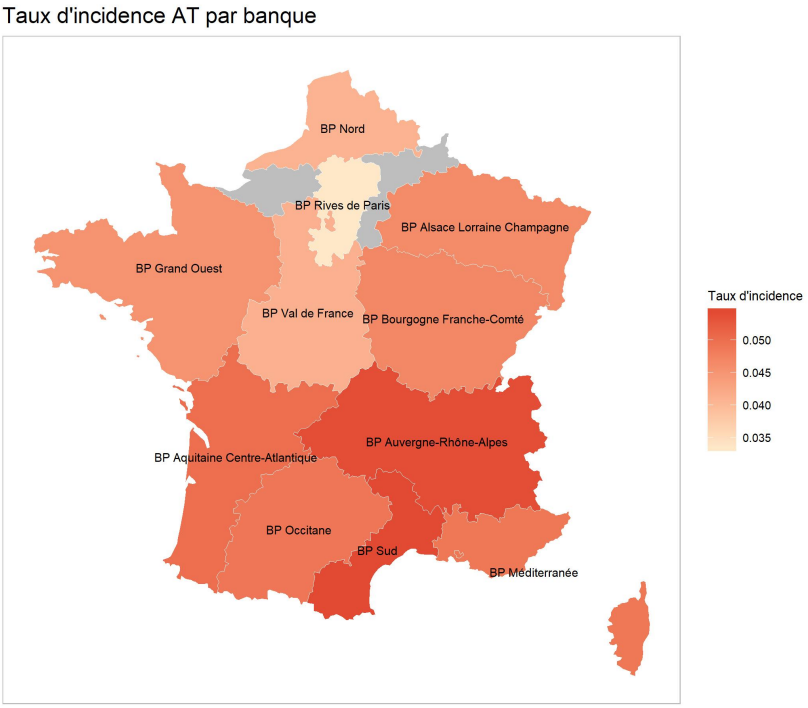

Figure 4.2: Taux d'incidence AT par entité bancaire

Ici, le résultat précédent observé sur les régions est en partie confirmé et complété.

Nous retrouvons bien le fait que les entités bancaires du nord sont associées à un niveau de risque plus faible (en particulier pour les BP Val de France, Rives de Paris et Nord). Contrairement au graphique des régions, les établissements BP Auvergne-Rhône-Alpes et Sud sortent nettement du lot en présentant des taux d'incidence très élevés.

Conclusion sur les variables géographiques

Il est intéressant de constater que les variables géographiques mettent bien en avant des disparités parmi les individus du portefeuille. Nous retenons ces résultats et viendrons les discuter plus tard dans ce mémoire.

Pour la suite de l'analyse descriptive, nous retrouverons dans tous les graphiques suivants deux courbes distinctes :

— Une courbe bleue, qui représente le taux d'incidence.
— Une courbe orange, qui représente l'exposition (le nombre d'individus) au sein du portefeuille.

Influence de l'âge

Poursuivons donc notre étude avec l'observation du taux d'incidence en incapacité par âge.

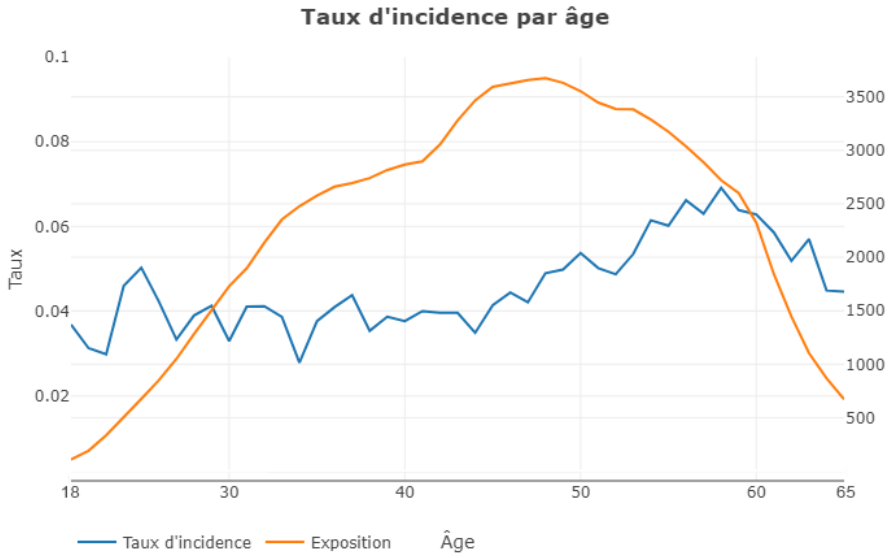

Figure 4.3: Taux d'incidence en incapacité par âge

Nous constatons que l'incidence augmente avec l'âge de l'assuré. Pour être plus précis, il semblerait que celle-ci soit relativement constante (en tout cas il est difficile d'identifier une tendance) pour les âges inférieurs à 45 ans, et qu'elle augmente fortement à partir de cette âge-là pour finalement diminuer légèrement à l'approche de la retraite.

Ces résultats sont plutôt logiques puisque la santé des individus tend à diminuer avec l'âge. En outre, nous pourrions interpréter le pic d'incidence pour les individus âgés de 20 à 25 ans par le fait que ceux-là ont des comportements plus risqués d'un point de vue risque accidentel.

Influence de l'ancienneté

Nous étudions maintenant l'influence de l'ancienneté sur l'incidence en arrêt de travail. Par ancienneté nous entendons le nombre d'années pendant lesquelles les individus sont restés dans notre portefeuille d'assurés.

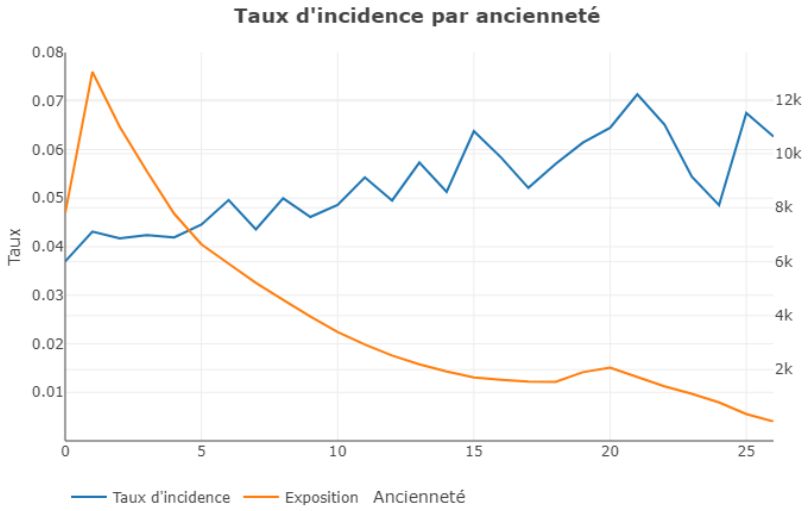

Figure 4.4: Taux d'incidence AT par ancienneté

À la lecture du graphique, il semblerait que le taux d'incidence en arrêt de travail tend à augmenter avec l'ancienneté des individus. Néanmoins, ce résultat est à tempérer pour plusieurs raisons :

— D'une part, Les pics d'incidence dessinés pour les anciennetés supérieures à 10 ans témoignent d'une exposition trop peu élevée pour pouvoir interpréter correctement les résultats.
— D'autre part, l'ancienneté est probablement fortement corrélée avec l'âge. En effet, un individu qui a 15 ans d'ancienneté en portefeuille est probablement un individu âgé.

Ainsi, l'impact de l'ancienneté n'est pas encore confirmé et la variable sera réétudiée dans une des prochaines étapes du mémoire.

Influence de la catégorie socio-professionnelle

Nous poursuivons notre étude avec l'observation du taux d'incidence par CSP.

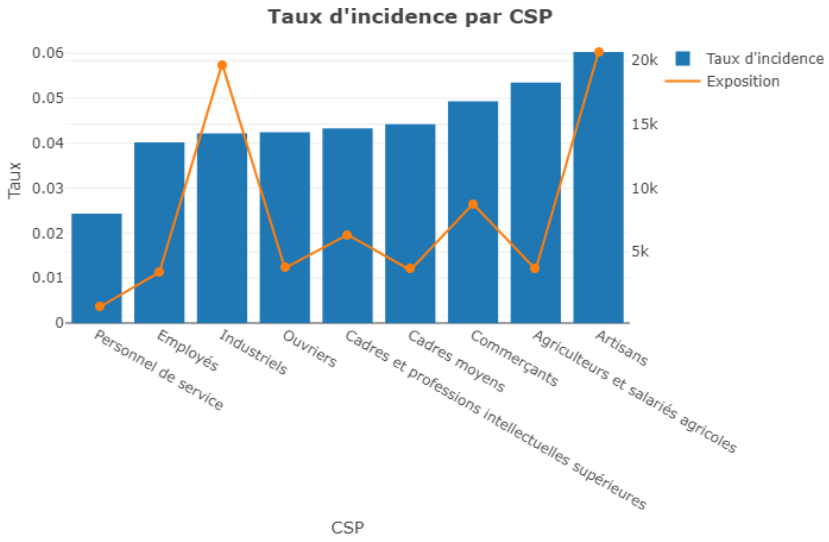

Figure 4.5: Taux d'incidence en incapacité par Catégorie Socio-Professionnelle

Nous constatons que la catégorie socio-professionnelle des individus semble avoir un impact sur l'incidence en arrêt de travail. En effet, nous retrouvons d'une part des classes peu soumises au risque comme les personnels de service ou les employés ; et d'autre part des classes plus exposées comme les agriculteurs ou les artisans.

C'est donc une variable que nous souhaitons a priori conserver pour notre étude et il sera intéressant de s'intéresser plus en détail à son impact dans une des prochaines parties du mémoire.

Influence de la sélection médicale

Nous poursuivons notre étude avec l'étude du taux d'incidence en arrêt de travail en fonction de la sélection médicale des individus.

Sélection médicale	Taux de sinistralité	Exposition
Oui	4.98%	19 931
Non	4.64%	84 195

Table 4.1: Taux d'incidence en fonction de la sélection médicale

Il semblerait que le portefeuille soit assez ancien pour que l'effet de la sélection médicale se soit estompé. En effet, la variable ne semble plus avoir d'impact sur l'incidence puisque les taux présentés sont très proches. Nous réétudierons ses effets plus en détail dans une prochaine partie du mémoire.

Influence du montant garanti

En ce qui concerne l'étude de l'impact du montant garanti sur l'incidence en arrêt de travail, notons tout d'abord que les individus ont été regroupés par tranche de montants garantis d'exposition environ.

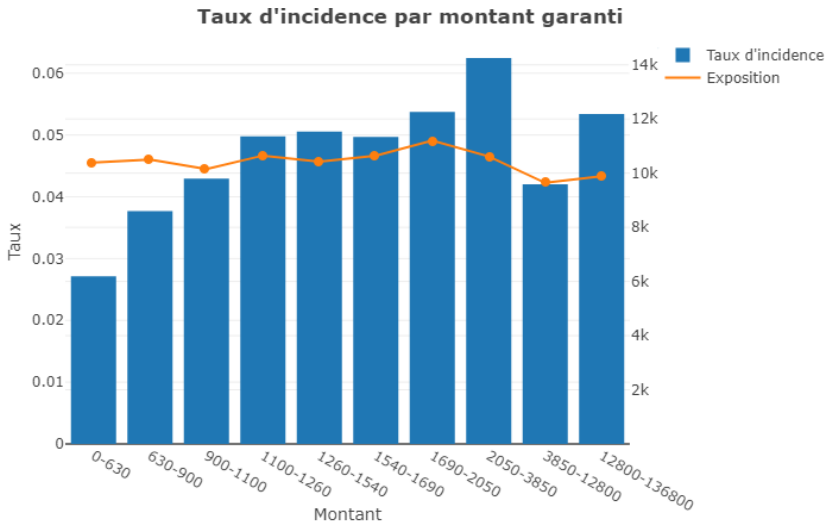

Figure 4.6: Taux d'incidence AT par tranche de montant garanti

Il semblerait que le montant garanti ait un fort impact sur l'incidence puisque, pour de faibles montants garantis, nous observons des taux d'incidence autour de 3% alors que, pour des montants garantis plus importants, nous observons des taux pouvant monter jusqu'à 6%.

Influence de la franchise

Nous retrouvons trois types de franchise distincts dans notre portefeuille d'assurés :

— « Accident seul à 7 jours » qui offre aux individus une protection contre l'accident seulement avec une durée de franchise de 7 jours.
— « Accident à 7 jours et Maladie à 15 jours » qui offre aux individus une protection contre l'accident avec une durée de franchise de 7 jours et la maladie avec une durée de franchise de 15 jours.
— « Accident à 30 jours et Maladie à 30 jours » qui offre aux individus une protection contre l'accident avec une durée de franchise de 30 jours et la maladie avec une durée de franchise de 30 jours.

Franchise	Taux de sinistralité	Exposition
Accident seul à 7 jours	3.28%	6 464
Accident à 7 jours et Maladie à 15 jours	5.05%	53 088
Accident à 30 jours et Maladie à 30 jours	4.57%	43 892

Table 4.2: Taux d'incidence par franchise

Les chiffres observés sont cohérents, puisque les deux règles suivantes sont respectées :

— Moins il y a de risques couverts, moins l'incidence est élevée.
— Plus la durée de franchise est courte, plus l'incidence est élevée.

Influence du montant de prime

Nous terminons l'étude du taux d'incidence en fonction du montant de prime. Bien que nous n'utiliserons pas la variable *montant de prime* comme variable explicative de nos modèles, il est toujours utile pour nous de vérifier la cohérence de la tarification sur le portefeuille existant.

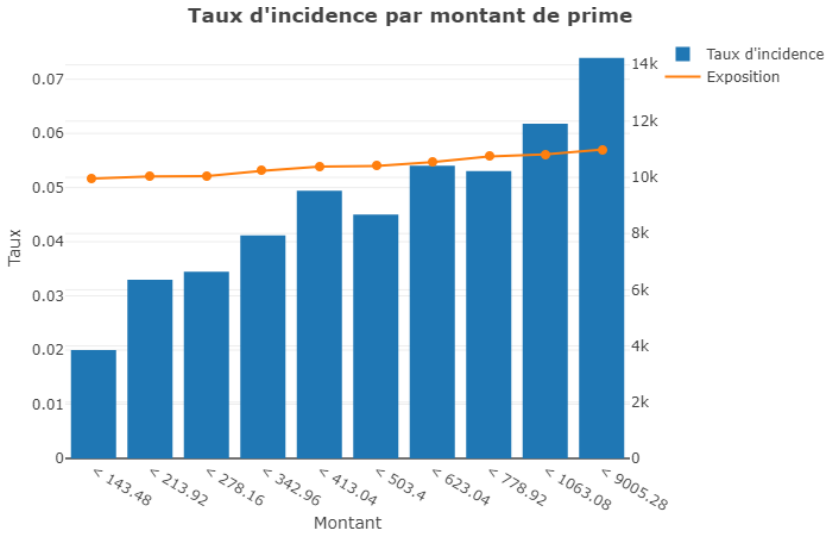

Figure 4.7: Taux d'incidence AT par montant de prime

Encore ici, des tranches de montants de prime ont été créées de telle sorte qu'elles regroupent à chaque fois environ 10% des individus du portefeuille.

Nous constatons que le taux d'incidence semble bien augmenter avec le montant de prime payé par les individus, ce qui prouve la significativité de la prime actuelle puisqu'elle met bien en évidence des groupes d'individus avec des profils de risque distincts.

4.2 Conclusion

L'objectif de l'analyse descriptive présentée dans cette partie était de se construire une première opinion sur les différentes variables qui pouvaient avoir un impact sur l'incidence en incapacité.

Par l'étude méthodique des variables qui constituent notre base, nous sommes parvenus à mettre en évidence des caractéristiques risquées et moins risquées. À ce stade du mémoire, voici ce à quoi ressemble le profil d'un assuré peu risqué :

Région	Ile-de-France, Hauts-de-France, Normandie
Établissement	Rives de Paris, Nord, Val de France
Âge	$<$ 45 ans
Ancienneté	$<$ 10 ans
Profession	Personnel de service, Employé
Montant garanti	Faible

Table 4.3: Caractéristiques d'un individu peu risqué

A l'inverse, voici ce à quoi ressemble le profil d'un assuré risqué :

Région	Auvergne-Rhône-Alpes, Nouvelle-Aquitaine, Occitanie, Corse
Établissement	Auvergne-Rhône-Alpes, Sud
Âge	$>$ 45 ans
Ancienneté	$>$ 10 ans
Profession	Agriculteur, Artisan
Montant garanti	Élevé

Table 4.4: Caractéristiques d'un individu risqué

Ces résultats sont très intéressants mais restent à nuancer puisqu'à ce stade, nous n'avons pas encore étudié les structures de corrélations qui lient les différentes variables [6].

Dans la prochaine partie, nous mettons en place un nouveau type d'approche avec l'objectif de discuter les résultats obtenus par analyse descriptive.

6. Nous conservons cette phase de l'analyse pour une partie ultérieure du mémoire.

5 Méthodes d'apprentissage non-supervisé

Précédemment, nous avons mis en évidence par une analyse descriptive élémentaire des caractéristiques spécifiques qui pouvaient faire qu'un individu serait plus ou moins exposé au risque. Dans cette partie du mémoire, nous aurons l'approche suivante : utiliser des algorithmes de la famille des méthodes d'apprentissage non-supervisé pour obtenir des classifications de nos données. Les objectifs sont les suivants :

— D'une part, nous pouvons effectuer des travaux de différents ordres sur des variables explicatives de la base (complétude des valeurs manquantes, regroupement des modalités).
— D'autre part, nous pouvons étudier des groupes d'individus qui présentent des caractéristiques similaires vis à vis de la sinistralité en arrêt de travail.

Nous commencerons ainsi par présenter les algorithmes utilisés en introduisant sommairement la théorie informatique et mathématique.

5.1 Méthodes d'apprentissage non-supervisé

Par méthode d'apprentissage non-supervisé (2017) [9] est entendu tout type de méthode qui consiste à apprendre sans superviseur. C'est à dire qu'il s'agit, non-pas de prédire de la réalisation d'une variable ou de calculer un score de réussite, mais plutôt d'extraire des classes ou des groupes d'individus présentant des caractéristiques communes.

Dans le cadre de notre étude, l'objectif sera d'isoler des groupes d'individus qui ont des profils de risques similaires. Nous chercherons donc à créer une partition de notre sinistralité et nous étudierons les modalités prépondérantes des variables explicatives pour chaque classe d'individus créées.

Nous aurons donc recours à plusieurs méthodes d'apprentissage non-supervisé connues, que nous présenterons dans les premières sous-parties suivantes.

5.1.1 Classification Ascendante Hiérarchique

La Classification Ascendante Hiérarchique (2005) [16], abrégée CAH, est une procédure non-supervisée qui consiste à émettre une classification hiérarchique des individus. C'est à dire que la construction des groupes d'individus aux caractéristiques similaires produit une structure arborescente et nous pouvons définir une profondeur de classification.

Pour deux individus distincts, il existe toujours au moins une profondeur de classification qui fait que ces individus appartiennent à la même classe. Il reste désormais à définir le critère de ressemblance entre les individus. C'est à dire sur quoi est-ce que nous nous basons pour dire que deux individus sont proches.

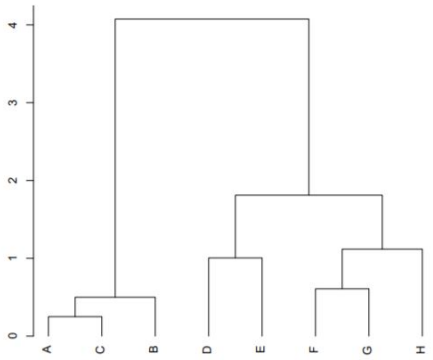

Figure 5.1: Exemple de classification produite par l'algorithme CAH

Dans la cadre de notre étude, nous mesurerons la proximité des individus en utilisant la distance euclidienne. Supposons un nombre de variables explicatives n et deux individus distincts x et y, la distance euclidienne entre les deux individus est définie de la manière suivante :

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Il est important d'introduire ensuite les notions d'inertie. Considérons N individus regroupés en K groupes avec I variables explicatives. La réalisation de de la variable i pour l'individu n appartenant au groupe k est ainsi notée : x_{ikn}. On a la relation suivante :

$$\sum_{n=1}^{N}\sum_{i=1}^{I}(x_{in} - \bar{x}_i)^2 = \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{I}(x_{ikn} - \bar{x}_{ik}^2) + \sum_{k=1}^{K}\sum_{i=1}^{I}(\bar{x}_{ik} - \bar{x}_i)^2$$

Définissons ce que signifie chacun des termes :

— Le premier terme correspond à l'inertie totale. C'est à dire que nous mesurons la variabilité globale de toutes les variables.
— Le deuxième terme correspond à l'inertie intra-classe. C'est à dire que nous mesurons la variabilité au seins des classes de toutes les variables, et ce pour chaque classe.
— Le dernier terme correspond à l'inter-classe. C'est à dire que nous mesurons la variabilité globale entre les classes qui constituent les nouvelles observations.

Nous pouvons ainsi mesurer la qualité d'une classification en utilisant l'inertie. En effet, si le rapport entre l'inertie inter-classe et l'inertie totale est proche de 1, cela signifie que la classification parvient bien à capter les variations globales des données.

Nous préciserons maintenant une méthode de production de l'aborescence décrite précédemment : la méthode de Ward (2006) [14].

Celle-ci consiste à partir d'une situation initiale ou chaque individu constitue une classe en tant que telle, puis à agréger les individus entre eux de sorte à minimiser à chaque étape la diminution de l'inertie inter-classe.

— **Initialisation :** 1 classe = 1 individu (Inertie inter. = Inertie totale)
— **A chaque étape :** agréger les classes qui minimisent la diminution de l'inertie inter.

Cette méthode présente entre autres l'avantage d'être simple à mettre en place. Dans la sous-partie suivante, nous parlons d'une autre méthode de classification que nous utiliserons : la méthode des K-Means.

5.1.2 Méthode des K-Means

Définition

K-means (2018) [4] est un algorithme non-supervisé de classification non-hiérarchique. Il permet de regrouper en K clusters distincts les observations de notre base de données. Les données similaires se retrouveront dans un même cluster. Une même observation ne pourra appartenir à deux clusters différents.

Similarité des données

Pour pouvoir regrouper un jeu de données en K clusters distincts, l'algorithme K-Means a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. Ainsi, deux données qui se ressemblent, auront une distance de similarité faible, alors que deux objets différents auront une distance de séparation élevée. Dans notre cas, nous utiliserons la même mesure de distance que celle introduite dans la partie de Classification Ascendante Hiérarchique : la distance euclidienne.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Choix du nombre de clusters

Il n'y a pas de règles précises pour définir le nombre de clusters K à choisir. Un nombre K trop élevé peut conduire à un partitionnement trop fragmenté des données, ce qui empêchera de découvrir des schémas intéressants. Par contre, un nombre de clusters K trop faible conduira à avoir, potentiellement, des clusters trop généralistes contenant beaucoup de données. Dans ce cas, nous n'aurons pas non-plus de conclusion instructive.

Pour une même base de données, il n'existe pas d'unique classification possible. La difficulté résidera donc dans le fait de choisir un nombre de clusters K qui permettra de mettre en lumière des schémas intéressants de relation entre les données. Malheureusement, il n'existe pas de procédé automatisé pour trouver le bon nombre de clusters.

La méthode la plus répandue pour choisir le nombre de clusters est de lancer l'algorithme avec différentes valeurs de K et d'étudier les rapports d'inertie [7].

Principe algorithmique

- **Entrée :** K le nombre de cluster à former ; les données à grouper.
- **Initialisation :** Choisir aléatoirement K points. Ces points sont les centres des clusters.
- **Répéter :** Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche au son centre ; Recalculer le centre de chaque cluster et en modifier le centre.
- **Jusqu'à :** Convergence ou stabilisation de l'inertie totale de la population.

Optimums locaux

En analysant la façon de procéder de l'algorithme de K-means, nous remarquons que pour une même base de données, nous pouvons avoir des partitionnements différents. En effet, L'initialisation des tous premiers centres de clusters est complètement aléatoire. Par conséquent, l'algorithme trouvera des clusters différents en fonction de cette première initialisation aléatoire. De ce fait, la configuration des clusters trouvés par K-Means peut ne pas être optimale. Nous utilisons le terme d'optimum local.

Afin de pallier aux problèmes des optimums locaux, il suffit de lancer K-means plusieurs fois sur le jeu de données (avec le même nombre K de clusters et des initialisations initiales différentes) et d'étudier la composition des clusters qui se forment.

Maintenant que nous connaissons les principes de fonctionnement clés de l'algorithme K-Means, nous présentons un autre algorithme de la famille des méthodes d'apprentissage non-supervisé : l'Analyse en Composantes Principales.

5.1.3 Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) (2013) [13] est une méthode d'apprentissage non-supervisé qui permet de transformer des variables corrélées entres elles en un nouvel ensemble de variables décorrélées. Ces nouvelles variables sont appelées « Composantes Principales ».

Dans le cadre de notre étude, cette méthode nous permettra de réduire le nombre de dimensions de notre base de travail. Dans ce nouvel espace constitué des axes principaux, nous pourrons ensuite mesurer de nouvelles distances entre les individus.

7. cf. page 29 pour la définition des différents types d'inertie dans un jeu de données.

Pour déterminer ces axes principaux, l'algorithme essaie de mettre en lumière les axes qui expliquent le mieux la dispersion générale des données. Il s'agit à la fois d'une approche géométrique (puisque les variables sont représentées dans un nouvel espace de travail) et d'une approche statistique (puisque la recherche portant sur les nouveaux axes permet de mieux expliquer la variance globale des données).

Maintenant que nous avons brièvement introduit l'Analyse en Composantes Principales, nous présentons une autre méthode d'apprentissage non-supervisé : l'Analyse Factorielle des Données Mixtes.

5.1.4 Analyse Factorielle des Données Mixtes

La base de données sur laquelle porte notre étude contient à la fois des variables qualitatives et des variables quantitatives : les données sont dites mixtes. Pour les analyser, nous pouvons avoir recours à un Analyse Factorielle des Données Mixtes (AFDM) (2017) [15].

Cet algorithme permet d'étudier les différents schémas de corrélation qui régissent l'existence de nos variables et de mettre en évidence des classifications d'individus aux profils similaires. Face à des données mixtes, nous pouvons procéder :

— soit en transformant la nature de l'un des types de variables pour réaliser une ACP (qui ne s'applique qu'à des données quantitatives) ou une Analyse des Correspondances Multiples (qui s'applique uniquement aux variables qualitatives).
— soit en conservant la nature de chaque variable afin d'opter pour l'AFDM.

L'ACP et l'ACM sont des algorithmes intéressants pour étudier des données homogènes mais présentent des faiblesses si leurs périmètres respectifs sont étendus : la première requiert une réduction des données ; la seconde impose des contraintes sur le nombre d'observations et sur le découpage des variables quantitatives en classes.

L'AFDM permet de combiner les avantages de ces deux méthodes. Le principe est d'utiliser la normalisation de l'ACM sur les variables qualitatives afin d'équilibrer l'influence de chaque ensemble de variables et d'effectuer une ACP sur la matrice des données devenues quantitatives.

Maintenant que nous avons introduit toutes les différentes méthodes d'apprentissage non-supervisé que utilisées dans ce mémoire, nous passons à l'étape d'application avec dans un premier temps, la tentative de réduction des modalités des variables qualitatives de notre base de données.

5.2 Réduction des modalités des variables qualitatives

Dans notre base de données, nous disposons de plusieurs variables qualitatives que nous pouvons utiliser pour mettre en évidence des caractéristiques qui rendraient un individu plus ou moins exposés au risque.

Ultérieurement, nous entamerons des étapes de modélisation d'une loi d'incidence. À ce moment-là, nous voulons bénéficier d'une exposition suffisante dans notre base de travail pour que l'étude reste cohérente.

Dans cette sous-partie, nous tentons d'appliquer les algorithmes de classification présentés précédemment pour réduire le nombre de modalités de nos variables qualitatives et ainsi garantir un niveau d'exposition minimal pour la modélisation.

Pour effectuer cette réduction, nous étudierons les modalités proches au sens des profils de risque. Pour chaque variable, nous travaillerons donc avec une base des taux de sinistralité par année entre 2014 et 2019.

5.2.1 Région

Nous commençons par créer la base de travail. C'est une base simple ; l'idée est d'effectuer des regroupements élémentaires que nous challengerons par la suite.

nom_region <chr>	tx_sin_14 <dbl>	tx_sin_15 <dbl>	tx_sin_16 <dbl>	tx_sin_17 <dbl>	tx_sin_18 <dbl>	tx_sin_19 <dbl>
Auvergne-Rhône-Alpes	0.05747126	0.04388817	0.05314497	0.04543902	0.04436039	0.03697121
Bourgogne-Franche-Comté	0.04321839	0.04654511	0.04451039	0.04854369	0.03559719	0.04236499
Bretagne	0.05089059	0.05056890	0.04889976	0.03860072	0.04970414	0.03406326
Centre-Val de Loire	0.03495630	0.05290323	0.04727273	0.03758542	0.02961276	0.03076923
Grand Est	0.04889179	0.04126547	0.04294918	0.04104478	0.04942085	0.03015873
Hauts-de-France	0.04183536	0.03337454	0.03206651	0.03243848	0.03402854	0.02777778
Ile-de-France	0.02723268	0.03775204	0.03448276	0.02951137	0.03377309	0.03290993
Normandie	0.06561680	0.01813472	0.03141361	0.04221636	0.03166227	0.02185792
Nouvelle-Aquitaine	0.04964200	0.05633075	0.05029428	0.04190981	0.04552509	0.04014228
Occitanie	0.04674010	0.04715803	0.05028463	0.05297368	0.04461590	0.04474506
Pays de la Loire	0.04895833	0.04569055	0.03917526	0.03510532	0.03544061	0.02923977
Provence-Alpes-Côte d'Azur	0.04819277	0.05542545	0.04276069	0.03790751	0.04318182	0.04148148

12 rows

Figure 5.2: La base de travail pour la réduction des modalités de région

À ce stade, nous avons pris la décision de fusionner les modalités de région « Corse » et « Provence-Alpes-Côte d'Azur ». En effet, comme l'exposition par année entre 2014 et 2019 pour la région Corse était trop faible pour être considérée à part, nous avons décidé de la regrouper avec une région qui présentait un profil de risque similaire.

Après avoir construit notre base de travail, nous effectuons un test de corrélation entre les taux d'incidence par année. En effet, si certains sont trop corrélés, alors les résultats du modèle de classification que nous appliquerons pourraient être biaisés.

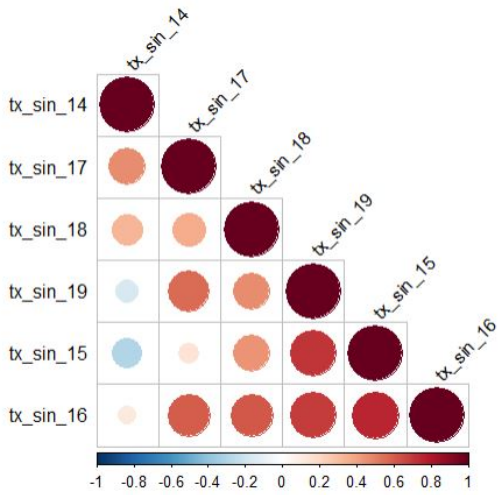

Figure 5.3: Corrélation entre les variables de sinistralité par région

Nous observons parfois des taux de corrélations élevés mais aucun ne dépasse le seuil arbitraire de 80%. Ainsi, nous décidons de procéder à la modélisation sans retirer aucune variable de la base.

Analyse en Composante Principale

La première étape consiste à appliquer une ACP. Le tableau récapitulatif des valeurs propres obtenus est le suivant :

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.17398612	63.479722	63.47972
comp 2	0.86530024	17.306005	80.78573
comp 3	0.61370550	12.274110	93.05984
comp 4	0.29119467	5.823893	98.88373
comp 5	0.05581347	1.116269	100.00000

Figure 5.4: Tableau des valeurs propres de l'ACP

Nous pouvons aussi tracer la contribution de chaque dimension à la variabilité totale :

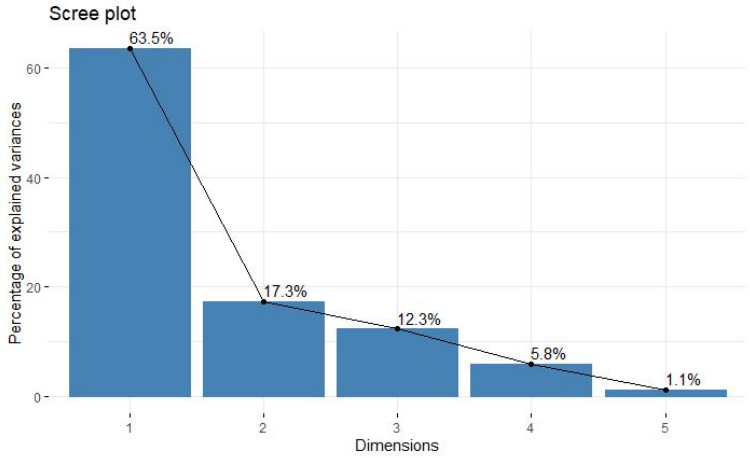

Figure 5.5: Contribution de chaque dimension à expliquer la variabilité totale

Pour choisir le nombre de dimensions que nous décidons de conserver, nous avons le choix entre plusieurs méthodes. Ici, nous décidons de ne pas utiliser la méthode du coude. En effet, nous constatons que les trois premières dimensions permettent d'expliquer 93% de la variabilité totale.

Étant capables d'interpréter correctement les résultats dans un espace à 3 dimensions, et le score de variabilité étant très élevé, nous décidons de retenir les trois premières dimensions identifiées par l'ACP.

Nous pouvons représenter ensuite la projection des modalités (qui constituent les individus du modèles) de la variable Région sur le nouvel espace à deux dimensions engendré par les deux premières composantes de notre Analyse en Composante Principale.

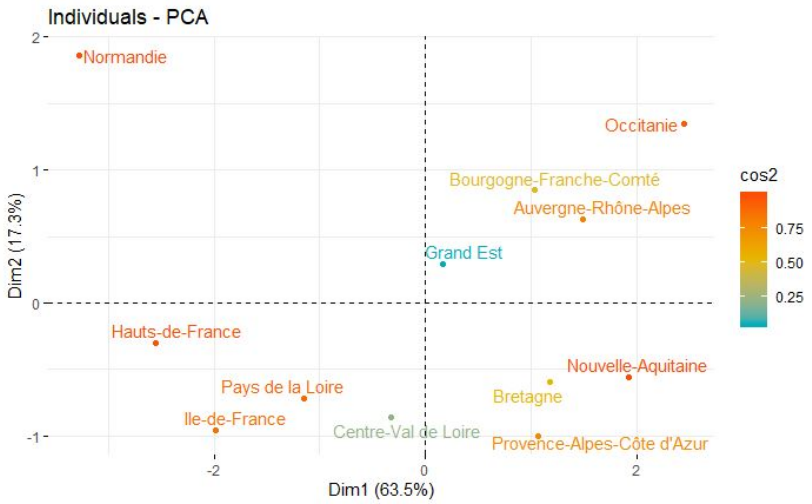

Figure 5.6: Position des individus sur l'espace des dimensions de l'ACP

Certains individus sont mal représentés par ces deux premières composantes : c'est le cas des régions Grand Est et Centre-Val de Loire (ce qui valide notre choix de conserver non-pas seulement les deux premières composantes de l'ACP mais en fait les trois premières). Ainsi, nous ne pouvons pas correctement interpréter leur position sur le graphique.

En ce qui concerne les autres régions, nous observons leur proximité en terme de distance euclidienne pour isoler des premiers groupes d'individus aux profils similaires. Nous distinguons quatre groupes disposés dans chaque coin du graphique. Les régions qui constituent ces groupes sont proches, et nous allons confirmer cela en appliquant la méthode de Ward pour les classifier.

Après application de la méthode de Ward, nous obtenons les clusters d'individus suivants :

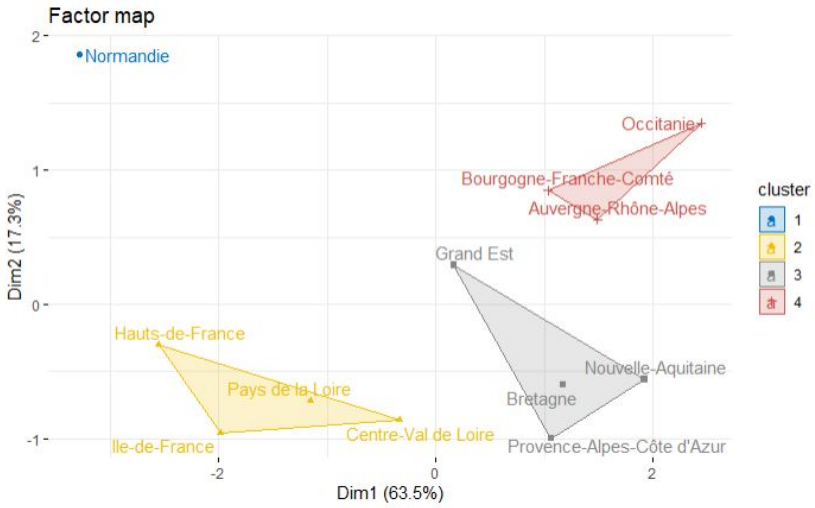

Figure 5.7: Représentation de la classification sous forme de clusters

Les groupes créés semblent cohérents avec l'analyse descriptive que nous avions effectuée sur la distribution de la sinistralité en fonction des régions françaises. En effet, nous retrouvons bien :

- Un groupe de régions à risque faible : Centre-Val de Loire, Hauts-de-France, Ile-de-France et Pays de la Loire.
- Un groupe de régions à risque élevé : Bourgogne-Franche-Comté, Auvergne-Rhône-Alpes et Occitanie.
- Un groupe de régions neutres constitué des régions restantes.

De plus, la Normandie doit être mise à part puisqu'elle est la région du portefeuille où nous avons l'exposition la plus faible ; les variations de la sinistralité en son sein sont donc plus difficilement interprétables.

Pour terminer notre étude sur la variable région, nous calculons à partir des nouvelles classes définies l'inertie inter-classe et l'inertie totale [8]. Il apparaît l'égalité suivante :

$$\frac{\text{inertie inter-classe}}{\text{inertie totale}} = 73\% \tag{5.1}$$

Cela signifie que les classes créées parviennent à bien représenter la variabilité des données.

Pour respecter le cadre défini lors de la présentation de la méthode de Ward dans une partie ultérieure du mémoire, nous représentons en début de page suivante ces groupes sous la forme d'une arborescence.

8. cf. page 29 pour l'introduction aux différents types d'inertie.

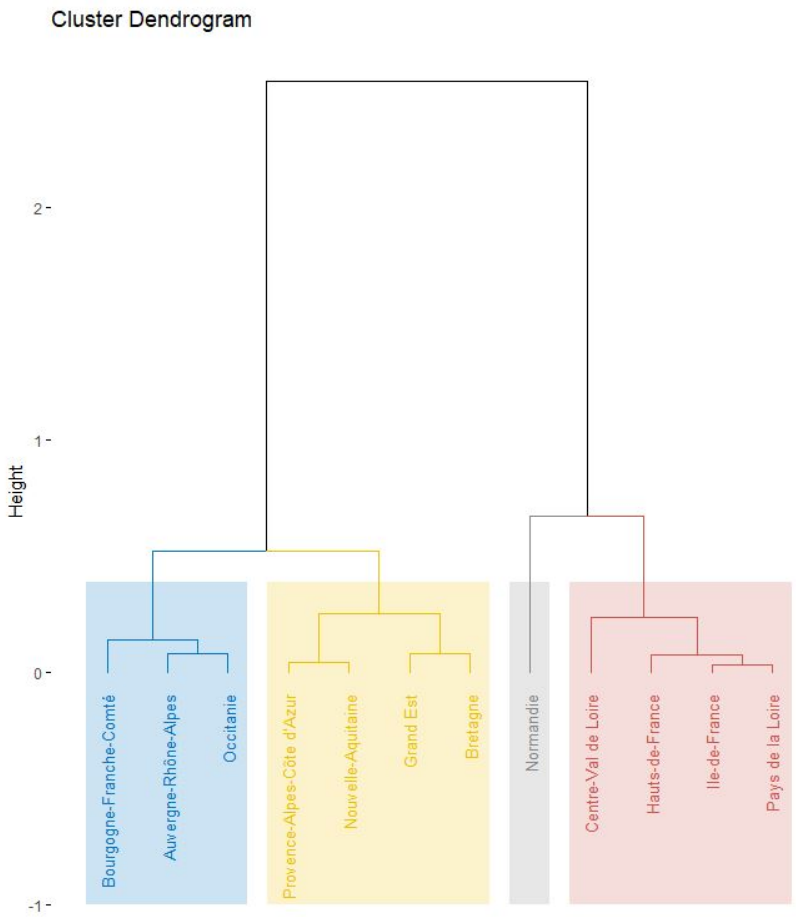

Figure 5.8: Représentation de la classification sous forme d'une arborescence

5.2.2 Banque

Pour la variable Banque, le même procédé que celui-présenté précédemment a été réappliqué. C'est à dire que :

1. Nous avons créé une base de travail ayant pour variables explicatives les taux de sinistralité annuels entre 2014 et 2019.
2. Nous avons étudié les taux de corrélations entre les variables de la base et retiré les variables pour lesquelles la corrélation était supérieure au seuil arbitraire de 80%.
3. Nous avons appliqué un algorithme d'ACP pour d'une part, réduire les dimensions de notre base de travail ; d'autre part, mettre en évidence les liens entre les modalités par projection sur un espace restreint.
4. Nous avons appliqué la méthode de Ward pour obtenir une classification hiérarchique des modalités de la variable banque.

Comme précédemment, nous présentons le graphique de répartition des modalités de la variable sur les deux premières dimensions de l'espace mis en évidence par l'ACP.

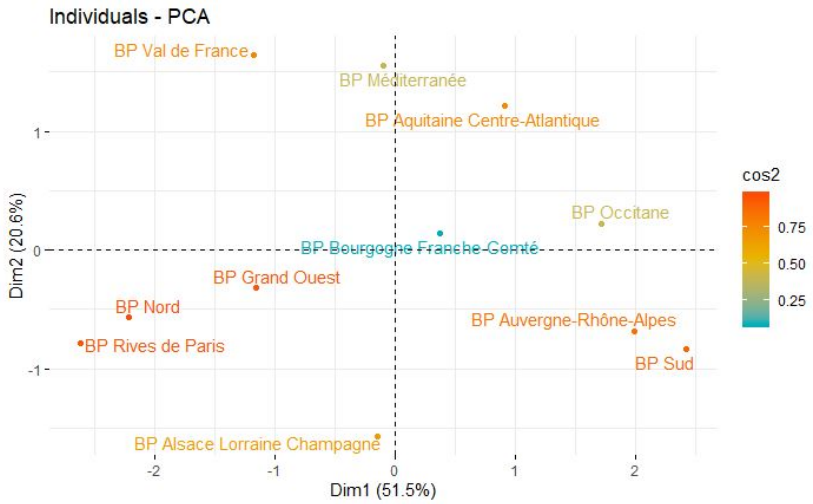

Figure 5.9: Position des modalités de la variable banque sur l'espace de dimension réduit défini par l'ACP

Nous constatons l'apparition de certaines proximités déjà observées aux étapes antérieures du mémoire :

— Les BP Auvergne-Rhône-Alpes et Sud apparaissaient toutes les deux liées à un risque élevé auparavant et apparaissent proche dans le graphique des individus.
— Les BP Nord et Rives de Paris apparaissaient toutes les deux liées à un risque très faible auparavant et apparaissent proches dans le graphique des individus.

Par application de la méthode de Ward, nous mettons en évidence la classification des modalités de la variable :

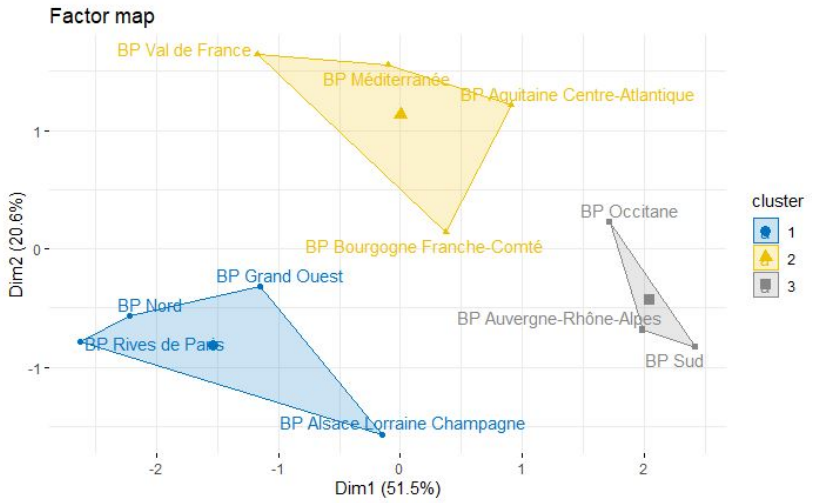

Figure 5.10: Représentation de la classification sous forme de clusters

Nous représentons aussi cette classification sous forme d'arborescence.

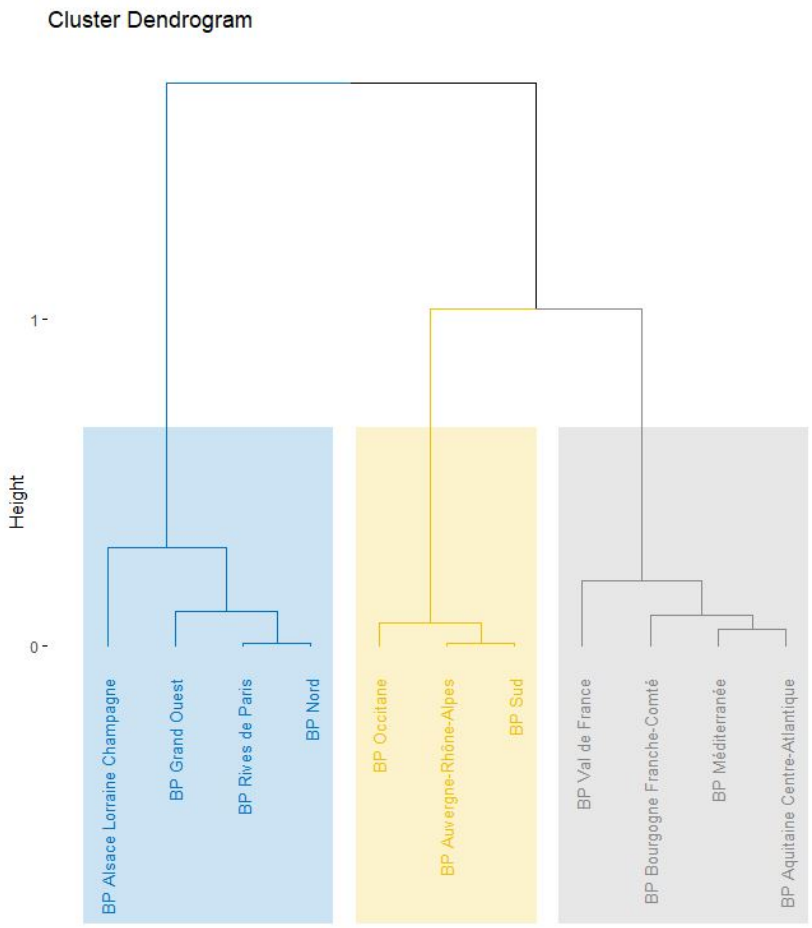

Figure 5.11: Représentation en arborescence de la classification de la variable Banque

Nous obtenons trois groupes que l'on peut ordonner par niveau de risque : un groupe de banques à risque faible (en bleu), un groupe de banques à risque élevé (en jaune), un groupe neutre (en gris). Nous obtenons aussi le résultat suivant :

$$\frac{\text{inertie inter-classe}}{\text{inertie totale}} = 45\% \tag{5.2}$$

A priori, cette classification des établissements bancaires ne parvient pas à expliquer la variabilité globale de la sinistralité.

5.2.3 Catégorie Socio-Professionnelle

Pour terminer cette partie de réduction des modalités des variables qualitatives, nous essaierons de procéder à un regroupement des modalités de la variable CSP.

Comme précédemment, nous veillons à bien effectuer toutes les étapes intermédiaires qui mènent à la classification finale et obtenons les résultats présentés ci-après.

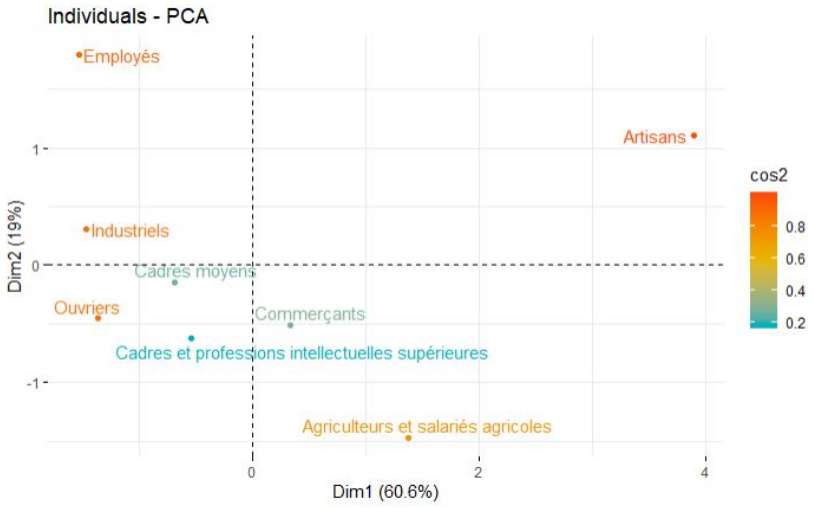

Figure 5.12: Position des individus CSP sur l'espace de l'ACP

Dans le cadre de l'étude mise en place dans cette sous-partie, nous avons retiré la catégorie *Personnel de service*. En effet, l'exposition était trop faible pour permettre d'étudier correctement les évolutions de la sinistralité annuelle entre 2014 et 2019.

Contrairement aux précédentes variables, nous constatons que le nombre de modalités mal représentées sur l'espace engendré par les deux premières composantes de l'ACP est plus important. Nous pouvons tout de même identifier des catégories isolées comme :

— Les Artisans, qui apparaissaient comme la CSP la plus exposée au risque dans l'analyse descriptive.
— Les Employés, qui apparaissaient comme la catégorie (avec une exposition suffisante) la moins exposée.

Par application de la méthode de Ward [9], nous obtenons l'arborescence présentée en début de page suivante.

9. cf. page 29 pour l'introduction à la méthode

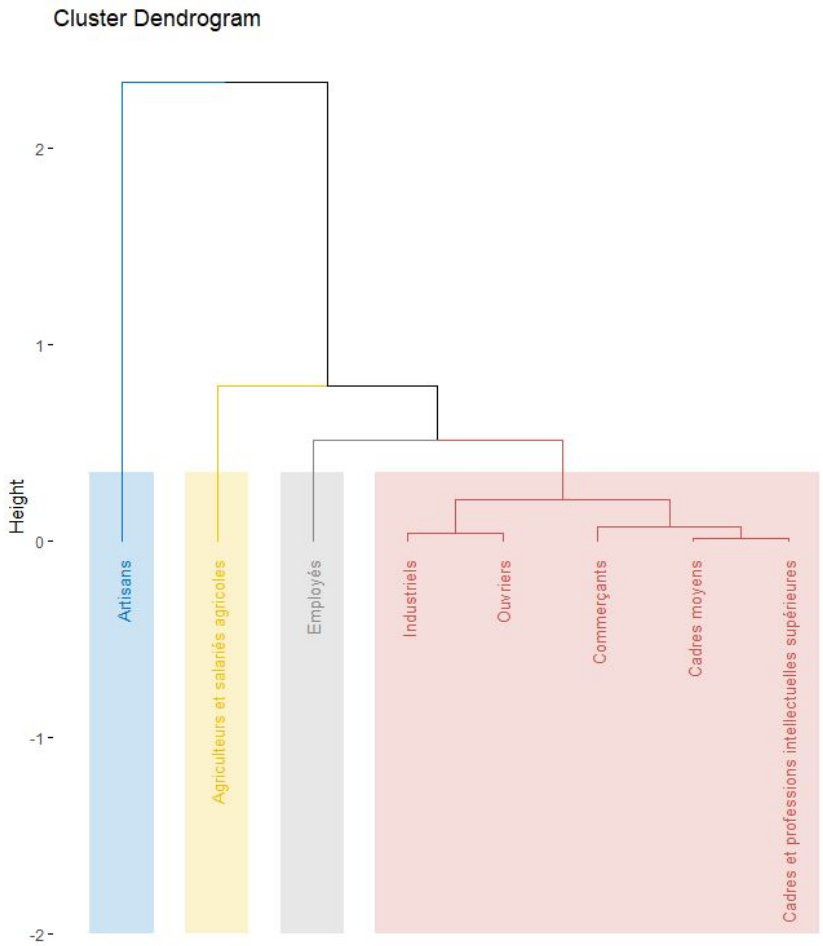

Figure 5.13: Représentation en arborescence des classes de modalités de la CSP

Nous obtenons au final cinq groupes de Catégories Socio-Professionnelles distincts. Nous remarquons que les Artisans se distinguent très tôt du reste des modalités ce qui signifie que leur profil de risque se détache bien de celui des autres individus.

Nous constatons aussi que les Agriculteurs constituent un groupe à part entière. Dans la partie d'analyse descriptive, ceux-là étaient le deuxième groupe le plus exposé au risque derrière les Artisans. De la même manière, les Employés sont une classe à part entière. Dans l'analyse descriptive, ceux-là étaient la catégorie la moins exposée au risque avec une exposition raisonnable.

Nous constatons que toutes les autres CSP sont regroupées dans la même classe, ce qui signifie qu'il est difficile d'isoler des profils de risque différents parmi celles-ci. Nous observons aussi le résultat suivant :

$$\frac{\text{inertie inter-classe}}{\text{inertie totale}} = 80\% \tag{5.3}$$

Les nouvelles classes de CSP créées parviennent donc bien à expliquer la variabilité totale des données.

5.2.4 Conclusion

Dans cette partie, les objectifs étaient les suivants :

1. Se familiariser avec certaines méthodes d'apprentissage non-supervisé.
2. Tenter de réduire les modalités des variables qualitatives.

Comme indiqué en introduction, la réduction des modalités accroît la significativité de la modélisation puisque nous nous assurons d'avoir un niveau d'exposition toujours acceptable.

Par application d'algorithmes de classification, nous sommes parvenus à créer des groupes de modalités pertinents pour les variables *Région* et *Catégorie Socio-Professionnelle*. Dans les deux cas, la part de variabilité totale expliquée par les groupes était importante et ceux-là semblaient pertinents. En effet, les groupes mis en évidence sont venus confirmer les résultats obtenus dans la partie d'analyse descriptive.

Pour la variable *Établissement*, nous ne sommes en revanche pas parvenus à créer des groupes de modalités significatifs. La part de variabilité totale expliquée était trop faible.

Dans les parties ultérieures du mémoire, nous travaillerons avec les variables et leurs modalités dans leur entièreté. En effet, c'est seulement au moment de la modélisation que nous ferons le choix des groupes de modalité que nous décidons de conserver. Ce choix dépendra des résultats obtenus jusqu'à cette étape mais aussi des résultats que nous présentons dans les parties suivantes.

5.3 Application à la création de groupes d'individus similaires

Dans cette partie, nous tenterons d'appliquer les méthodes d'apprentissage non-supervisé présentées précédemment avec pour but la création de classe d'individus (les assurés du portefeuille) similaires en terme de profil de risque. Nous sortons ainsi du cadre de la réduction des modalités et tentons d'élargir le cadre d'application des modèles de classification.

Notre démarche dans cette partie se construira de la manière suivante :

1. Dans un premier temps, nous procéderons à une Analyse Factorielle des Données Mixtes afin de mettre en évidence les structures de corrélation des variables de notre base et de réduire les dimensions de travail de notre base d'étude.
2. Dans un seconds temps, nous appliquerons l'algorithme des K-Means pour créer des classes d'individus similaires projetés dans le nouvel espace défini par AFDM.
3. Pour terminer, nous procéderons à une analyse descriptive des différents groupes d'individus créés et essaierons de mettre en lumière des caractéristiques propres aux individus en fonction de leur niveau d'exposition au risque.

5.3.1 Analyse Factorielle des Données Mixtes

L'objectif de cette première sous-partie est de mettre en évidence les structures de corrélation entre les variables de notre base de travail, puis d'identifier un sous-espace de travail sur lequel seront projetés les individus de notre portefeuille et à partir duquel les futures classes d'individus seront créées.

L'un des avantages de procéder à une analyse factorielle des données mixtes est d'étudier en même temps les variables quantitatives et qualitatives de notre base [10].

Dans le cadre de cette étude, nous faisons le choix de ne conserver que les variables qualitatives et quantitatives les plus importantes et les plus susceptibles d'apporter des résultats intéressants. Voici ce à quoi ressemble notre base de travail :

	age	anciennete	mnt_gar	nom_region	nom_etab	lib_csp	sel_med	duree_franchise
1:	55	20	1253.03	Normandie	BP Grand Ouest	Artisans	N	30
2:	60	20	2578.53	Centre-Val de Loire	BP Val de France Cadres et professions intellectuelles supérieures	O	30
3:	56	19	1551.34	Normandie	BP Grand Ouest	Artisans	N	30
4:	64	10	1137.30	Ile-de-France	BP Rives de Paris	Commerçants	N	30
5:	60	19	2578.53	Auvergne-Rhône-Alpes	BP Auvergne-Rhône-Alpes	Industriels	O	30

111167:	46	0	10800.00	Nouvelle-Aquitaine	BP Aquitaine Centre-Atlantique	<NA>	N	715
111168:	31	0	14400.00	Pays de la Loire	BP Grand Ouest	<NA>	N	30
111169:	29	0	7343.52	Auvergne-Rhône-Alpes	BP Auvergne-Rhône-Alpes	<NA>	N	715
111170:	24	0	7200.00	Auvergne-Rhône-Alpes	BP Auvergne-Rhône-Alpes	<NA>	N	715
111171:	30	0	9791.36	Auvergne-Rhône-Alpes	BP Auvergne-Rhône-Alpes	<NA>	N	715

Figure 5.14: Base de travail pour l'AFDM

10. cf. page 32 pour l'introduction au procédé.

Choix du nombre de composantes principales à conserver

Après application de l'algorithme, il s'agit de décider du nombre de composantes dimensionnelles à conserver. Pour cela, nous pouvons nous aider du tableau récapitulatif des valeurs propres et de la variabilité expliquée par composante :

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.24376565	6.35322364	6.353224
Dim.2	2.01462588	5.70441425	12.057638
Dim.3	1.98275623	5.61417533	17.671813
Dim.4	1.97437872	5.59045442	23.262268
Dim.5	1.94453424	5.50594974	28.768217
Dim.6	1.90605043	5.39698282	34.165200
Dim.7	1.85349883	5.24818295	39.413383
Dim.8	1.83077010	5.18382653	44.597210
Dim.9	1.76863721	5.00789721	49.605107
Dim.10	1.60894475	4.55572792	54.160835
Dim.11	1.23814121	3.50579750	57.666632
Dim.12	1.08923602	3.08417237	60.750805
Dim.13	1.04580768	2.96120502	63.712010
Dim.14	1.02984745	2.91601362	66.628023
Dim.15	1.01767146	2.88153729	69.509561
Dim.16	1.00388519	2.84250146	72.352062
Dim.17	0.99366531	2.81356386	75.165626
Dim.18	0.98772257	2.79673699	77.962363
Dim.19	0.98337925	2.78443887	80.746802
Dim.20	0.96627554	2.73600970	83.482812
Dim.21	0.95420237	2.70182452	86.184636
Dim.22	0.91265255	2.58417618	88.768812
Dim.23	0.82486269	2.33559915	91.104411
Dim.24	0.76747436	2.17310408	93.277515
Dim.25	0.65698907	1.86026490	95.137780
Dim.26	0.41607483	1.17811611	96.315896
Dim.27	0.33787292	0.95668734	97.272584
Dim.28	0.23317457	0.66023392	97.932818
Dim.29	0.20076606	0.56846921	98.501287
Dim.30	0.17542556	0.49671758	98.998004
Dim.31	0.12814278	0.36283636	99.360841
Dim.32	0.08499484	0.24066293	99.601504
Dim.33	0.05304735	0.15020361	99.751707
Dim.34	0.03941270	0.11159707	99.863304
Dim.35	0.02539920	0.07191784	99.935222

Figure 5.15: Tableau des valeurs propres

Nous décidons d'appliquer le critère de Kaiser [11] et de conserver ainsi seulement les composantes pour lesquelles la valeur propre associée est supérieure à 1. Cela indique que la composante principale concernée représente plus de variance par rapport à une seule variable d'origine. Nous conservons donc les seize premières.

Analyse des structures de corrélations des variables

À cette étape, nous pouvons étudier les structures de corrélations des variables par projection sur les deux premières composantes principales de l'AFDM. Bien que les variables ne soient pas toujours bien représentées, il est utile à ce stade de formuler une première interprétation a priori du comportement de celles-ci dans la base.

11. cf. xxx

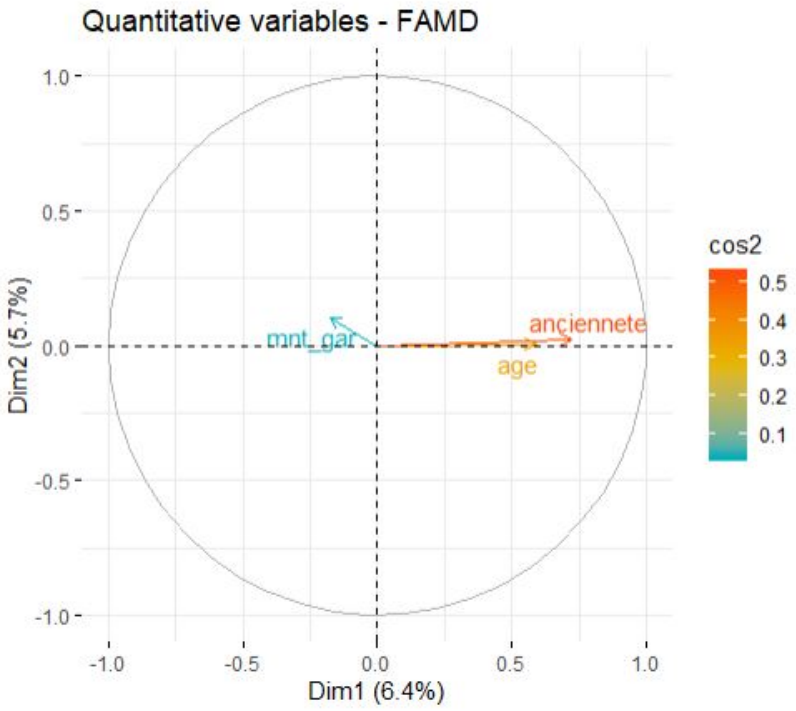

Figure 5.16: Analyse des corrélations des variables quantitatives

Comme indiqué sur le graphique, nous disposons de trois variables quantitatives dans notre base de travail : l'âge, l'ancienneté et le montant garanti.

Nous constatons que les variables anciennetés et âge semblent fortement corrélées. En effet, elles sont toutes les deux relativement bien représentées par les deux premières composantes principales. De plus, les vecteurs les représentants sont alignés et de même sens.

En ce qui concerne les variables qualitatives, nous en avons à notre disposition un bien plus grand nombre. Lors de la représentation en deux dimensions, ce seront les modalités de ces variables qui seront représentés.

Ainsi, nous n'allons non pas nous contenter d'interpréter la seule projection sur les deux premières dimensions mais aussi les projections sur d'autres composantes, dans le but de capter un maximum d'information sur les structures de corrélation entre les différentes variables qualitatives.

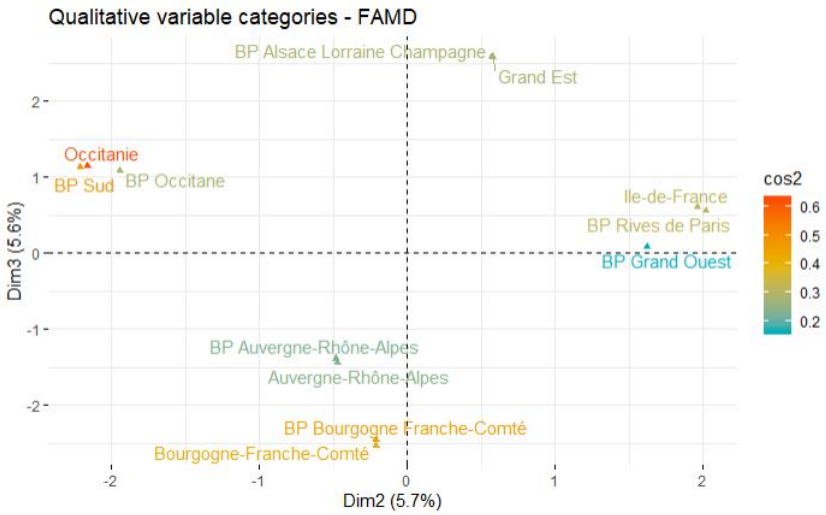

Figure 5.17: Analyse des corrélations des variables qualitatives (Composantes 2 et 3)

Nous constatons à la lecture du graphique qu'il semble y avoir un fort schéma de corrélation entre variables de régions et variables d'entités bancaires. En effet, les banques et les zones auxquelles elles correspondent se retrouvent aux mêmes endroits du graphe. Ainsi, il semblerait par exemple qu'un individu habitant en Île-de-France présente les mêmes caractéristiques qu'un individu ayant souscrit son contrat à la banque BP Rives de Paris.

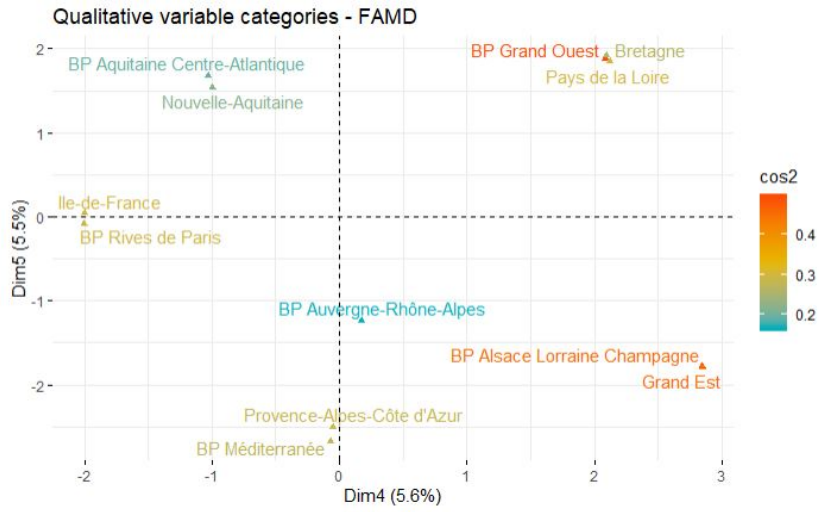

Figure 5.18: Analyse des corrélations des variables qualitatives (Composantes 4 et 5)

Ce résultat se confirme ici. En effet, nous retrouvons bien à nouveau les paires de banques et de régions associées. Cette observation confirme nos remarques sur les taux d'expatriation [12]. La grande majorité des individus semble vivre dans la région de dominance de leur banque ce qui justifie la corrélation entre ces deux variables.

12. cf. page 15.

Néanmoins, pour un certain nombre d'entités bancaires, plusieurs régions apparaissent associées (c'est le cas de la BP Grand Ouest sur le graphique). Ainsi, nous pourrions observer des effets comportementaux à l'échelle des régions et pas nécessairement à l'échelle des banques et inversement.

Nous représenterons ci-dessous un graphique résumé de tous les résultats précédent, qui combine variables qualitatives et variables quantitatives.

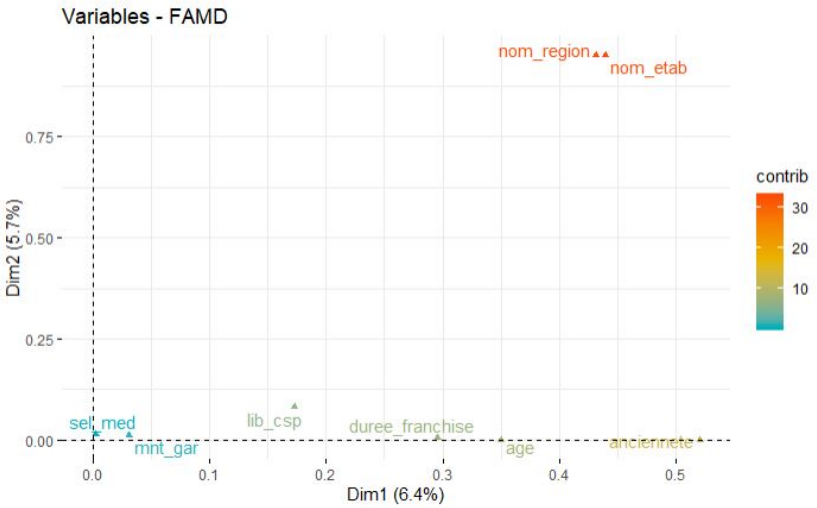

Figure 5.19: Variables qualitatives et quantitatives

De l'analyse du graphique, nous pouvons tirer plusieurs conclusions :

— Il existe une corrélation apparente entre les régions et les entités bancaires.
— Les variables de catégories socio-professionnelles, durée de franchise, âge et ancienneté dessinent bien des groupes d'individus distincts.
— Les variables de sélections médicales et de montant garanti sont globalement mal représentés sur les dimensions de l'AFDM et ne permettent pas ou peu de différencier les individus entre eux.

Coloration des individus

Avant de procéder à la partition de notre portefeuille, nous représentons la projection de tous les assurés sur les deux premières composantes de l'AFDM et effectuons des colorations en fonction de certaines des variables qualitatives.

L'intérêt pour nous est d'une part, de mieux comprendre les données qui constituent notre portefeuille d'assurance ; d'autre part, d'identifier des schémas dans la répartition des modalités des variables qualitatives au sein de la population.

Nous représentons d'abord une coloration des individus par catégorie socio-professionnelle.

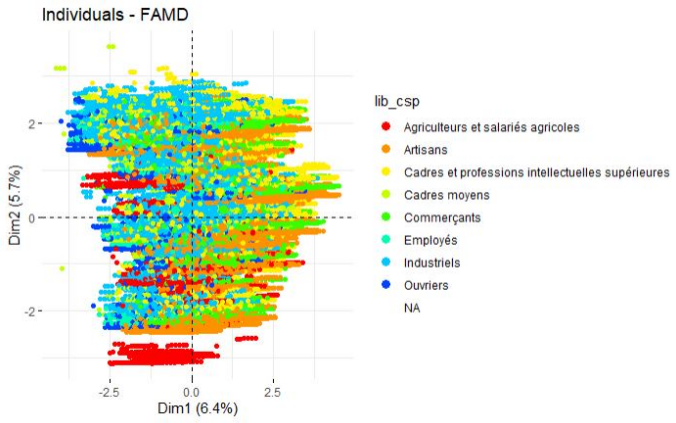

Figure 5.20: Coloration des individus par CSP

Nous représentons ensuite une coloration des individus par région et par entité bancaire.

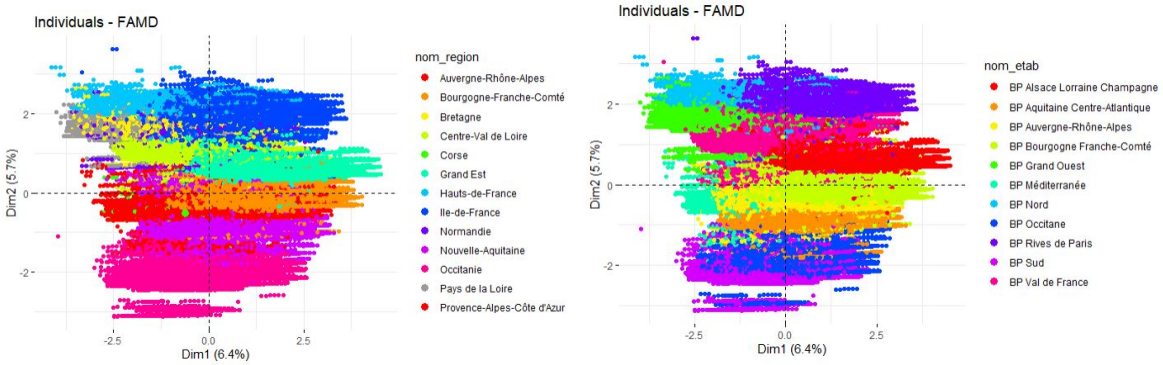

Figure 5.21: Coloration des individus par région et par banque

L'étude de ces graphiques se fait en deux dimensions : cette représentation peut s'avérer trompeuse puisque nous conservons en réalité seize dimensions uniques pour la classification future. Toutes les variables ne sont ainsi pas nécessairement bien représentées par les axes affichés ; les points visibles ne sont que des projections qui ne reflètent pas toujours parfaitement la réalité.

Par étude des colorations, nous constatons qu'il est difficile d'isoler les modalités les unes par rapport aux autres que ce soit pour la CSP, la Région ou l'Établissement. En effet, la plupart des points (qui représentent les individus) sont superposés et il paraît difficilement concevable que l'on puisse déterminer avec certitude la région d'appartenance d'un assuré par exemple à partir de sa seule position sur les deux premières composantes principales.

Néanmoins, nous observons quand même de vraies disparités de répartition entre les modalités d'une même variable et, même s'il paraît compliqué d'isoler des modalités par rapport à d'autres, il apparaît quand même que certaines modalités sont plus représentées à certains endroits du graphiques qu'à d'autres. Ainsi, cela n'enlève en aucun cas de l'intérêt aux variables en tant que variables explicatives et il sera intéressant de mieux comprendre l'impact de ces dernières sur la sinistralité.

5.3.2 Classification et analyse des groupes d'individus

Comme annoncé précédemment, nous passons maintenant à la phase de création des groupes d'individus aux caractéristiques similaires et à l'étude de la sinistralité pour chacun de ces groupes.

Pour ce faire, nous utiliserons l'étude menée à l'aide de l'Analyse Factorielle de Données Mixtes. En effet, par application de l'algorithme d'AFDM, nous avons pu réduire les dimensions de notre espace de travail tout en conservant une bonne modélisation de la variabilité totale de nos données. Les assurés de notre portefeuille, auparavant définis par un espace à trente-cinq dimensions, sont représentés maintenant par leur projection sur un espace à seulement seize dimensions.

C'est sur cet espace à seize dimensions que nous allons appliquer un algorithme des K-Means [13] afin d'isoler des individus aux caractéristiques similaires. Pour rappel, l'algorithme des K-Means fonctionne simplement en calculant les distances euclidiennes entre tous les individus et en groupant les individus qui présentent les distances les plus faibles entre eux.

Pour appliquer l'algorithme des K-Means, il est également nécessaire de choisir un nombre final de classes. Idéalement, nous aimerions disposer de trois grandes classes d'individus : les individus peu exposés, les individus neutres et les individus exposés. Néanmoins, si l'on choisit de renseigner un paramètre égal à 3 pour l'application de l'algorithme, nous n'allons au final rien obtenir de significatif mais simplement une représentation moyenne à échelle réduite du portefeuille global.

Ainsi, nous choisissons de procéder de la manière suivante :

1. D'abord, nous appliquerons un algorithme des K-Means sur un nombre de classes égal à 30. Nous validerons également le modèle par étude de l'inertie inter-classe.
2. Ensuite, nous étudierons la répartition de la sinistralité entre ces nouvelles classes et les fusionneront en 3 grandes classes, ordonnée par risque.
3. Enfin, nous procéderons à une analyse descriptive de chaque classe pour mettre en évidence des caractéristiques propres à chacune d'elles.

13. cf. page 30 pour l'introduction au modèle.

Application de K-Means

Quant à l'application de l'algorithme, il est important de préciser que les observations ont été pondérées par leur temps de présence en portefeuille. Ensuite, comme précisé avant, nous obtenons au final bien trente classes d'individu distinctes et procédons au calcul des inerties inter-classes, intra-classes et totale [14]. Nous obtenons finalement :

$$\frac{\text{inertie inter-classe}}{\text{inertie totale}} = 79.73\% \tag{5.4}$$

Cela signifie que les classes d'individus mises en évidence par l'algorithme permettent de bien expliquer la variabilité totale des données. Il s'agit néanmoins de ne pas surestimer les performances de l'algorithme. En effet, il faut être conscient que la réduction des dimensions par l'analyse factorielle des données mixtes a dans un premier temps entrainé la perte d'une partie expliquée de la variabilité totale ; l'action est réitérée ici par application de K-Means.

Cette perte de variabilité expliquée lors de l'application des algorithmes importe finalement peu puisque notre objectif principal est d'obtenir des classes d'individus avec des disparités en terme de profil de risque. À partir de là, nous pourrons procéder comme nous l'expliquions à l'analyse descriptive de ces classes et ainsi mettre en évidence des caractéristiques propres à une sinistralité plus forte ou plus faible.

Création des classes de sinistralité

Pour créer trois classes de sinistralité à partir des trentes classes obtenues après application de K-Means, il convient simplement de fixer des seuils arbitraires de sinistralité en fonction de la sinistralité globale du portefeuille. Avant toute chose, voici la distribution de la sinistralité par cluster créé par K-Means :

cluster	tx_sin	N
26	2.87	1323.86
7	3.24	616.64
8	3.39	1180.65
9	3.41	5813.03
25	3.59	2976.55
20	3.65	1889.79
30	3.65	2686.71
16	3.68	895.97
4	3.79	2667.93
24	4.10	6029.35
12	4.23	4255.18
10	4.31	2972.78
5	4.39	2959.32
13	4.61	1062.05
14	4.62	9944.91
29	4.65	1161.78
22	4.66	3474.94
27	4.70	7313.04
3	4.96	6674.66
11	5.13	7595.03
2	5.23	1681.91
15	5.43	2100.07
1	5.50	1090.86
23	5.58	2366.58
6	5.60	9220.51
18	5.89	763.99
17	6.07	5403.79
21	6.39	3836.09
19	6.59	516.14
28	6.84	1300.90

Figure 5.22: Distribution de la sinistralité au sein des clusters

14. cf. page 29 pour l'introduction aux notions d'inertie

Nous observons que la sinistralité semble être bien distribuée parmi toutes les classes créées par K-Means. En effet, les taux d'incidence par classe varient entre 2.87% et 6.84%. À partir de là, nous souhaitons créer trois grands groupes d'individus plus ou moins exposés au risque. Pour ce faire, nous partons simplement du taux de sinistralité global sur le portefeuille (qui est de 4.7%) et nous définissons arbitrairement les 3 groupes de risque suivants :

— Groupe peu risqué pour les classes d'individus ayant un taux d'incidence inférieur à 4%.
— Groupe neutre pour les classes d'individus ayant un taux d'incidence entre 4% et 5.4%.
— Groupe risqué pour les classes d'individus ayant un taux d'incidence supérieur à 5.4%.

Nous vérifions ensuite que l'exposition est assez élevée dans chacune des classes.

Groupe peu risqué	**Groupe neutre**	**Groupe risqué**
20 052	55 125	26 599

Table 5.1: Exposition par groupe

La majorité des individus se situent dans le groupe neutre mais l'exposition reste bien suffisante dans les groupes risqués et peu risqués. Ainsi, nous pouvons procéder comme expliqué précédemment à l'analyse descriptive de chacun de ces groupes [15] et essayer d'identifier des caractéristiques déterminantes pour l'incidence.

Analyse descriptive du groupe de sinistralité faible

À l'instar de ce que nous avons fait dans la partie d'analyse descriptive [16], nous chercherons ici à mettre en avant des caractéristiques propres aux individus peu risqués par l'étude des répartition des modalités des variables de notre base de travail. Dans tous les graphiques suivants, nous comparerons donc les distributions observées à la distribution globale du portefeuille. Par positionnement, nous serons capables de tirer des conclusions sur les profils de risque des assurés.

Ainsi, dans tous les graphiques suivants, la courbe bleue représentera la distribution de la sinistralité au sein du cluster. La courbe orange quant à elle, représentera la distribution de la sinistralité au sein du portefeuille global.

15. En réalité, nous nous limiterons à l'analyse du groupe peu risqué et du groupe risqué puisque le groupe neutre ne représente qu'une moyenne du portefeuille.
16. cf. page 20.

Nous commençons par étudier la distribution de l'âge et de l'ancienneté.

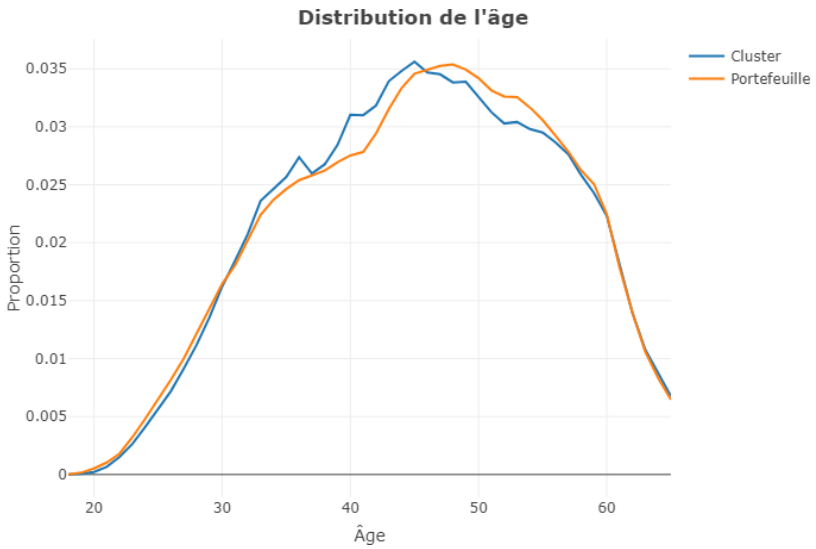

Figure 5.23: Distribution de l'âge pour le cluster de sinistralité faible

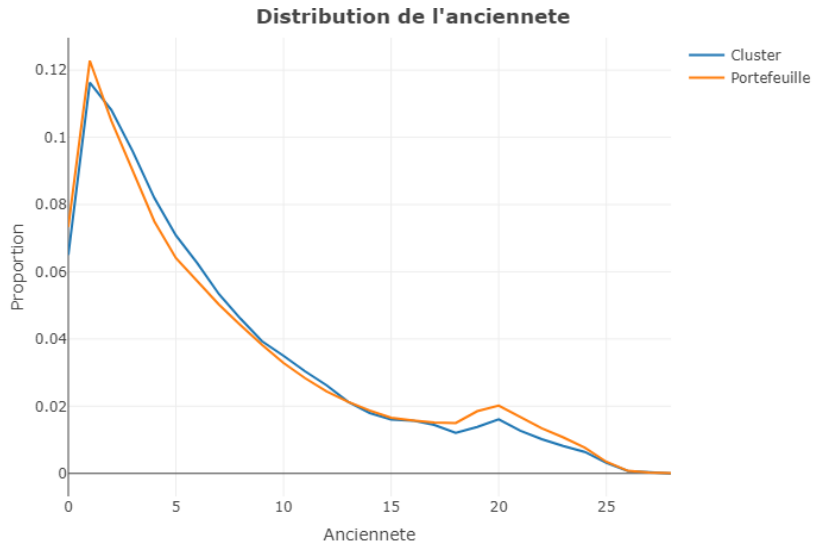

Figure 5.24: Distribution de l'âge pour le cluster de sinistralité faible

Nous constatons ici que l'âge et l'ancienneté se répartissent pour le cluster de sinistralité faible comme pour le portefeuille au global.

Ainsi, malgré le phénomène logique d'augmentation de la sinistralité avec l'âge observé dans la partie des statistiques descriptives, nous ne pouvons pas ici tirer de conclusion sur l'influence de la variable quant à l'incidence en arrêt de travail. La conclusion est la même pour la variable ancienneté.

Nous étudions ensuite la distribution des régions au sein du cluster de sinistralité faible.

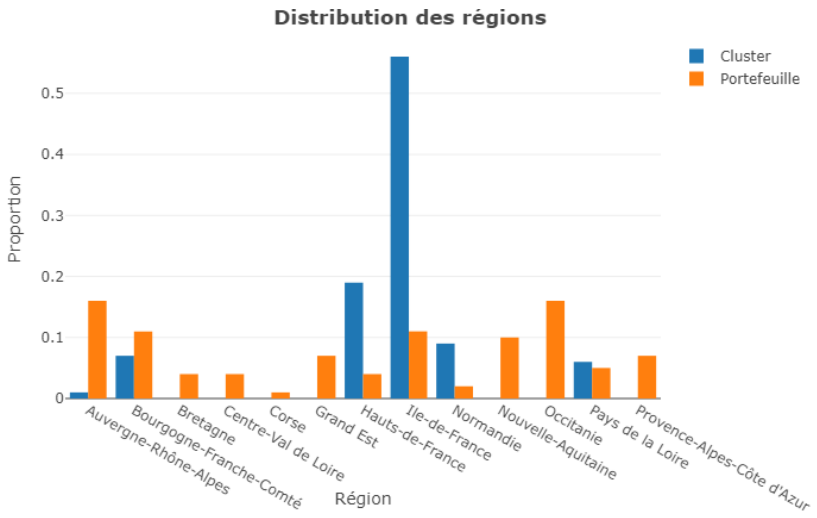

Figure 5.25: Distribution des régions pour le cluster de sinistralité faible

En ce qui concerne la distribution des régions, le résultat est très intéressant. En effet, la grande majorité du portefeuille de sinistralité faible est composée d'individus vivants en Ile-de-France, Hauts-de-France ou Normandie.

De plus, certaines régions ne sont pas ou quasiment pas représentés comme : Auvergne-Rhône-Alpes, Bretagne, Centre-Val-de-Loire, Corse, Grand Est, Nouvelle-Aquitaine, Occitanie, Provence-Alpes-Côte d'Azur.

Ces résultats sont intéressants d'une part puisqu'ils montrent une distribution non-équilibrée au sein des régions et ainsi permettent d'isoler des régions a priori plus ou moins risquées par rapport aux autres ; d'autre part car ils viennent confirmer les résultats intermédiaires que nous avons observés dans la partie de statistiques descriptives.

Les individus vivants dans des régions du Nord de la France apparaissent ainsi moins risqués que les individus vivants dans d'autres régions.

Nous étudions maintenant la distribution des établissements.

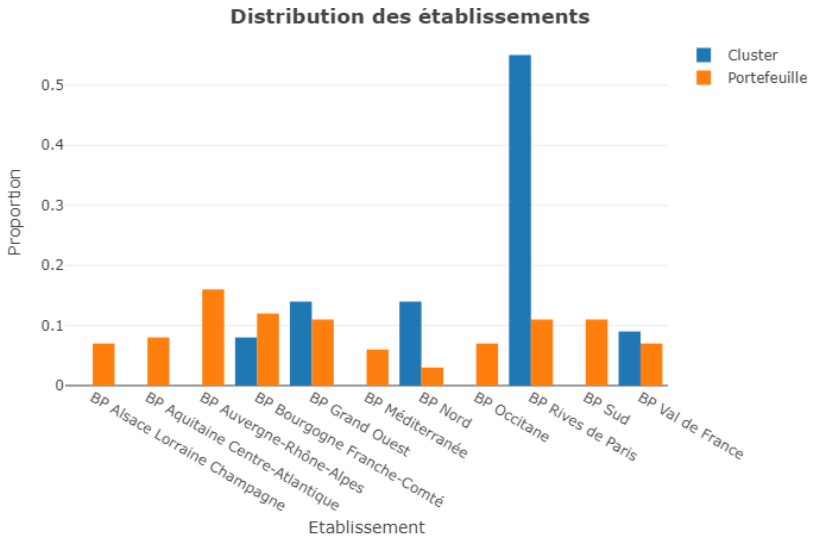

Figure 5.26: Distribution des banque pour le cluster de sinistralité faible

Le résultat obtenu sur les régions est confirmé ici par la distribution des banques au sein du cluster de sinistralité faible. En effet, les banques les plus représentées sont les banques parisiennes et les banques du nord. À l'inverse, les autres banques et en particulier les banques du sud ne sont pas du tout représentées.

Nous étudions maintenant la distribution des montants garantis et de la sélection médicale.

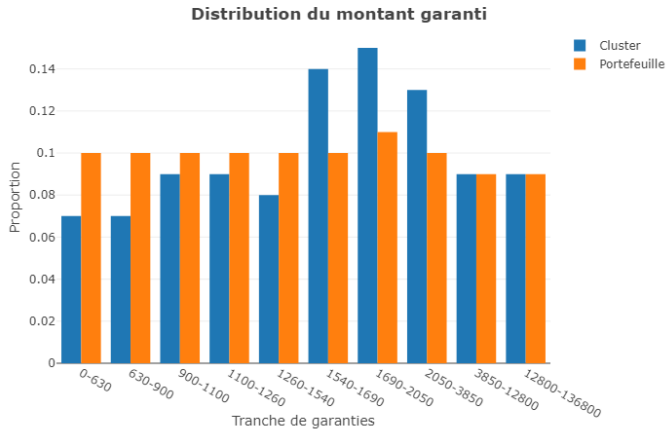

Figure 5.27: Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité faible

En ce qui concerne les montants garantis, nous observons que certains sont légèrement plus représentés dans le cluster de sinistralité faible que dans le portefeuille. Néanmoins, cela reste trop faible pour que nous puissions en tirer des conclusions.

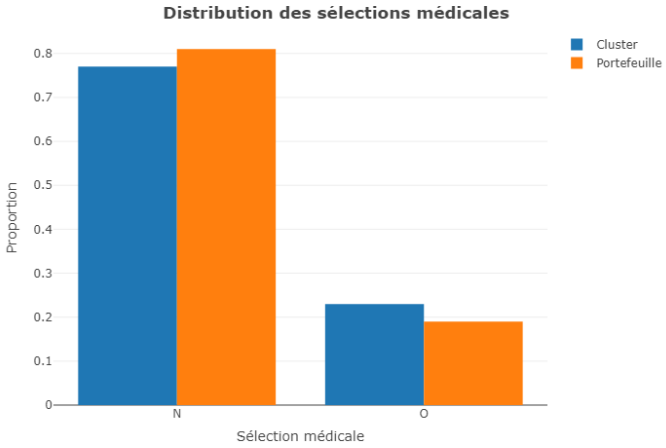

Figure 5.28: Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité faible

Pour la sélection médicale, nous constatons que la distribution au sein du cluster est sensiblement la même que celle du portefeuille. Nous ne pouvons donc pas tirer de vraies conclusions sur l'effet de la sélection médicale.

Nous étudions maintenant la distribution des Catégories Socio-Professionnelles.

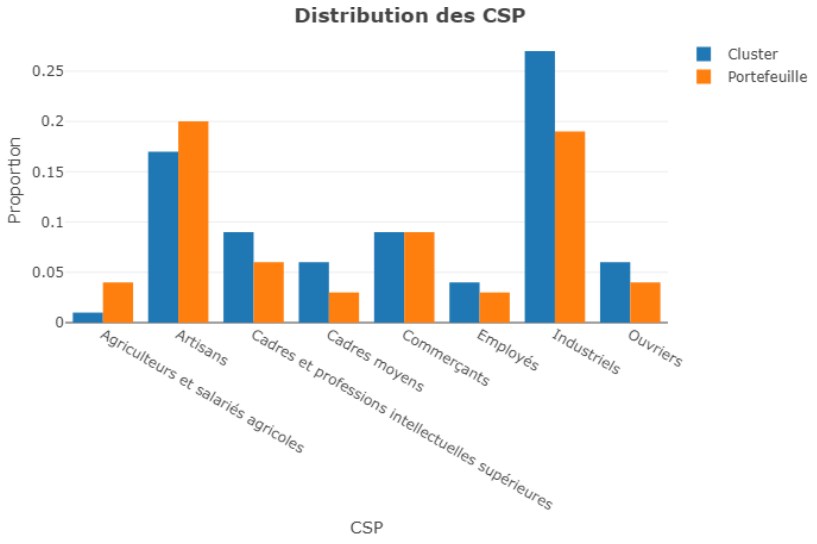

Figure 5.29: Distribution des CSP pour le cluster de sinistralité faible

En ce qui concerne la distribution des catégories socio-professionnelles, il semblerait que les Industriels soient plus représentés dans le cluster de sinistralité faible que dans le portefeuille. Le même résultat se présente pour les Cadres moyens, les Employés et les Ouvriers.

À l'inverse, les Artisans sont légèrement sous-représentés. De même pour les Agriculteurs. Ceux-là apparaissent ainsi comme des CSP plus risquées.

Bien que ces résultats confirment les observations faites précédemment, nous ne pouvons pas tirer de réelles conclusions ici puisque les écarts restent faibles.

Analyse descriptive du groupe de sinistralité forte

Maintenant que nous avons analysé les caractéristiques des individus du cluster de sinistralité faible, procédons à l'analyse des caractéristiques des individus du cluster de sinistralité forte.

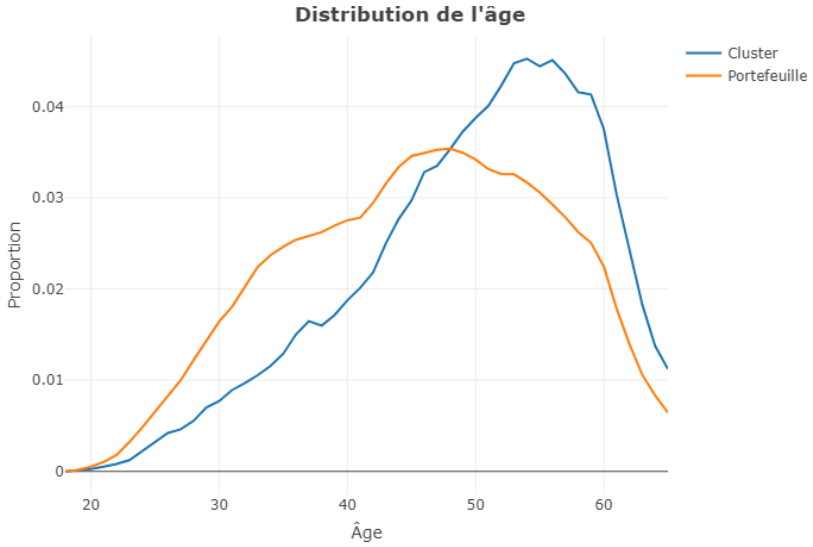

Figure 5.30: Distribution de l'âge pour le cluster de sinistralité forte

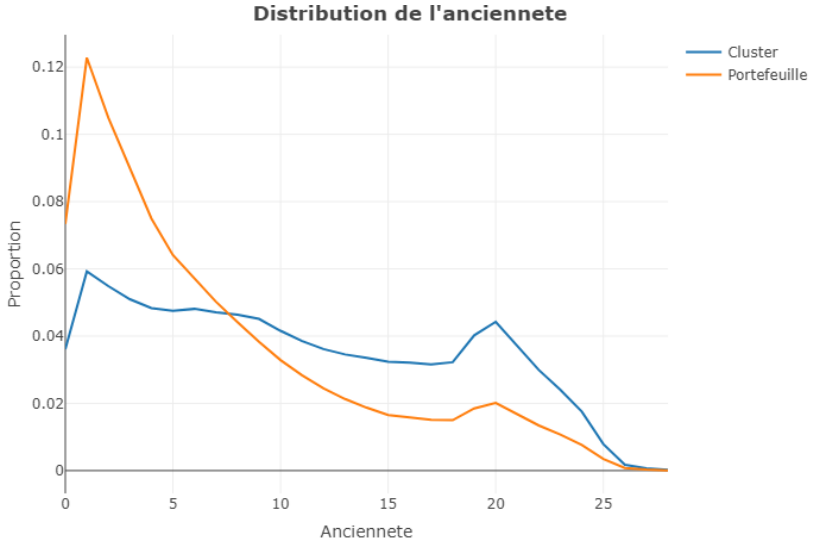

Figure 5.31: Distribution de l'âge pour le cluster de sinistralité forte

La courbe bleue représente à nouveau la distribution au sein du groupe et la courbe orange la distribution au sein du portefeuille global.

Nous constatons que la répartition de l'âge est très fortement marquée vers la droite. La part d'individus âgés dans le groupe d'individus risqué est donc beaucoup plus forte que la part d'individus âgés dans le portefeuille global. Cela vient confirmer les résultats que nous observions dans la partie d'analyse descriptive.

En ce qui concerne l'ancienneté, le résultat est là encore intéressant. Nous constatons qu'on trouve dans le groupe d'individus risqués beaucoup moins d'individus à ancienneté faible que dans le portefeuille global. Le résultat est au final assez cohérent pour deux raisons : d'une part, plus le temps passé en portefeuille augmente plus la probabilité d'avoir déjà eu un sinistre augmente ; d'autre part, les individus avec une ancienneté forte ont nécessairement un âge plus avancé.

Nous poursuivons avec l'analyse de la distribution des régions.

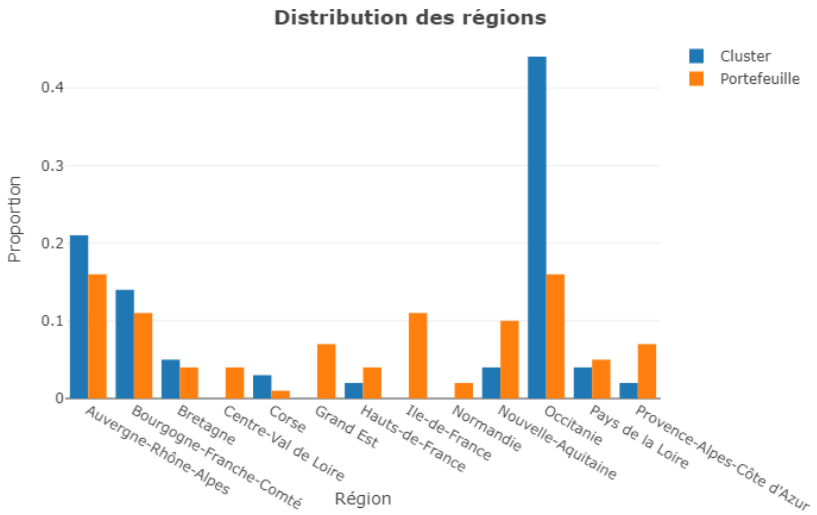

Figure 5.32: Distribution des régions pour le cluster de sinistralité forte

En ce qui concerne la distribution des régions, le résultat est remarquable puisqu'il s'agit du parfait opposé du résultat obtenu sur le groupe de sinistralité faible.

Nous retrouvons bien une surreprésentation des régions du sud (en particulier ici des régions Occitanie et Corse) et une sous-représentation des régions du nord.

Néanmoins, la région Provence-Alpes-Côte d'Azur est également sous-représentée ce qui pourrait signifier qu'elle n'est pas assez risquée pour être dans le groupe de sinistralité forte mais trop risquée pour être dans le groupe à sinistralité faible.

Nous continuons avec l'analyse de la distribution des établissements.

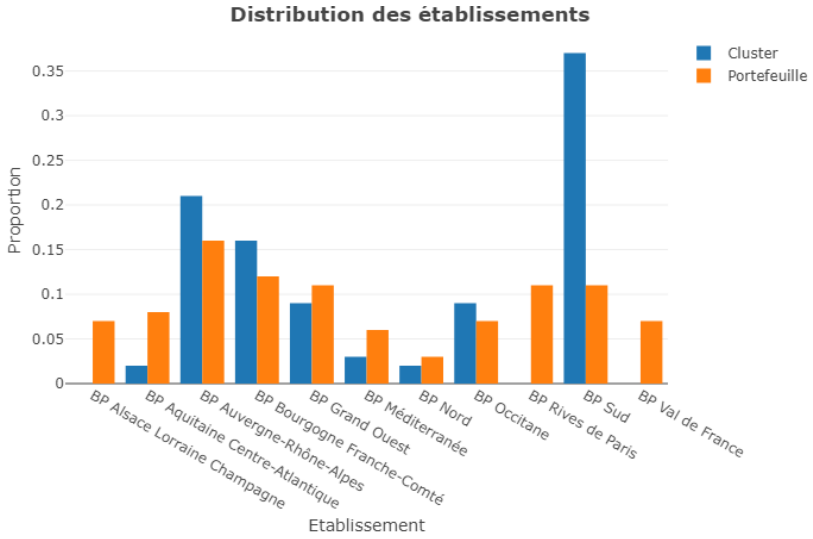

Figure 5.33: Distribution des banques pour le cluster de sinistralité forte

Ici encore, les résultats observés sur les régions sont confirmés par la distribution observée sur les banques. Les banques du sud sont en effet les plus représentées et les banques du nord sont les moins représentées.

Nous continuons avec l'anayse de la distribution des montants garantis.

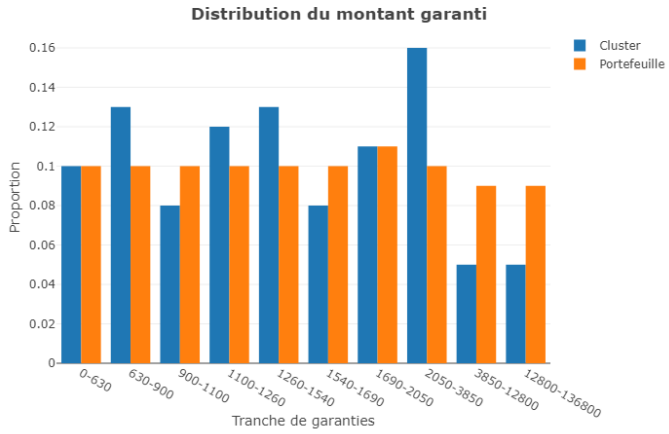

Figure 5.34: Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité forte

Nous constatons qu'il est difficile d'isoler des groupes de montants garantis spécifiques pour le groupe de sinistralité forte. Le résultat est en fait sensiblement le même que celui pour le groupe de sinistralité faible. Certes, nous observons certaines disparités dans la répartition mais celles-ci restent relativement faibles et il n'est pas évident d'établir un lien direct avec la sinistralité.

Nous continuons avec l'étude de la distribution de la sélection médicale.

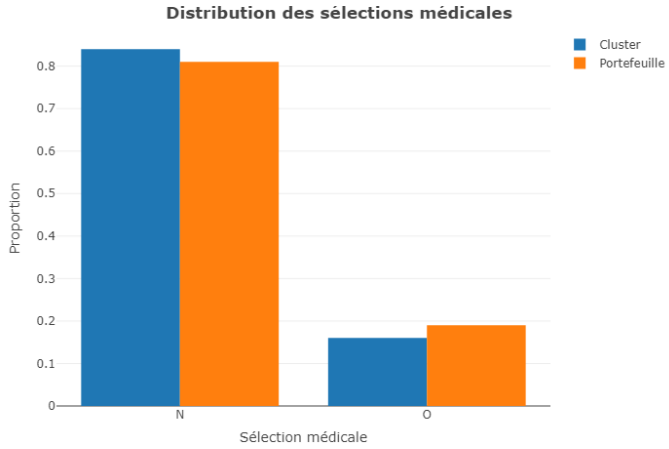

Figure 5.35: Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité forte

Nous observons que la distribution au sein du groupe de risque correspond relativement à la distribution globale du portefeuille. Nous ne pouvons donc pas tirer des conclusions sur l'impact de la variable sur la sinistralité ; ce résultat est cohérent avec ceux obtenus précédemment.

Nous étudions maintenant la distribution des Catégories Socio-Professionnelles.

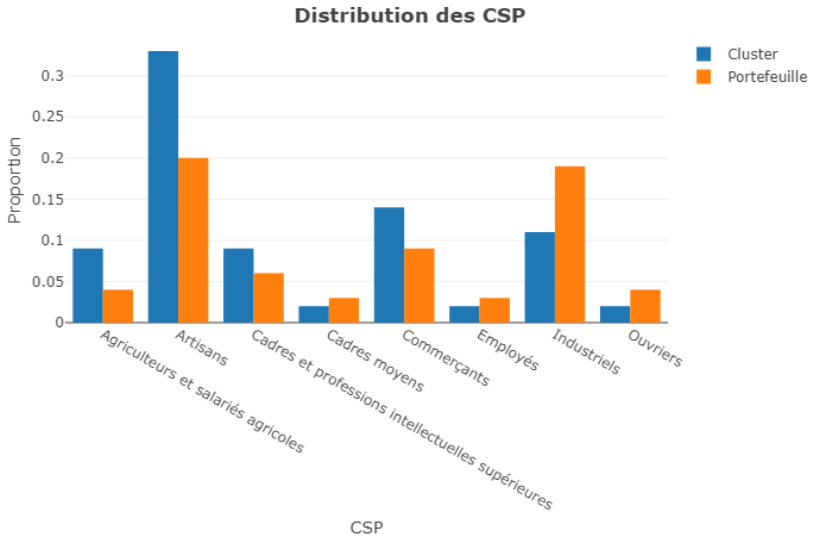

Figure 5.36: Distribution des CSP pour le cluster de sinistralité forte

Les résultats obtenus sur les catégories socio-professionnelles confirment égalementles résultats précédents. Nous observons que les Artisans, les Commerçants et les Agriculteurs sont surreprésentés dans le groupe de sinistralité forte. De la même manière, certaines catégories apparaissent sous-représentées. C'est le cas des Industriels par exemple, qui ressortaient dans les études précédentes comme une catégorie peu risquée.

5.3.3 Conclusion

Concernant la partie d'application de méthodes d'apprentissage non-supervisé à la création de groupes d'individus similaires -

En conclusion, nous sommes bien parvenus à identifier des groupes d'assurés qui présentaient des profils de risque similaires au sein du portefeuille. Nous avons pu identifier des modalités de variables qui semblent caractériser la propension d'un individu à subir un arrêt de travail. Nous pouvons résumer les résultats dans le tableau suivant :

Variable	**Groupe peu risqué**	**Groupe risqué**
Âge	$<$ 45ans	$>$ 45ans
Ancienneté	Faible	Forte
Région	Nord de la France	Sud de la France
Banque	Banques du nord	Banques du sud
CSP	Industriels ou Employés	Artisans ou Agriculteurs

Table 5.2: Caractéristiques des groupes de risque identifiés

Les résultats obtenus confirment :

— D'une part, les résultats obtenus dans la partie d'analyse descriptive.
— D'autre part, la pertinence des groupes de modalités créés en début de partie.

Avant de passer à l'étape de modélisation de la loi d'incidence, nous allons tenter de discuter encore une fois les résultats obtenus. Pour ce faire, nous utiliserons un modèle d'apprentissage supervisé.

6 Méthodes d'apprentissage supervisé

Dans cette partie, l'objectif est de terminer la phase d'identification de variables discriminantes pour l'incidence en arrêt de travail.

Pour rappel, nous avons déjà procédé à une analyse descriptive de notre portefeuille d'assurés, ainsi qu'à l'application de divers modèles d'apprentissage non-supervisé. Nous avons obtenu des résultats significatifs que nous souhaitons discuter une ultime fois dans cette partie.

Ici, nous porterons notre attention sur une méthode d'apprentissage supervisé. Une méthode d'apprentissage supervisé est un algorithme de type prédictif. Jusqu'à maintenant, les algorithmes que nous avons utilisés étaient des algorithmes de classification : par groupement des données, nous étions en mesure de tirer des conclusions.

Dans cette partie, nous chercherons à prédire l'incidence en incapacité pour des individus donnés, en fonction des caractéristiques qui leurs sont propres.

À ce stade, il est important de rappeler que, bien que l'objectif du mémoire soit lié à la prédiction de l'incidence en arrêt de travail, nous n'utiliserons pas les méthodes présentées ici directement à des fins utiles de tarification. L'objectif unique de cette partie est un objectif complémentaire aux travaux effectués jusqu'ici. C'est à dire que nous continuons de chercher à identifier des caractéristiques qui permettent de déterminer si les individus sont plus susceptibles de subir des arrêts de travail.

6.1 Les Modèles Linéaires Généralisés

Dans cette partie du mémoire, nous vous présenterons une méthode d'apprentissage supervisé que nous voulions utiliser : une approche par Modèles Linéaires Généralisés (2000) [5] . Même si nous n'en parlerons pas dans ce mémoire, une approche par utilisation de différents modèles de Machine Learning aurait pu être envisagée. Nous suivrons les étapes suivantes :

1. Dans cette première sous-partie, nous introduirons les Modèles Linéaires Généralisés dans leur ensemble.
2. Nous parlerons ensuite plus précisément du modèle que nous avons choisi d'utiliser.
3. Enfin, nous entamerons la phase applicative et vous présenterons également les résultats que nous avons obtenus.

Avant d'introduire en détail les modèles linéaires généralisés, nous évoquerons brièvement le modèle linéaire classique, avec l'objectif de bien comprendre le cheminement vers la construction d'un GLM.

Le modèle linéaire classique

Le modèle linéaire classique consiste à établir une relation de type linéaire entre une variable à expliquer (aléatoire) et des variables explicatives (déterministes). Sur le plan conceptuel, le modèle se compose d'une partie « fixe » et d'une partie « aléatoire » qui restitue la variabilité de l'objet étudié.

L'équation du modèle se présente ainsi :

$$Y = \beta_0 + \sum_{k=1}^{n} \beta_k X_k + \epsilon \tag{6.1}$$

Avec :

— Y la variable à expliquer.
— $(\beta_k)_{k=1,\dots,n}$ les coefficients du modèle à déterminer.
— ϵ un terme d'erreur servant à quantifier l'écart entre la valeur observée et la valeur estimée.

Le modèle linéaire généralisé

L'intérêt du modèle linéaire généralisé (GLM) quant à lui est qu'il permet aux termes d'erreur d'être distribués autrement que simplement selon une loi normale. En outre, il permet aux variables explicatives d'être reliées à la variable explicative par le biais d'une fonction de lien, et en autorisant l'amplitude de la variance de chaque mesure d'être une fonction de sa valeur prédite.

Le modèle linéaire généralisé consiste en 3 éléments principaux :

— Des variables à expliquer $(Y_k)_{k=1,\dots,n}$ qui font partie de la famille exponentielle.
— Un prédicteur linéaire.
— Une fonction de lien.

Nous allons maintenant présenter sommairement chacun de ces éléments. L'objectif est de comprendre le fonctionnement du modèle avant de passer à la phase applicative et à l'interprétation des résultats obtenus.

Des variables à expliquer issues de la famille exponentielle

Dans le cadre d'application du modèle, nous supposons les variables $(Y_k)_{k=1,\dots,n}$ indépendantes et de fonctions de densité qui font partie de la famille exponentielle. Elles s'écrivent :

$$f(y, \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right) \tag{6.2}$$

Avec :

— a, b et c des fonctions définies sur R ou R^2.
— θ le paramètre de la moyenne.
— ϕ le paramètre de dispersion.

Il est intéressant de prendre aussi en compte les deux égalités suivantes :

— $E(Y) = b'(\theta)$
— $Var(Y) = b''(\theta) \times a(\phi)$

Un prédicteur linéaire

Le prédicteur linéaire noté η est une quantité représentative des informations des variables indépendantes du modèle. Il correspond à la valeur estimée des données par la fonction de lien ; c'est une combinaison linéaire (d'où son nom) des paramètres inconnus que l'on notera β. Si nous écrivons les coefficients de la combinaison linéaire dans une matrice de variables indépendantes X, nous obtenons l'expression suivante :

$$\eta = X\beta \tag{6.3}$$

Une fonction de lien

La fonction de lien permet de représenter le lien entre le prédicteur linéaire et l'espérance de la fonction de distribution. Avec $(Y_i)_{i=1,...,n}$ les variables à expliquer, en notant $\mu_i = E(Y_i)$ et g la fonction de lien du modèle, nous obtenons l'égalité suivante :

$$\eta_i = g(\mu_i) \quad \forall i = 1, ..., n \tag{6.4}$$

Après avoir introduits les premiers concepts importants relatifs à l'utilisation d'un GLM, nous passons à la phase d'application.

6.2 Application du modèle

Dans cette sous-partie, nous présenterons les résultats que obtenus en appliquant un modèle GLM sur notre base de travail. Nous structurerons notre approche de la manière suivante :

1. Dans un premier temps, nous décrirons la base de données sur laquelle nous appliquerons le modèle ; de ce que nous cherchons à prédire ; du modèle GLM spécifique que nous choisissons et expliquerons ce choix.
2. Dans un deuxième temps, nous détaillerons le processus et les retraitements effectués pour parvenir aux résultats finaux ; que nous présenterons ensuite.

6.2.1 Base de données, objectifs et choix du modèle

Dans le cadre de l'application d'un modèle GLM, il a fallu légèrement modifier notre base de données et identifier la variable à prédire.

Pour rappel, la base de travail actuelle offre une vision client par année de survenance. Jusqu'ici, la variable utilisée pour compter le nombre de sinistres était une variable de comptage représentée par un entier naturel supérieur ou égal à 0. Nous devons redéfinir maintenant l'objectif à atteindre en appliquant un modèle GLM.

Comme nous l'avons souligné précédemment, l'objectif que nous cherchons à atteindre par l'application d'un modèle GLM n'est pas un modèle parfait de prédiction de l'incidence, mais plutôt un intermédiaire d'interprétation qui permettrait dans la continuité de ce qui a été fait précédemment, de poursuivre jusqu'à son terme la phase d'identification des variables discriminantes pour l'incidence en arrêt de travail.

C'est pourquoi l'un des premiers choix forts préalables à l'application du modèle a été de prédire une variable binaire valant 0 si l'assuré n'a pas eu de sinistres et 1 si l'assuré en a eu au moins un sur la période. Les raisons de ce choix sont motivées par les explications suivantes :

— D'une part, le modèle est rendu plus simple d'application par ce changement. En effet, les modèles GLM sont mieux adaptés pour prédire des variables "top" plutôt que des variables entières pour lesquelles il y aura beaucoup de valeurs nulles (même si des solutions existent).
— D'autre part, prédire une variable "top" nous rapproche des objectifs de l'application du modèle. En effet, nous parviendrons quand même à identifier des variables et des caractéristiques qui rendent les individus plus susceptibles de subir un arrêt de travail.

Le choix de la variable à prédire effectué, il reste à sélectionner la fonction de lien du modèle GLM que nous utiliserons. Comme la variable à prédire est une variable qui prend ses valeurs dans l'ensemble $[|0,1|]$, nous opterons pour une fonction de lien *logit*. En la notant g, on a l'égalité suivante :

$$g(p) = \ln\left(\frac{p}{1-p}\right) \tag{6.5}$$

La dernière étape avant de passer à l'application du modèle consiste à retraiter les variables explicatives de la base pour que celles-ci soient correctement gérées par le modèle GLM. Les variables quantitatives ne sont pas modifiées ; les variables qualitatives sont quant à elles discrétisées en fonction du nombre de modalités qu'elles possèdent.

Nous attribuons aussi une pondération à chaque observation de la base en fonction du temps passé en portefeuille (plus un individu reste longtemps en portefeuille plus l'observation qu'on en fait à un instant t a un poids élevé).

Maintenant que nous avons clairement fixé les objectifs de cette partie, choisi les caractéristiques de notre modèle et modifié les variables de notre base de données en conséquence, nous pouvons entamer la partie d'application du modèle.

6.2.2 Processus et retraitements

Dans le cadre du traitement de sujets de modélisation, il arrive souvent que la base de données initiale soit divisée en deux bases distinctes. L'une sert pour l'apprentissage et l'autre pour le test ; en d'autres termes, nous calibrons notre modèle sur certaines données et vérifions que celui-ci est pertinent en le testant sur des nouvelles données. L'objectif en suivant cette procédure est de vérifier la significativité et le caractère prédictif du modèle.

Dans notre cas, nous choisissons de passer cette étape et de calibrer notre modèle sur l'ensemble de la base de données. Les raisons qui motivent notre choix sont les suivantes :

— L'objectif de cette partie n'étant pas tant de prédire, mais plutôt d'identifier des variables discriminantes, nous n'éprouvons pas le besoin d'utiliser des statistiques supplémentaires pour vérifier le caractère prédictif du modèle.
— Nous souhaitons identifier les variables discriminantes en utilisant le maximum de données dont nous disposons. En calibrant le modèle sur l'ensemble de la base, nous obtiendrons de meilleurs résultats.

La première étape consiste ainsi en l'application du code GLM, en respectant les modalités définies précédemment. Nous obtenons ainsi la première série de résultats suivants :

```
Call:
glm(formula = top_sin ~ ., family = binomial(link = "logit"),
    data = total, weights = vec_poids, na.action = na.pass)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5408  -0.3156  -0.2734  -0.2261   3.0898

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -4.084425   0.119052 -34.308  < 2e-16 ***
grp_age31-35                                    -0.051757   0.080372  -0.644 0.519594
grp_age36-39                                     0.004707   0.080530   0.058 0.953391
grp_age40-42                                    -0.062342   0.085965  -0.725 0.468326
grp_age43-45                                    -0.040530   0.082909  -0.489 0.624950
grp_age46-48                                     0.114906   0.080165   1.433 0.151752
grp_age49-51                                     0.227480   0.079803   2.851 0.004365 **
grp_age52-55                                     0.350923   0.076918   4.562 5.06e-06 ***
grp_age56-59                                     0.514050   0.079596   6.458 1.06e-10 ***
grp_age60-65                                     0.413841   0.088043   4.700 2.60e-06 ***
grp_anciennete1                                  0.128833   0.077718   1.658 0.097380 .
grp_anciennete2                                  0.077919   0.081043   0.961 0.336325
grp_anciennete3                                  0.043059   0.084467   0.510 0.610213
grp_anciennete4-5                                0.044647   0.078962   0.565 0.571783
grp_anciennete6-7                                0.074977   0.084158   0.891 0.372978
grp_anciennete8-9                                0.035258   0.090015   0.392 0.695292
grp_anciennete10-13                              0.113609   0.086455   1.314 0.188819
grp_anciennete14-19                              0.178535   0.091538   1.950 0.051130 .
grp_anciennete20-28                              0.258722   0.096979   2.668 0.007634 **
grp_gar630-900                                   0.187457   0.085590   2.190 0.028511 *
grp_gar900-1100                                  0.498211   0.082299   6.054 1.42e-09 ***
grp_gar1100-1260                                 0.540525   0.080489   6.716 1.87e-11 ***
grp_gar1260-1540                                 0.542628   0.080891   6.708 1.97e-11 ***
grp_gar1540-1690                                 0.651047   0.079991   8.139 3.98e-16 ***
grp_gar1690-2050                                 0.664600   0.080312   8.275  < 2e-16 ***
grp_gar2050-3850                                 0.801298   0.082348   9.731  < 2e-16 ***
grp_gar3850-12800                                0.485866   0.084596   5.743 9.28e-09 ***
grp_gar12800-136800                              0.729257   0.080766   9.029  < 2e-16 ***
sel_medO                                        -0.000141   0.041620  -0.003 0.997297
nom_regionAuvergne-Rhône-Alpes                  -0.140097   0.180046  -0.778 0.436500
nom_regionBourgogne-Franche-Comté               -0.071530   0.205481  -0.348 0.727758
nom_regionBretagne                               0.208020   0.248483   0.837 0.402502
nom_regionCentre-Val de Loire                    0.202861   0.132818   1.527 0.126670
nom_regionCorse                                  0.168659   0.255972   0.659 0.509962
nom_regionGrand Est                             -0.340010   0.277193  -1.227 0.219965
nom_regionHauts-de-France                       -0.193571   0.181762  -1.065 0.286890
nom_regionNormandie                             -0.143027   0.248455  -0.576 0.564840
nom_regionNouvelle-Aquitaine                     0.246037   0.158363   1.554 0.120272
nom_regionOccitanie                              0.049460   0.180868   0.273 0.784502
nom_regionPays de la Loire                       0.179330   0.245610   0.730 0.465303
nom_regionProvence-Alpes-Côte d'Azur             0.001474   0.195966   0.008 0.994000
nom_etabBP Alsace Lorraine Champagne             0.680281   0.275668   2.468 0.013596 *
nom_etabBP Aquitaine Centre-Atlantique           0.127696   0.169649   0.753 0.451627
nom_etabBP Auvergne-Rhône-Alpes                  0.615026   0.181872   3.382 0.000721 ***
nom_etabBP Bourgogne Franche-Comté               0.410583   0.203615   2.016 0.043750 *
nom_etabBP Grand Ouest                           0.166991   0.240884   0.693 0.488155
nom_etabBP Méditerranée                          0.355740   0.204772   1.737 0.082342 .
nom_etabBP Nord                                  0.483473   0.193153   2.503 0.012312 *
nom_etabBP Occitane                              0.351112   0.181744   1.932 0.053372 .
nom_etabBP Sud                                   0.513706   0.186360   2.757 0.005842 **
nom_etabBP Val de France                         0.065736   0.121061   0.543 0.587132
lib_cspAgriculteurs et salariés agricoles       -0.007689   0.090797  -0.085 0.932511
lib_cspArtisans                                  0.139478   0.048836   2.856 0.004290 **
lib_cspCadres et professions intellectuelles supérieures -0.186427   0.075064  -2.484 0.013007 *
lib_cspCadres moyens                            -0.059767   0.091277  -0.655 0.512603
lib_cspCommerçants                              -0.062906   0.064513  -0.975 0.329514
lib_cspEmployés                                 -0.091591   0.096234  -0.952 0.341224
lib_cspIndustriels                              -0.119043   0.049752  -2.393 0.016723 *
lib_cspOuvriers                                 -0.064973   0.090570  -0.717 0.473141
franchiseACC 7J                                 -0.601453   0.079844  -7.533 4.97e-14 ***
franchiseAcc 30J / Mal 30J                      -0.338686   0.039103  -8.661  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35883  on 111170  degrees of freedom
Residual deviance: 35276  on 111110  degrees of freedom
AIC: 35403

Number of Fisher Scoring iterations: 6
```

Figure 6.1: Premiers résultats après application du GLM

Notons que les interprétations se font par rapport à l'individu de référence suivant : un individu entre 18 et 35 ans, avec une ancienneté inférieure à 1 an, un montant garanti inférieur à 630€, sans sélection médicale, de la région Île-de-France, ayant souscris son contrat dans un établissement des Rives de Paris, de la catégorie professionnel Personnel de service, avec une franchise de type Accident 7 jours, Maladie 15 jours.

Nous pouvons maintenant interpréter et d'effectuer les retraitements nécessaires pour parvenir aux résultats finaux. Nous procédons de la manière suivante :

— Par lecture de la $p-value$, en dernière colonne des résultats affichés précédemment, nous pouvons déduire de la significativité des variables explicatives dans le modèle.
— Comme nous travaillons au seuil de confiance 95%, les variables explicatives avec une $p-value$ supérieure à 5% ne seront pas interprétées. En effet, nous ne pouvons pas conclure à propos de l'impact de celles-ci sur le comportement de la variable à expliquer.
— Deux possibilités s'offrent ensuite à nous : en fonction des résultats, nous pourrons retirer la variable du modèle ou simplement procéder à un regroupement de ses modalités.
— Les variables explicatives avec une $p-value$ inférieure à 5% sont des variables significatives pour le modèle et nous pouvons affirmer qu'elles ont un impact réel sur le comportement de la variable à expliquer.

Après application et suivi de ces principes, nous parvenons aux résultats suivants :

```
Call:
glm(formula = top_sin ~ grp_age + grp_gar + grp_etab + grp_csp +
    franchise, family = binomial(link = "logit"), data = total,
    weights = vec_poids)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5113  -0.3136  -0.2736  -0.2299   3.0976

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -4.07315    0.08623 -47.237  < 2e-16 ***
grp_age49-51                      0.24222    0.05235   4.627 3.72e-06 ***
grp_age52-55                      0.37279    0.04669   7.984 1.41e-15 ***
grp_age56-59                      0.54306    0.04876  11.138  < 2e-16 ***
grp_age60-65                      0.46029    0.05870   7.841 4.46e-15 ***
grp_gar630-900                    0.21356    0.08414   2.538 0.011142 *
grp_gar900-1100                   0.49790    0.08152   6.108 1.01e-09 ***
grp_gar1100-1540                  0.54778    0.07205   7.603 2.90e-14 ***
grp_gar1540-2050                  0.65595    0.07155   9.168  < 2e-16 ***
grp_gar2050-3850                  0.83808    0.07815  10.723  < 2e-16 ***
grp_gar3850-12800                 0.48932    0.08310   5.888 3.90e-09 ***
grp_gar12800-136800               0.71707    0.07930   9.043  < 2e-16 ***
grp_etabBP Val de France & Nord   0.25211    0.07546   3.341 0.000835 ***
grp_etabBP Aquitaine & Méditerranée  0.39809  0.06794   5.860 4.64e-09 ***
grp_etabBP Alsace, Bourgogne & Ouest  0.35367  0.06197  5.708 1.15e-08 ***
grp_etabBP Occitane               0.46117    0.08097   5.696 1.23e-08 ***
grp_etabBP Auvergne & Sud         0.53795    0.06210   8.662  < 2e-16 ***
grp_cspArtisans                   0.18549    0.03948   4.698 2.62e-06 ***
grp_cspCadres sup                -0.13906    0.06966  -1.996 0.045889 *
grp_cspIndustriels               -0.09461    0.04373  -2.164 0.030478 *
franchiseACC 7J                  -0.62145    0.07892  -7.874 3.43e-15 ***
franchiseAcc 30J / Mal 30J       -0.29721    0.03476  -8.550  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35883  on 111170  degrees of freedom
Residual deviance: 35319  on 111149  degrees of freedom
AIC: 35371

Number of Fisher Scoring iterations: 6
```

Figure 6.2: Résultats retraités du GLM

Voici ce que nous pouvons déduire des résultats obtenus :

— Les individus d'âge inférieur à 49 ans sont difficilement identifiables par rapport à leur sinistralité ; néanmoins, celle-ci augmente significativement pour les âges supérieurs.
— Le montant garanti semble bien être corrélé positivement avec l'incidence ; plus celui-ci est élevé, plus la sinistralité est forte.

— Par rapport à la BP Rives de Paris, les établissements ou groupes d'établissements bancaires les plus risqués sont, dans l'ordre : BP Val de France, BP Alsace Bourgogne Ouest, BP Aquitaine Méditerranée, BP Occitane, BP Auvergne Sud.
— Par rapport aux autres catégories socio-professionnelles, les Industriels et les Cadres supérieurs semblent peu risqués. À l'inverse, les Artisans semblent quant à eux soumis à un risque très élevé.
— Les franchises les plus risquées sont dans l'ordre : Accident 7 jours, Accident 30 jours Maladie 30 jours, Accident 7 jours Maladie 30 jours.

Il est intéressant de remarquer que ces résultats sont en accord avec l'ensemble de ceux obtenus précédemment (analyse descriptive et méthodes d'apprentissage non-supervisé).

Nous observons toutefois que la variable *Région* apparaît non-significative. Son homologue, la variable *Établissement* apparaît quant à elle significative.

6.2.3 Conclusion

Par application des modèles linéaires généralisés, nous avons obtenu des résultats cohérents en partie avec ceux présentés dans les parties précédentes. La disparité majeure que nous observons concerne la variable *Région*, puisqu'elle apparaît ici non-significative et donc non-discriminante.

La variable *Établissement* est retenue, nous conservons donc le principe d'influence de la localisation géographique de l'assuré sur l'incidence en incapacité.

Au cours de la phase de modélisation à venir, nous tenterons dans un premier temps de créer des groupes d'individus plus ou moins risqués. Pour ce faire, nous nous baserons sur les caractéristiques de risque que nous avons identifiées jusque-là.

Néanmoins, nous devons aussi respecter des critères d'exposition : pour que les étapes de modélisation aient un sens, il est nécessaire de bénéficier d'une exposition suffisante. À la vue des derniers résultats obtenus par GLM, nous prenons les décisions suivantes :

— Les variables *Franchise* et *Montant de garantie* seront mises de côté pour la modélisation et ré-interviendront plutôt au moment de la tarification.
— Certaines modalités des variables présentant des profils de risque similaires seront regroupées.

Après regroupement, nous obtenons les résultats suivants :

```
Call:
glm(formula = top_sin ~ grp_age + grp_gar + grp_etab + grp_csp +
    franchise, family = binomial(link = "logit"), data = total,
    weights = vec_poids)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4851  -0.3129  -0.2728  -0.2320   3.0686

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -3.59100    0.04864 -73.827  < 2e-16 ***
grp_age49-51                     0.25136    0.05227   4.809 1.52e-06 ***
grp_age52-55                     0.38725    0.04645   8.337  < 2e-16 ***
grp_age56-59                     0.56138    0.04830  11.622  < 2e-16 ***
grp_age60-65                     0.48579    0.05799   8.377  < 2e-16 ***
grp_gar900-1540 & 3850-12800     0.40535    0.04851   8.356  < 2e-16 ***
grp_gar1540-3850 & 12800-136800  0.59845    0.04778  12.526  < 2e-16 ***
grp_etabBP Rives de Paris       -0.38027    0.05896  -6.449 1.12e-10 ***
grp_etabBP Val de France & Nord -0.13619    0.05609  -2.428   0.0152 *
grp_etabBP Auvergne & Sud        0.15339    0.03552   4.318 1.58e-05 ***
grp_cspArtisans                  0.19729    0.03917   5.037 4.73e-07 ***
grp_cspCadres sup & Industriels -0.09965    0.03954  -2.520   0.0117 *
franchiseACC 7J                 -0.62815    0.07879  -7.973 1.55e-15 ***
franchiseAcc 30J / Mal 30J      -0.27989    0.03411  -8.205 2.31e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35883  on 111170  degrees of freedom
Residual deviance: 35340  on 111157  degrees of freedom
AIC: 35377

Number of Fisher Scoring iterations: 6
```

Figure 6.3: Résultats finaux du GLM

Les groupes présentés ci-dessus correspondent bien aux groupes mis en évidence dans :

1. La partie d'analyse descriptive.
2. La partie de réduction des modalités des variables qualitatives.
3. La partie d'identification de groupes d'assurés semblables.

Nous abordons donc la modélisation sur la base de ces observations, et détaillons à présent la suite des des étapes.

Deuxième partie

Modélisation

7 Introduction

7.1 Définition des objectifs

Étant parvenus à l'identification de variables discriminantes pour l'incidence en arrêt de travail, il s'agit maintenant de passer à la phase de modélisation. Nous procéderons en quatre temps :

Dans un premier temps, nous aurons pour objectif de créer trois groupes de risque (individus peu risqués, individus neutres, individus risqués) d'exposition équivalente et en se basant sur les variables précédemment identifiées.

Ensuite, nous procéderons grâce à l'estimateur de Kaplan-Meier à la modélisation de la loi d'incidence pour le groupe d'individus de risque neutre.

Nous nous baserons sur cette loi pour positionner par application du modèle de Cox les lois d'incidence des deux autres groupes de risque.

Enfin, nous adapterons ces lois pour qu'elles correspondent à la forme classique utilisée dans les outils de tarification du service. À ce moment-là, nous serons capables de comparer nos résultats avec ceux issus des lois déjà existantes.

Avant de passer à la création des groupes de risque, il s'agit de sélectionner parmi les variables discriminantes identifiées précédemment celles que nous conserverons pour notre étude. Il est en effet impossible de toutes les conserver, puisque nous avons besoin d'un certain niveau d'exposition minimal afin que l'application des modèles ait un sens.

En outre, afin de respecter le cadre défini par les outils de tarification que nous utiliserons dans la dernière partie du mémoire, nous préférerons créer une loi d'incidence par âge et par franchise. Nous conserverons ainsi la variable *Âge* comme variable discriminante. Pour traiter de la variable *Franchise* en revanche, nous procéderons dans un premier temps à la modélisation pour une durée de franchise particulière et ensuite à l'extrapolation à toutes les durées et tous les types de franchise [17].

17. Nous détaillerons ces étapes dans une partie ultérieure du mémoire.

7.2 Création des groupes de risque

En plus des variables citées précédemment, nous avons choisi, dans le cadre de notre étude, de retenir les variables *Établissement* et *Catégorie socio-professionnelle*.

Nous avons vu auparavant qu'il semblerait que les individus ayant souscrit un contrat d'assurance proposé par les entités bancaires du nord de la France présentent une incidence plus faible. De la même manière, nous avons observé une disparité des risques selon les différentes catégories socio-professionnelles, avec les *Artisans* apparaissant comme catégorie la plus risquée. Ces variables semblent donc adéquates pour la création de groupes d'individus plus ou moins risqués.

Pour conserver une exposition raisonnable au sein des groupes d'individus que nous allons étudier, nous conserverons les variables *Établissement* et *CSP* avec leurs modalités regroupées de la même manière qu'à la sortie du modèle GLM [18]. C'est à dire que :

— Les modalités de la variable *Établissement* seront les suivantes : BP Rives de Paris, BP Val de France et Nord et Autres BP.
— Les modalités de la variable *Catégorie socio-professionnelle* seront les suivantes : Artisans, Cadres sup Industriels et Autres CSP.

Avant de créer des groupes de risque d'exposition environ égale, nous allons étudier les effets croisés des variables citées précédemment. En effet, bien que :

— Les individus plus âgés semblent plus risqués.
— Les artisans semblent plus risqués.
— Les individus ayant souscrit à un contrat dans une entité bancaire du nord de la France semblent plus risqués.

Nous devons vérifier si les artisans âgés ayant souscrit à un contrat dans une entité bancaire du nord de la France sont des individus plus risqués que la moyenne. C'est pourquoi, après avoir procédé à un regroupement des individus par tranches d'âge d'exposition égale, nous calculons les taux d'incidence en fonction des variables sélectionnées.

18. cf. page 69.

Nous regroupons les résultats dans un tableau :

grp_etab	grp_csp	tx_18-48	expo_18-48	tx_49-51	expo_49-51	tx_52-55	expo_52-55	tx_56-59	expo_56-59	tx_60-65	expo_60-65
BP Rives de Paris	Autres CSP	2,73	2493	2,05	438	3,28	579	4,51	554	4,16	456
BP Rives de Paris	Artisans	2,62	839	2,76	289	3,21	405	3,53	368	4,43	338
BP Val de France & Nord	Cadres sup & Industriels	3,43	1719	2,34	299	3,95	278	3,96	277	2,76	217
BP Rives de Paris	Cadres sup & Industriels	2,95	2673	6,01	482	3,7	594	4,91	447	4,28	373
BP Val de France & Nord	Autres CSP	3,51	4047	3,54	480	5,24	629	6,87	480	5,23	344
Autres BP	Cadres sup & Industriels	3,73	7390	5,62	1442	4,7	1723	5,7	1368	4,66	1029
Autres BP	Autres CSP	3,95	18395	5,01	2695	5,82	3109	5,3	2586	5,32	1878
BP Auvergne & Sud	Cadres sup & Industriels	4,64	3059	4,21	616	5,34	861	7,67	652	3,81	446
BP Auvergne & Sud	Autres CSP	4,59	11050	5,07	1399	6,66	1531	6,75	1214	5,93	843
BP Val de France & Nord	Artisans	3,46	692	7,61	183	5,37	242	10,3	271	9,11	164
Autres BP	Artisans	4,82	4400	5,78	1332	6,36	1981	8,87	1894	7,1	1365
BP Auvergne & Sud	Artisans	5,08	2499	7,77	746	8,25	1030	8,96	915	7,75	645

Figure 7.1: Taux d'incidence après croisement des modalités des variables explicatives

Nous pouvons ensuite calculer un taux moyen par ligne, trier par ordre croissant et identifier des groupes d'exposition environ égale :

grp_etab	grp_csp	tx_18-48	expo_18-48	tx_49-51	expo_49-51	tx_52-55	expo_52-55	tx_56-59	expo_56-59	tx_60-65	expo_60-65	taux moyen	expo tot	expo groupe	Groupe de risque
BP Rives de Paris	Autres CSP	2,73	2493	2,05	438	3,28	579	4,51	554	4,16	456	3,10	4520		
BP Rives de Paris	Artisans	2,62	839	2,76	289	3,21	405	3,53	368	4,43	338	3,17	2239		
BP Val de France & Nord	Cadres sup & Industriels	3,43	1719	2,34	299	3,95	278	3,96	277	2,76	217	3,37	2790	33050	A
BP Rives de Paris	Cadres sup & Industriels	2,95	2673	6,01	482	3,7	594	4,91	447	4,28	373	3,67	4569		
BP Val de France & Nord	Autres CSP	3,51	4047	3,54	480	5,24	629	6,87	480	5,23	344	4,06	5980		
Autres BP	Cadres sup & Industriels	3,73	7390	5,62	1442	4,7	1723	5,7	1368	4,66	1029	4,35	12952		
Autres BP	Autres CSP	3,95	18395	5,01	2695	5,82	3109	5,3	2586	5,32	1878	4,46	28663	34297	B
BP Auvergne & Sud	Cadres sup & Industriels	4,64	3059	4,21	616	5,34	861	7,67	652	3,81	446	4,98	5634		
BP Auvergne & Sud	Autres CSP	4,59	11050	5,07	1399	6,66	1531	6,75	1214	5,93	843	5,06	16037		
BP Val de France & Nord	Artisans	3,46	692	7,61	183	5,37	242	10,3	271	9,11	164	6,04	1552	34396	C
Autres BP	Artisans	4,82	4400	5,78	1332	6,36	1981	8,87	1894	7,1	1365	6,20	10972		
BP Auvergne & Sud	Artisans	5,08	2499	7,77	746	8,25	1030	8,96	915	7,75	645	6,89	5835		

Figure 7.2: Taux d'incidence après croisement des modalités des variables explicatives (2)

Pour faciliter la lecture des données numériques du tableau, nous avons surligné :

- En vert foncé les taux d'incidence très faibles par rapport à la moyenne.
- En vert clair les taux d'incidence faibles par rapport à la moyenne.
- En blanc les taux d'incidence dans la tranche moyenne.
- En rouge clair les taux d'incidence élevés par rapport à la moyenne.
- En rouge foncé les taux d'incidence très élevés par rapport à la moyenne.

Nous pouvons ensuite effectuer plusieurs observations, qui viennent confirmer nos hypothèses précédentes et répondre à certaines de nos interrogations :

Dans un premier temps, nous constatons que l'incidence semble bien augmenter avec l'âge de la population assurée.

Ensuite, par lecture des couleurs nous constatons que : les individus jeunes et appartenant à des groupes peu risqués (en haut à gauche du tableau) présentent des taux d'incidence très faibles par rapport à la moyenne ; les individus âgés et appartenant à des groupes risqués (en bas à droite du tableau) présentent des taux d'incidence très élevés par rapport à la moyenne.

Nous observons aussi que les effets croisés des variables ne présentent pas de schémas complexes : une combinaison de modalités risquées donne bien naissance à un individu risqué ; une combinaison de modalités non-risquées donne bien naissance à un individu peu risqué.

Les groupes de risques mis en évidence sont d'exposition environ égale (34 000 individus). Il est intéressant de remarquer que la population du groupe B (groupe de risque neutre) est constitué en majeure partie d'individus aux modalités : *Autres BP* et *Autres CSP*. Cela signifie que les distinctions de modalités retenues semblent bien pertinentes puisqu'elles permettent de mettre en évidence des disparités de sinistralité au sein du portefeuille.

Maintenant que nous avons identifié trois groupes de risque pertinents, nous pouvons entamer la partie de modélisation de la loi d'incidence. Comme expliqué précédemment, nous nous concentrerons sur la modélisation de la loi pour le groupe de risque neutre (groupe B) et positionnerons ensuite les groupes A et C (de risque faible et élevé) par rapport à celui-ci par application du modèle de Cox.

8 Modélisation pour le groupe de risque neutre

Dans cette partie, nous introduirons les différents modèles utilisés pour la modélisation de la loi d'incidence du groupe B (groupe de risque neutre) et présenterons les résultats obtenus après application. Notre approche s'effectuera de la manière suivante :

Dans un premier temps, nous transformerons notre base de travail pour que celle-ci se prête à l'application des modèles de survie. Nous découperons également notre base de travail en une base d'entraînement et une base de test sur laquelle nous validerons notre modèle.

Dans un second temps, nous calculerons les taux d'incidence bruts par âge par application du modèle de Kaplan-Meier.

Dans un troisième temps, nous effectuerons un lissage des taux en utilisant la méthode de Whittaker-Henderson.

Dans un quatrième temps, nous validerons les taux finaux obtenus en calculant des ratios observés / attendus sur la base test isolée en étape 1.

Enfin, nous ré-effectuerons un apprentissage du modèle ; sur la base complète cette fois-ci, pour bénéficier de la quantité maximum d'information disponible et s'assurer de coïncider au mieux avec la réalité.

8.1 Calcul des taux d'incidence bruts

Avant de passer au calcul des taux d'incidence bruts, nous introduisons les notions de modèles de survie et présentons l'estimateur de Kaplan-Meier que nous utiliserons dans cette partie.

8.1.1 Présentation de l'estimateur de Kaplan-Meier

En notant T la variable aléatoire positive qui décrit le temps qui s'écoule entre la souscription du contrat par l'assuré et l'entrée en incapacité, et en notant x l'âge de ce-dernier, nous obtenons l'égalité suivante :

$$ {}_tq_x = P(T + t | T > x) \tag{8.1}$$

avec ${}_tq_x$ la probabilité que l'assuré d'âge x ait un arrêt de travail avant d'atteindre l'âge $x + t$. Dans le cadre de notre étude, nous cherchons à estimer le taux d'incidence à un an à l'âge x que nous noterons q_x. Nous allons maintenant introduire quelques notions basiques d'analyse de survie qui nous permettrons de calculer q_x.

En notant F la fonction de répartition de la variable aléatoire T, les fonctions utilisées dans les modèles de survie sont les suivantes :

— La fonction de survie S définie par : $S(t) = 1 - F(t) = P(T > t)$ qui est une fonction décroissante continue. Nous avons les égalités suivantes : $S(0) = 1$, $\lim_{t\to+\infty} S(t) = 0$ et $f(t) = \frac{dF(t)}{dt} = \frac{-dS(t)}{dt}$ avec f la fonction de densité de T.

— La fonction de hasard h définie par : $h(t) = \frac{f(t)}{S(t)}$. Elle correspond au taux d'incidence « instantané » et nous avons les égalités suivantes :

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T \leq t + \Delta t | T > t}{\Delta t} = \lim_{\Delta t\to 0} \frac{P(t \leq T \leq t + \Delta t)}{S(t)\Delta t} \tag{8.2}$$

— La fonction de hasard cumulé H définie par : $H(t) = \int_0^t h(x)dx$.

De toutes ces égalités, nous déduisons que :

$$S(t) = \exp(-H(t)) \tag{8.3}$$

Toutes ces fonctions caractérisent ainsi entièrement la distribution de T. Dans notre cas précis, pour calculer les taux bruts d'incidence en arrêt de travail, nous utiliserons l'estimateur de Kaplan-Meier (1958) [7]. Celui-ci nous permet d'écrire la fonction de survie estimée $\hat{S}(t)$ sous la forme suivante :

$$\hat{S}(t) = \prod_{t_i<t} (1 - \frac{d_i}{n_i}) \tag{8.4}$$

Avec :

— d_i le nombre d'arrêts de travail observé au temps t_i.
— n_i le nombre d'individus assurés à risque au temps t_i.

Ainsi, nous pouvons déduire l'expression de q_x pour tout âge x :

$$\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)} \tag{8.5}$$

Une fois calculés les différents taux d'incidence q_x par âge, il est intéressant de construire des intervalles de confiance pour augmenter la pertinence de nos résultats finaux. Pour ce faire, nous utiliserons l'estimateur de Hoem qui propose un calcul des intervalles de confiance simple et cohérent.

L'estimateur de Hoem propose l'écriture suivante : $\hat{q}_x(t) = \frac{N(x,t)}{E(x,t)}$ où :

— $E(x,t)$ correspond à l'exposition à l'âge x et à l'instant t.
— $N(x,t)$ correspond au nombre de sorties à l'âge x et à l'instant t.

En supposant que les taux d'incidence en arrêt de travail sont constants au cours du temps (nous reparlons de cette hypothèse au paragraphe suivant), nous avons, pour tout âge x, la relation suivante :

$$\hat{q}_x = \frac{N(x)}{E(x)} \tag{8.6}$$

En utilisant ensuite l'estimateur des taux bruts de Hoem, nous pouvons obtenir l'écriture de l'intervalle de confiance à 95% suivante :

$$IC^{95\%}(\hat{q}_x) = \hat{q}_x \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{E(x)}} \tag{8.7}$$

où $t_{1-\frac{\alpha}{2}}$ correspond au quantile d'ordre $\frac{\alpha}{2}$ de la loi normale centrée réduite et α au niveau de risque 5%.

Comme indiqué plus haut, il est important de vérifier la stabilité des taux d'incidence par année. Si les taux ne sont pas stables, la pertinence de l'utilisation des modèles de survie est remise en cause. Heureusement, dans le cadre notre étude, la distribution des taux d'incidence en arrêt de travail présente bien une vraie stabilité sur toutes les années qui serviront de base d'apprentissage au modèle. Nous pouvons ainsi continuer la phase de modélisation.

8.1.2 Création de la base de travail

Pour la modélisation d'une loi d'incidence en arrêt de travail, nous avons pris la décision lors de l'étude d'utiliser l'estimateur des Kaplan-Meier que nous venons de vous présenter. Néanmoins, pour qu'il soit cohérent d'appliquer un modèle de survie à proprement parler, il est nécessaire de passer par une étape de transformation de notre base de travail.

En effet, si nous consid"rons un assuré qui souscrirait son contrat à une certaine date t, ce que nous essayons de modéliser est le temps entre cette date de souscription et la date de survenance de son arrêt de travail s'il existe.

Là où l'utilisation d'un modèle de survie pose problème, c'est que l'individu assuré en question, à la fin de son arrêt de travail, reprendra dans la plupart des cas normalement le travail et devra donc réintégrer notre base de modélisation. Nous modéliserons en fait les différentes périodes d'activité successives des assurés, qui seront considérées mathématiquement comme leurs périodes de survie.

D'un point de vue informatique, un individu ayant souscrit à un contrat et ayant eu un unique arrêt de travail observé, sera considéré comme deux individus distincts :

— Le premier étant rentré en portefeuille à la date d'entrée de l'individu original, et étant sorti du portefeuille à la date de début d'arrêt de travail.
— Le deuxième étant rentré en portefeuille à la date de reprise du travail de l'individu original, et étant sorti à la fin de la fenêtre d'observation.

Nous pouvons synthétiser cette explication dans un schéma :

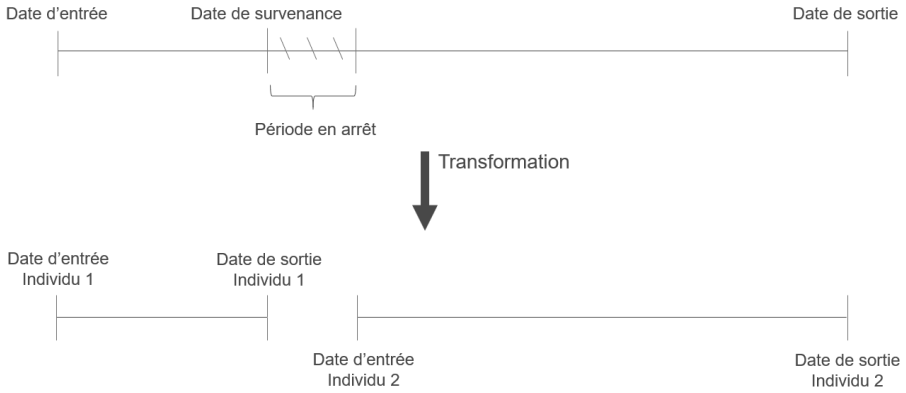

Figure 8.1: Transformation de la base en vue de l'application des modèles de survie

N'ayant néanmoins pas accès aux dates exactes de reprise des individus dans notre base de données, nous avons écrit un court algorithme de génération de durée aléatoire d'arrêt de travail afin de simuler une date de reprise hypothétique. Cet algorithme s'appuie sur les tables de maintien en incapacité [19] mises à disposition par le BCAC [20]. Ces tables de maintien proposent une approche assez circonspecte qui ajoute de la prudence à notre modélisation des taux.

Ceci n'est pas source de difficulté puisque nous restons dans une optique de tarification et non de meilleure estimation ; au contraire, cette marge de prudence ajoutée nous permet de nous assurer de ne pas produire un tarif trop inférieur aux lois réelles du marché.

Une fois la transformation de la base faite, nous pouvons passer à l'application de la méthode de Kaplan-Meier et à l'interprétation des résultats obtenus.

19. Nous aurons l'occasion de présenter ces tables dans une partie ultérieure du mémoire.
20. Bureau Commun des Assurances Collectives

8.1.3 Application

L'exposition du portefeuille étant centrée autour des âges 39 - 45 ans, et plutôt faible aux extrémités (individus de moins de 25 ans et de plus de 55 ans), nous prenons la décision de ne pas obtenir des taux estimés par Kaplan-Meier par âge, mais plutôt par tranche d'âge. Nous créons ainsi des tranches d'âge d'exposition environ égale et cherchons ensuite à estimer le taux d'incidence q_x présenté précédemment.

Il est également important de noter que nous nous concentrerons pour cette première étape de modélisation sur la franchise de type : *Accident 7 jours, Maladie 15 jours*. Comme expliqué précédemment, nous généraliserons dans une prochaine étape la loi d'incidence à tous les types de franchise. Voici les taux d'incidence bruts obtenus :

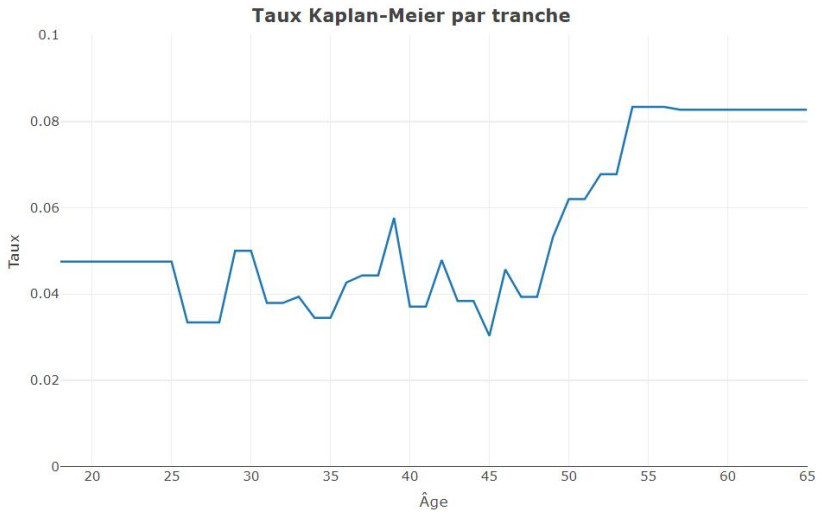

Figure 8.2: Taux d'incidence bruts par tranches d'âge

Nous observons une courbe de taux bruts qui semble logique :

- Les individus jeunes semblent relativement risqués : en effet, ceux-là sont plus susceptibles d'avoir des accidents que le reste de la population car moins prudents.
- Les individus âgés entre 25 et 47 ans présentent un profil de risque moyen. Bien que les taux soient fluctuants ici, nous corrigerons cela au moment du lissage.
- Pour les individus de plus de 47 ans, le taux d'incidence s'envole : ceux-là sont plus risqués car leur santé devient plus fragile.

8.2 Lissage des taux bruts

Il s'agit maintenant de passer à l'étape de lissage des taux bruts obtenus précédemment. Pour ce faire, nous utiliserons la méthode de Whittaker-Henderson que nous vous présenterons dans un premier temps ; puis que nous appliquerons dans un deuxième temps.

8.2.1 Présentation de la méthode de Whittaker-Henderson

L'intérêt du lissage des taux d'incidence réside dans le fait que les irrégularités présentées parmi les taux bruts obtenus précédemment ne sont pas forcément représentatives de la réalité mais sont plutôt dûes à des fluctuations d'échantillonnage au sein même de nos données. En effectuant un lissage de ces taux bruts, nous augmentons nos chances de représenter correctement l'incidence réelle.

Il existe deux principaux types de lissage :

— Les lissages paramétriques ; ici nous supposons que les taux observés sont distribués selon une loi mathématique connue.
— Les lissages non-paramétriques ; ici nous appliquons un certain nombre de traitements aux taux d'incidence bruts afin de parvenir à un résultat plus cohérent.

Dans le cadre de notre étude, nous avons choisi d'utiliser la méthode de Whittaker-Henderson (2020) [10] qui fait partie des méthodes de lissage non-paramétriques. Il existe bien évidemment d'autres méthodes de lissages non-paramétriques ; nous avons néanmoins jugé qu'il n'était pas nécessaire d'en parler dans ce mémoire.

Dans le cadre d'application de la méthode, nous prenons en compte deux critères importants :

— Un critère de fidélité ; celui-ci permet de mesurer la fiabilité de la représentation des données brutes par les données lissées.
— Un critère de régularité ; celui-ci permet de dénoter de l'aspect régulier ou non du lissage effectué.

Le principe de la méthode de Whittaker-Henderson est de rechercher des valeurs ajustées qui maximisent la fidélité et la régularité du lissage ; et donc qui minimisent les critères mathématiques de fidélité et de régularité.

Pour tout âge x, le critère de fidélité s'écrit :

$$F(\bar{q}) = \sum_{x=18}^{65} w_x (q_x - \hat{q}_x)^2 \tag{8.8}$$

où w_x est le poids attribué à chaque âge et où $\bar{q}$ représente les taux d'incidence lissés par âge. Le critère de régularité quant à lui s'écrit de la manière suivante :

$$S_k(\bar{q}) = \sum_{x=18}^{65-k} (\Delta^k q_x)^2 \tag{8.9}$$

où k est un paramètre intrinsèque au modèle permettant de faire varier la régularité du lissage et avec $\Delta^k q_x$ étant défini par l'égalité suivante :

$$\Delta^k q_x = \sum_{i=0}^{k} C_k^i (-1)^{k-i} q_{x+i} \tag{8.10}$$

Finalement, le but du lissage est de trouver la loi d'incidence $\bar{q}$ telle que la quantité $F(\bar{q}) + \alpha S_k(\bar{q})$ soit minimale, avec α un coefficient fixé par l'utilisateur permettant de jouer sur le poids de fidélité et de lissage.

8.2.2 Application

Dans le cadre de notre étude, nous initions le lissage avec le paramètre w_x comme étant la quantité $\frac{E_x}{E_{tot}}$, où :

— E_x représente l'exposition à l'âge x.
— E_{tot} représente l'exposition totale en portefeuille.

De cette manière, l'importance accordée dans le lissage à chaque âge est proportionnelle à la part de l'effectif total qu'il représente. Ensuite, nous étudions les différents résultats obtenus en faisant varier les autres paramètres mis à notre disposition par la méthode et choisissons ceux qui présentent des critères de fidélité et de régularité optimaux.

Finalement, nous obtenons les taux d'incidence en arrêt de travail lissés suivants :

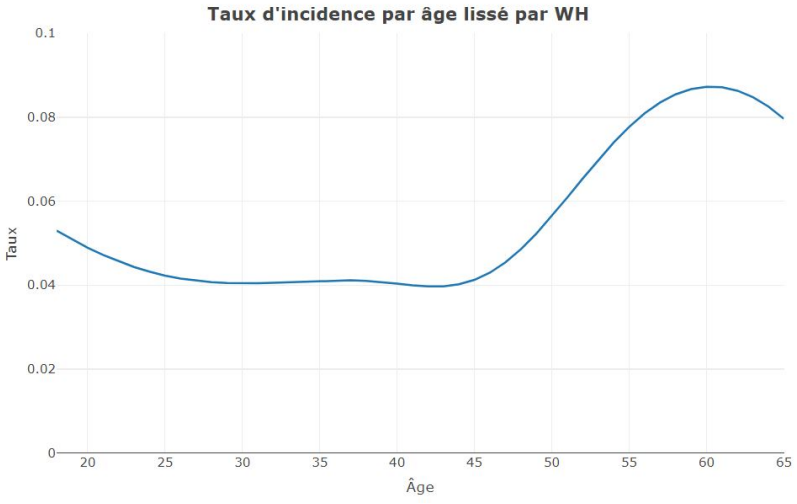

Figure 8.3: Taux d'incidence lissés par âge

Nous observons que les taux lissés suivent bien les tendances mises en évidence par les taux d'incidence bruts : nous retrouvons le pic pour les jeunes individus et la forte augmentation à partir de 50 ans.

8.3 Validation de la loi

Maintenant que nous avons obtenu les taux d'incidence lissés par âge, il s'agit de vérifier que ceux-ci sont cohérents en effectuant des prédictions sur la base test isolée en début de modélisation. Le procédé est le suivant :

1. Nous comptons empiriquement le nombre d'arrêts de travail survenus par âge dans la base test. C'est possible puisque c'est une base constituée de données connues.
2. Nous effectuons des prédictions sur les nombres d'arrêts de travail survenus par âge dans la base test en utilisant la loi modélisée.
3. Nous comparons les résultats obtenus par prédiction avec ceux observés empiriquement. En fonction de la justesse des écarts, nous validons ensuite la loi modélisée ou non.

Dans notre base d'apprentissage, nous retrouvons plusieurs années d'historique. Les individus présents dans la base d'apprentissage ne sont pas présents dans la base de test, même vus à une date ultérieure.

Si c'était le cas, cela entraînerait un biais : les individus présents dans la base de test présenteraient des caractéristiques trop proches de ceux présents dans la base d'entraînement. La prédiction serait ainsi faussée.

8.3.1 Calcul des observés sur attendus

Nous présentons ici les résultats que nous obtenons relatifs à la validation de la loi d'incidence lissée par âge. Voici le graphique que nous obtenons :

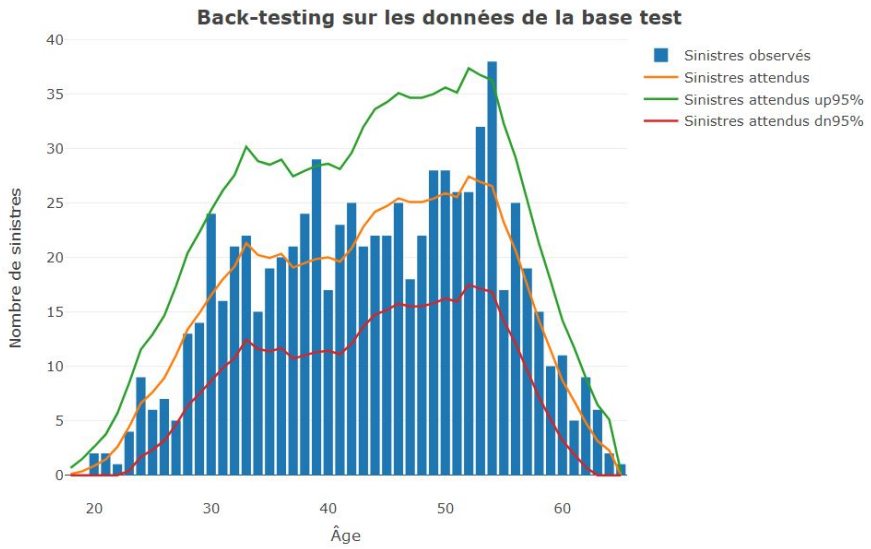

Figure 8.4: Observés / Attendus : validation de la loi pour le groupe B

Sont représentés :

— En bleu, le nombre de sinistres réellement observés par âge dans la base de test.
— En orange, le nombre de sinistres prédits par âge par la loi d'incidence.
— En rouge, le nombre de sinistres minimums à 95% prédits par la loi d'incidence.
— En vert, le nombre de sinistres maximums à 95% prédits par la loi d'incidence.

Par observation du graphique, nous constatons que le nombre de sinistres réellement observés est bien contenu dans l'intervalle créé par les nombres maximums et minimums de sinistres prédits par la loi, et ce à tout âge.

Ainsi, nous pouvons valider notre première modélisation des taux d'incidences par âge en arrêt de travail.

8.3.2 Ré-apprentissage sur la base de données complète

Maintenant que nous avons pu valider la loi créée à partir de la base d'apprentissage sur la base de test, il est dans notre intérêt de ré-effectuer une modélisation sur la base de données complètes.

En effet, ces étapes de modélisation et de validation nous ont permis de confirmer la pertinence de notre méthode ; en se limitant à un apprentissage sur la base d'entraînement, nous ne sélectionnons pas toutes les informations à notre disposition pour parfaire notre modèle.

En ré-effectuant toutes les étapes de modélisation (calcul des taux bruts par utilisation de l'estimateur de Kaplan-Meier, lissage par utilisation de la méthode de Whittaker-Henderson) cette fois-ci sur la base de données complètes, nous maximisons nos chances de capter au mieux tous les phénomènes réels liés à l'incidence en arrêt de travail.

Après cette nouvelle étape de modélisation, nous obtenons les taux d'incidence modélisés suivants :

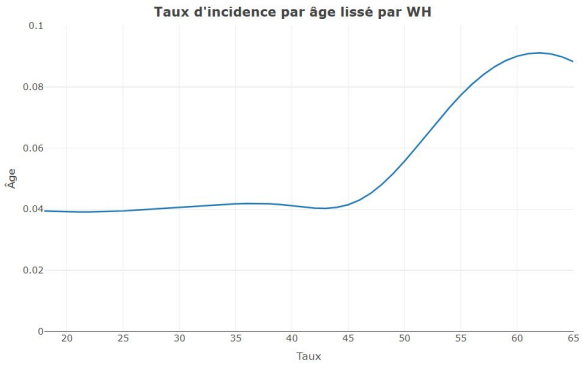

Figure 8.5: Taux d'incidence (base complète) lissés par âge pour le groupe B

9 Positionnement des taux d'incidence

Dans la partie précédente, nous sommes parvenus à obtenir des taux d'incidence lissés par âge créés à partir d'un apprentissage sur la base de données complète, pour la franchise *Accident 7 jours, Maladie 15 jours* et pour le groupe de risque neutre (*groupe B*). L'objectif ici est d'étendre cette modélisation aux deux autres groupes de risques :

— Le groupe d'individus peu risqués (*groupe A*).
— Le groupe d'individus risqués (*groupe C*).

Pour ce faire, nous utiliserons le modèle de Cox à risques proportionnels que nous vous présenterons dans un premier temps ; puis que nous appliquerons ensuite à notre loi d'incidence modélisée. Enfin, nous validerons les lois de tous les groupes de risques par calcul d'observés sur attendus.

9.1 Présentation du modèle de Cox à risques proportionnels

Dans le cadre de notre étude, il est intéressant d'utiliser le modèle de Cox (1975) [3] pour obtenir les taux d'incidence de tous les groupes de risque. En effet, nous aurions pu modéliser, en suivant l'exacte même méthode que précédemment, toutes les lois d'incidence pour tous les groupes de risque en créant à chaque fois de nouvelles bases d'apprentissage et de test.

Toutefois, il ne faut pas oublier que ces lois d'incidence seront utilisées dans la dernière partie du mémoire à des fins de tarification. Si nous disposons de trois lois différentes (une pour chaque groupe de risque), nous serons bien en mesure de calculer trois tarifs différents. La procédure en sera néanmoins fastidieuse.

Appliquer le modèle de Cox pour positionner les groupes de risques les uns par rapport aux autres nous permet ainsi d'obtenir des coefficients de passage (ou de proportionnalité) qui sont utilisables dans les formules de tarification. Ainsi, nous nous limiterons au calcul d'un seul et unique tarif auquel, en fonction des caractéristiques des individus, nous appliquerons un coefficient de majoration ou de réduction.

Les caractéristiques des loi d'incidence de chaque groupe étant similaires, il est cohérent d'essayer d'appliquer le modèle de Cox. En effet, le modèle de Cox est dit semi-paramétrique et se base sur l'hypothèse des risques proportionnels. Cela signifie que le comportement entre deux variables explicatives est supposé stable dans le temps. C'est une hypothèse forte que nous vérifierons lors de l'application du modèle.

Le modèle de Cox à risques proportionnels s'écrit de la manière suivante :

$$h(t|X_i) = h_0(t) \exp(\beta X_i) \tag{9.1}$$

où $h_0(t)$ représente le risque de référence. C'est à dire le taux d'incidence lissé par âge du groupe de risque neutre (groupe B), dans notre cas. La variable β représente quant à elle un vecteur de coefficients de régression calculés lors de l'application du modèle, et X_i les autres groupes de risques (groupe d'individus peu risqués et groupe d'individus risqués).

Nous parlons dans le modèle de risque proportionnel car le ratio des taux d'incidence entre deux groupes de risques différents est constant dans le temps :

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{\exp(\beta X_1)}{\exp(\beta X_2)} = \exp(\beta(X_1 - X2)) \tag{9.2}$$

Pour vérifier empiriquement cette hypothèse, nous pouvons étudier la log-linéarité des variables entre elles. Maintenant que le modèle de Cox a été introduit, nous passons à la phase d'application.

9.2 Application

Dans un premier temps, nous vérifions l'hypothèse de log-linéarité pour pouvoir valider la pertinence de l'utilisation du modèle de Cox. Celle-ci suppose que le coefficient β introduit précédemment évolue linéairement lorsque la variable associée augmente de une unité. Pour effectuer le test, nous traçons pour les deux groupes d'individus A et C :

$$log(h(t|X_i)) = log(h_0(t)) = \beta X_i \tag{9.3}$$

Pour les individus du groupe de risque A par rapport au groupe neutre B, nous obtenons le graphique suivant :

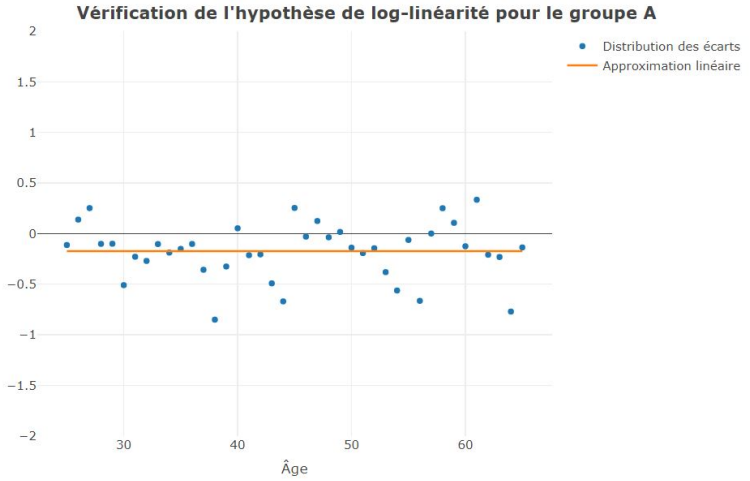

Figure 9.1: Vérification de l'hypothèse de log-linéarité pour le groupe A

Nous constatons que les points sont relativement alignés et tous à faible distance de l'approximation linéaire, qui est négative (le groupe A est moins risqué par définition que le groupe B. Pour les individus du groupe de risque C par rapport au groupe neutre B, nous obtenons le graphique suivant :

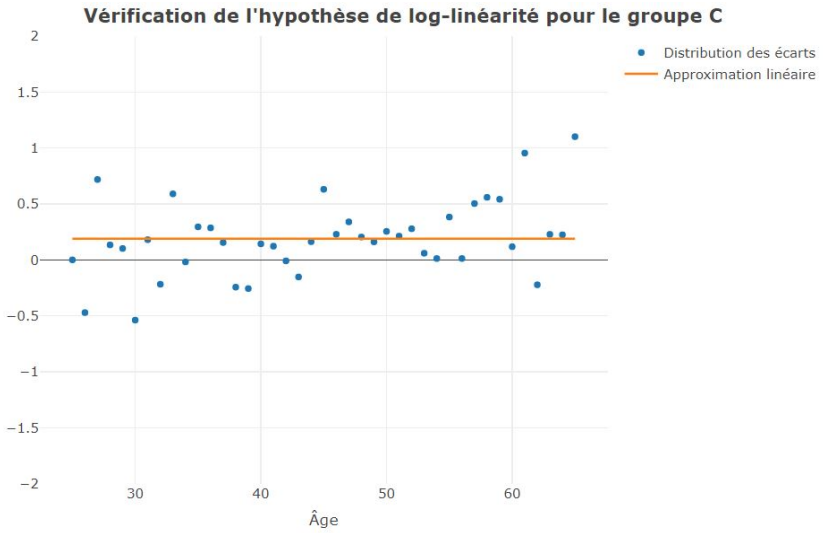

Figure 9.2: Vérification de l'hypothèse de log-linéarité pour le groupe C

Nous constatons que les points sont relativement alignés et tous à faible distance de l'approximation linéaire, qui est positive (le groupe C est plus risqué par définition que le groupe B. Pour les deux groupes, nous supposons donc l'hypothèse de log-linéarité vérifiée.

Nous pouvons donc appliquer le modèle de Cox à l'ensemble de nos données. Voici les résultats obtenus :

```
Call:
coxph(formula = Surv(age_entree, age_sortie, non_censure, type = "counting") ~
    grp_tarif, data = temp)

  n= 16988, number of events= 2183

               coef exp(coef) se(coef)      z Pr(>|z|)
grp_tarifA -0.19829   0.82013  0.05567 -3.562 0.000368 ***
grp_tarifC  0.19280   1.21264  0.04965  3.883 0.000103 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
grp_tarifA    0.8201     1.2193    0.7354    0.9147
grp_tarifC    1.2126     0.8246    1.1002    1.3366

Concordance= 0.54  (se = 0.006 )
Likelihood ratio test= 52.13  on 2 df,   p=5e-12
Wald test            = 51.58  on 2 df,   p=6e-12
Score (logrank) test = 52.05  on 2 df,   p=5e-12
```

Figure 9.3: Application du modèle de Cox

Nous faisons les interprétations suivantes :

— Pour les deux groupes de tarification, les *p-values* sont bien inférieures à 5% ; les résultats obtenus sur les coefficients sont donc bien significatifs.
— Pour le groupe A, le taux d'incidence h s'écrit : $h_A(x) = 0.8201 h_B(x)$. Cela correspond bien aux résultats attendus : le groupe A est par définition moins risqué que le groupe B ; les taux d'incidence sont donc inférieurs.
— Pour le groupe C, le taux d'incidence h s'écrit : $h_C(x) = 1.2126 h_B(x)$. Cela correspond bien aux résultats attendus : le groupe C est par définition plus risqué que le groupe B ; les taux d'incidence sont donc supérieurs.

Avant de passer aux étapes suivantes, nous cherchons d'abord à valider ces deux nouvelles lois d'incidence. Pour ce faire, nous suivons le même procédé que celui effectué pour le groupe B.

C'est à dire que nous allons tracer les nombres de sinistres prédits et réels par âge pour les individus associés aux groupes A et C issus d'une base de test isolée précédemment. En fonction de la justesse des prédictions, nous pourrons valider ou non les lois positionnées.

Nous commençons par tracer le graphique d'observés sur attendus pour le groupe d'individus peu risqués (*groupe A*).

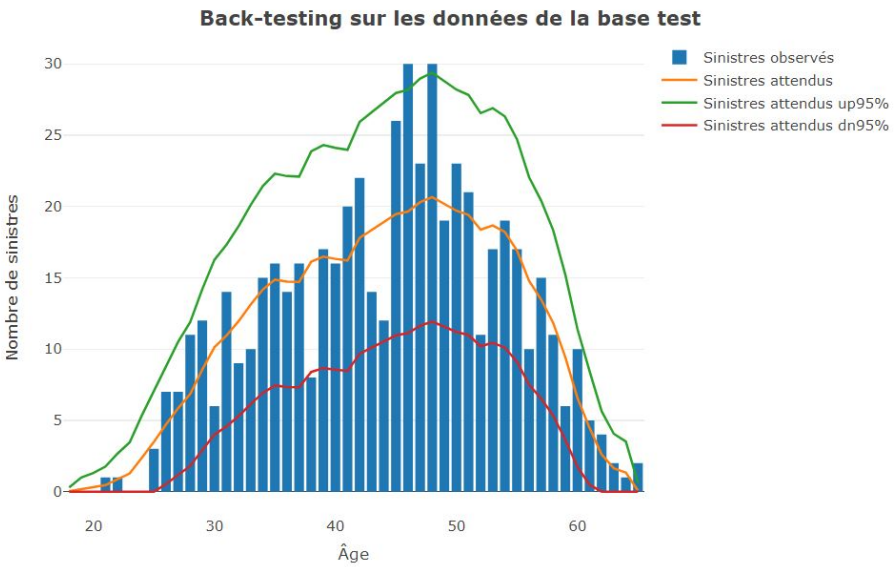

Figure 9.4: Validation de la loi positionnée pour le groupe A

Nous constatons que l'histogramme en bleu est bien contenu dans l'intervalle de confiance créé par les courbes vertes et jaunes. Nous pouvons ainsi valider la loi d'incidence pour le groupe A. Notons que le nombre maximal de sinistres observé culmine à 30, pour les individus d'âge 50 ans environ.

Nous continuons avec le graphique d'observés sur attendus pour le groupe d'individus plus risqués (*groupe C*).

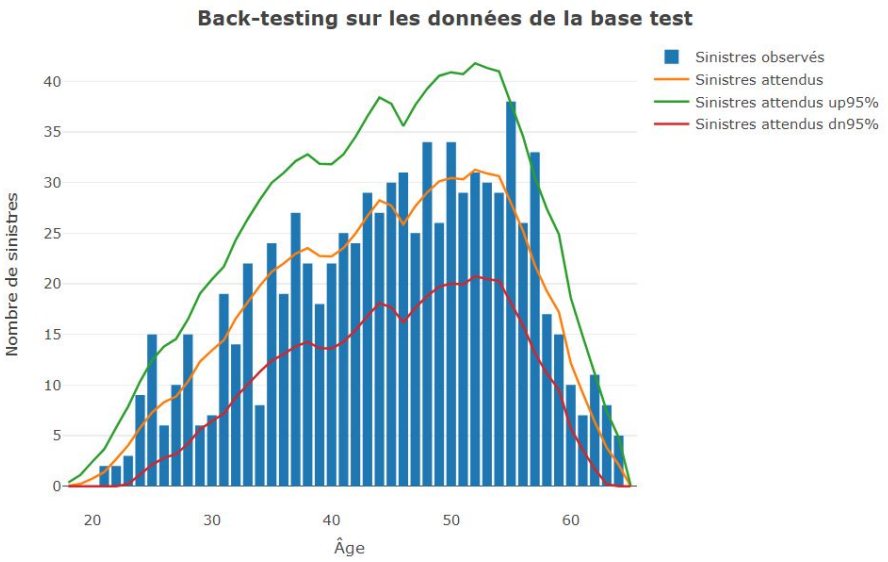

Figure 9.5: Validation de la loi positionnée pour le groupe C

Nous constatons que l'histogramme en bleu est bien contenu dans l'intervalle de confiance créé par les courbes vertes et jaunes. Nous pouvons ainsi valider la loi d'incidence pour le groupe C. Notons que le nombre maximal de sinistres observé culmine à 38, ce qui est environ 25% de plus que pour le groupe d'individus peu risqués, ce qui nous conforte dans la significativité de nos groupes de risque.

Comme les lois sont désormais validées, nous pouvons tracer les taux d'incidence de chacun des groupes :

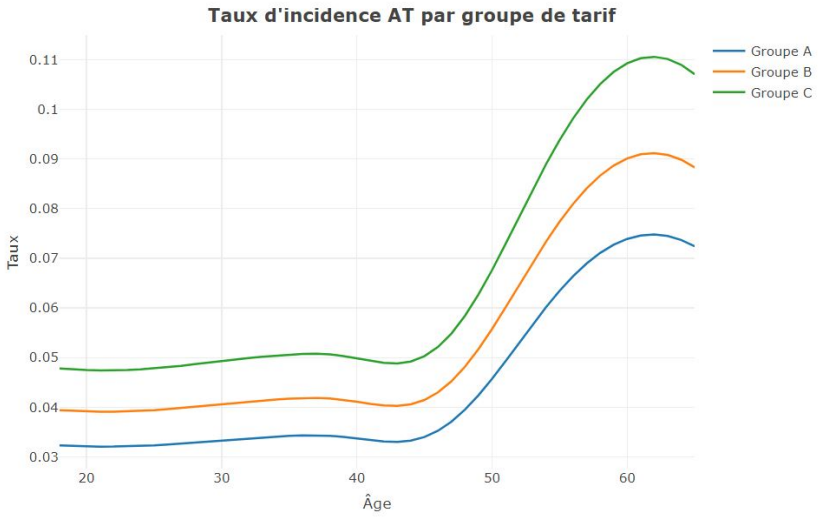

Figure 9.6: Taux d'incidence finaux pour tous les groupes de risque

Nous observons bien le positionnement relatif de groupes les uns rapport aux autres.

Troisième partie

Tarification

10 Introduction à la tarification

Dans cette nouvelle grande partie, nous parlerons de tarification. En effet, nous sommes jusqu'ici parvenus à :

— Identifier des variables discriminantes pour l'incidence en arrêt de travail.
— Modéliser des lois d'incidence par âge et par groupe de risque en fonction de ces différentes variables.

Nous passons maintenant à la partie applicative du mémoire où nous parlerons brièvement de tarification, et tenterons de tirer des conclusions des résultats que nous obtiendrons. Pour ce faire, nous découperons notre approche en 3 parties :

Dans un premier temps, nous vous présenterons l'offre pour laquelle nous cherchons à calculer un nouveau tarif ; et introduirons quelques notions mathématiques relatives au calcul des primes pures.
Dans un second temps, nous procéderons à une re-ventilation de nos taux d'incidence modélisés pour s'adapter au format de la tarification.
Dans un dernier temps, nous présentons et interpréterons les résultats que nous obtenons après calcul d'une nouvelle grille tarifaire.

Nous commençons tout de suite par introduire les méthodes de tarification en AT, et donc par parler de calcul de primes pures.

10.1 Description de l'offre

Pour le calcul des tarifs finaux, nous avons suivi les modalités de l'offre décrite par le produit *Prévoyance Pro*[21]. Celui-ci ayant récemment été lancé sur le marché, l'objectif était de challenger les tarifs existants.

Les options disponibles pour l'assuré sont les suivantes :

1. Celui-ci doit choisir s'il souhaite se couvrir contre le risque *Accident seul* ou contre les risques *Accident* et *Maladie*.
2. L'assuré doit opter pour une durée de franchise pour chacun des risques contre lesquels il est assuré. Cette durée peut être de sept, quinze ou trente jours.
3. L'assuré peut choisir une option d'hospitalisation. Cette option permet, en cas d'hospitalisation, de revoir à la baisse les durées de franchise qui passent à trois jours.
4. Enfin, l'assuré peut choisir d'ajouter en plus de la garantie ITT d'autres garanties comme l'IPP ou l'IPT. Pour chacune d'elles, celui-ci doit renseigner un montant de garanti.

21. cf. introduction du mémoire pour la présentation des produits.

10.2 Méthode de tarification

Nous utilisons pour la tarification en arrêt de travail un modèle de type Probabilité × Charges. Dans ce modèle, les tarifs en incapacité sont définis par :

— Le taux d'incidence de l'arrêt de travail pour un âge, une cause, une franchise et un groupe de risque donné.
— La charge probable de l'arrêt de travail (appelée charge de sinistre).

Les primes pures sont déduites à partir de l'égalisation entre les engagements de l'assuré et les engagements de l'assureur. L'engagement de l'assuré est de payer la prime pure à terme échus. L'engagement de l'assureur est de payer le montant de la rente mensuelle en cas de maintien de l'assuré en incapacité au-delà de la durée de franchise prévue dans le contrat et dans la limite de durée contractuelle. Nous obtenons les égalités suivantes :

$$\text{Engagement de l'assuré} = \text{Engagement de l'assureur} \tag{10.1}$$

$$P = q_x \times CS(x, d, n, G) \tag{10.2}$$

avec :

— CS le montant probable de la charge de sinistre.
— q_x le taux d'incidence en incapacité à l'âge x.
— d le nombre maximum de versements mensuels de la rente d'incapacité.
— n la durée de franchise en jours.
— G le montant garanti défini dans les termes du contrat.

Comme nous l'écrivons dans l'équation, le montant probable de la charge de sinistre dépend de variables qui sont propres à chaque contrat d'assurance. Dans le cas de la précédente loi d'incidence, les données mentionnées en introduction de partie étaient suffisantes. Dans le cas de la nouvelle loi modélisée, nous cherchons également à obtenir l'information sur :

1. La Catégorie Socio-Professionnelle de l'individu.
2. L'établissement de souscription du contrat.

Ces deux variables, comme nous l'avons présenté dans les parties antérieures du mémoire, nous permettent d'attribuer à l'assuré un coefficient de majoration ou de minoration en fonction de son profil de risque. Ainsi, nous aurons les trois formules suivantes pour les trois groupes de risques définis précédemment :

$$PP_x^{ITT} = 0.8201\ q_x\ CS(x,d,n,G)\ \textit{(groupe peu risqué, A)} \tag{10.3}$$

$$PP_x^{ITT} = q_x\ CS(x,d,n,G)\ \textit{(groupe neutre, B)} \tag{10.4}$$

$$PP_x^{ITT} = 1.2126\ q_x\ CS(x,d,n,G)\ \textit{(groupe risqué, C)} \tag{10.5}$$

Nous pouvons, à partir de ces formules, calculer l'ensemble des primes pures par âge pour les assurés du portefeuille. Lors de ce calcul néanmoins, il arrive que les primes présentent des aspects trop bruts que nous voulons corriger pour présenter aux assurés une offre cohérente. Ainsi, afin de respecter l'augmentation des primes par âge, les taux d'incidence q_x doivent être lissés. L'objectif est de :

1. Faire en sorte que les cotisations soient croissantes par âge. De cette manière, nous évitons aux clients de vivre une baisse de tarif une année et une forte hausse l'année suivante.
2. Faire en sorte d'éviter d'avoir des variations de tarifs trop importantes pour une année donnée. Nous prévoyons ainsi un lissage sur plusieurs années.

L'intervalle d'âge à lisser et la méthode de lissage ne peuvent néanmoins être définis à l'avance. Un test de cohérence sur l'augmentation des primes par âge doit être effectué sur tous les âges tarifés.

Dans le cadre de notre étude, nous utiliserons un outil de tarification mis en place par la direction produits pour la prévoyance individuelle au sein de Natixis Assurances. Cet outil nous permettra de calculer tous les différents tarifs dont nous avons besoin, pour des paramètres d'entrée (franchise, âge, montant garanti, ...) différents.

Avant de passer à la phase de calcul des tarifs néanmoins, nous devons adapter nos lois d'incidence à cet outil de tarification. Pour ce faire, nous suivons certaines étapes précises que nous détaillons dans la prochaine partie.

11 Ventilation des taux d'incidence

Dans cette partie, nous parlerons de l'étape re-ventilation des taux d'incidence par âge. En effet, les taux que nous avons modélisés ne sont pas encore parfaitement adaptés à l'outil de tarification que nous allons utiliser.

Pour calculer des tarifs cohérents avec l'offre proposée par Natixis Assurances, il s'agit d'exprimer des taux d'incidence par âge selon les deux variables suivantes :

— Selon le type de risque couvert : l'accident ou la maladie.
— Selon la durée de franchise prévue pour chaque type de risque.

Nous allons donc détailler notre approche pour parvenir à ces taux d'incidence finaux. Dans un premier temps, nous effectuerons la distinction selon le type de risque couvert. Dans un deuxième temps, nous effectuerons l'interpolation de nos lois d'incidence à toutes les durées de franchise en utilisant les tables de maintien du BCAC [22].

11.1 Distinction Accident / Maladie

Pour effectuer la distinction sur l'incidence *Accident* ou *Maladie*, nous repartons des taux modélisés pour le groupe de risque neutre (*groupe B*) et la franchise de type *Accident 7 jours, Maladie 15 jours*.

Lorsque nous observons un sinistre dans notre base de données, celui-ci est défini par plusieurs variables comme : la date de survenance, le type de franchise, le montant de la prestation, etc. Une de ces variables nous intéresse particulièrement : c'est la cause de survenance. En utilisant cette variable en effet, nous sommes en mesure de rattacher un sinistre à une cause de type *Accident* ou de type *Maladie*.

Sachant que notre loi d'incidence par âge a été faite sans distinction des causes, il est envisageable maintenant de définir une loi d'incidence pour *l'Accident* et pour la *Maladie* en calculant une répartition empirique et en appliquant les coefficients obtenus aux taux globaux. Cette méthode nous permet d'obtenir finalement des taux d'incidence en arrêt de travail ventilés par type de survenance, ce qui est nécessaire pour la bonne application de l'outil de tarification.

Il est important de noter toutefois que nous n'aurions que difficilement pu modéliser à l'origine deux lois d'incidence pour ces deux types de survenance. En effet, l'exposition aurait été trop faible et les étapes de la modélisation moins sensées. Un des inconvénients de cette méthode néanmoins reste le fait que nous ne bénéficions pas de l'information sur le type de survenance pour tous les sinistres. Nous avons quand même assez d'information à notre disposition pour calculer une répartition empirique sensée.

22. Bureau Commun des Assurances Collectives

Pour les individus ayant souscrit à un contrat avec une franchise de type *Accident 7 jours, Maladie 15 jours*, nous obtenons la répartition par âge suivante :

Âge	% Accident	% Maladie			
18	48%	52%	42	48%	52%
19	48%	52%	43	48%	52%
20	48%	52%	44	41%	59%
21	48%	52%	45	41%	59%
22	48%	52%	46	33%	67%
23	48%	52%	47	33%	67%
24	48%	52%	48	33%	67%
25	48%	52%	49	30%	70%
26	48%	52%	50	30%	70%
27	48%	52%	51	30%	70%
28	48%	52%	52	34%	66%
29	48%	52%	53	34%	66%
30	55%	45%	54	34%	66%
31	55%	45%	55	34%	66%
32	55%	45%	56	34%	66%
33	55%	45%	57	34%	66%
34	53%	47%	58	34%	66%
35	53%	47%	59	34%	66%
36	53%	47%	60	34%	66%
37	53%	47%	61	34%	66%
38	51%	49%	62	34%	66%
39	51%	49%	63	34%	66%
40	51%	49%	64	34%	66%
41	48%	52%	65	34%	66%

Figure 11.1: Répartition Accident / Maladie par âge

Nous observons des résultats intéressants. En effet :

— Nous remarquons que, pour les jeunes individus, la répartition entre l'Accident et la Maladie est quasiment égale.
— Nous observons un faible pic d'incidence en Accident pour les individus âgés entre 30 et 40 ans. Cela pourrait traduire une période de résistance à la maladie.
— Nous constatons que, pour les individus les plus âgés, quasiment 70% des sinistres sont dûs à des maladies plutôt qu'à des accidents. Cela semble traduire le fait que les individus âgés ont une santé plus fragile et sont donc plus susceptibles de s'arrêter pour ces raisons.

À partir de ce moment-là, nous appliquons ces taux de répartition à notre loi d'incidence pour la franchise *Accident 7 jours, Maladie 15 jours*, et obtenons les taux présentés en page suivante.

Âge	acc_7			Âge	mal_15		
18	1,88%	42	1,93%	18	2,06%	42	2,10%
19	1,87%	43	1,93%	19	2,05%	43	2,10%
20	1,87%	44	1,68%	20	2,05%	44	2,38%
21	1,86%	45	1,72%	21	2,05%	45	2,43%
22	1,86%	46	1,43%	22	2,05%	46	2,86%
23	1,87%	47	1,51%	23	2,05%	47	3,01%
24	1,87%	48	1,61%	24	2,06%	48	3,21%
25	1,88%	49	1,54%	25	2,06%	49	3,63%
26	1,89%	50	1,66%	26	2,07%	50	3,91%
27	1,90%	51	1,79%	27	2,09%	51	4,21%
28	1,91%	52	2,20%	28	2,10%	52	4,26%
29	1,93%	53	2,35%	29	2,12%	53	4,55%
30	2,22%	54	2,50%	30	1,85%	54	4,83%
31	2,23%	55	2,60%	31	1,86%	55	5,13%
32	2,24%	56	2,72%	32	1,87%	56	5,37%
33	2,25%	57	2,83%	33	1,88%	57	5,58%
34	2,18%	58	2,91%	34	1,97%	58	5,75%
35	2,19%	59	2,98%	35	1,98%	59	5,88%
36	2,20%	60	3,03%	36	1,99%	60	5,98%
37	2,20%	61	3,06%	37	1,99%	61	6,03%
38	2,12%	62	3,07%	38	2,05%	62	6,05%
39	2,11%	63	3,05%	39	2,04%	63	6,03%
40	2,09%	64	3,02%	40	2,02%	64	5,96%
41	1,95%	65	2,97%	41	2,12%	65	5,86%

Figure 11.2: Taux d'incidence pour l'accident à 7 jours et la maladie à 15 jours

11.2 Interpolation à toutes les durées de franchise

Maintenant que nous avons obtenu les taux d'incidence par âge pour les deux types de risques identifiés en portefeuille, il s'agit d'obtenir ces taux pour toutes les durées de franchise possibles.

Pour rappel, l'objectif du mémoire est de challenger la loi précédemment modélisée en interne par le calcul d'un nouveau tarif que nous comparerons au nouveau tarif. Pour pouvoir effectuer cette comparaison, nous devons être en mesure de calculer différents tarifs en fonction de l'option choisi par l'assuré qui souscrit son contrat.

Plusieurs durées de garanties étant proposées, il s'agit d'avoir à disposition des taux d'incidence qui dépendent de ces durées de garanties. Pour ce faire, nos utiliserons la table de maintien en incapacité du BCAC [23] 2013.

23. Bureau Commun des Assurances Collectives

L'idée est la suivante :

— La table de maintien en incapacité du BCAC 2013 nous donne, sur un effectif de départ en incapacité à 10 000 individus, le nombre d'individus se maintenant en incapacité après trois, sept, quinze ou trente jours.
— À partir de là, nous pouvons calculer des coefficients d'extrapolation d'une durée de franchise à une autre.
— En appliquant ces coefficients à nos taux d'incidence, nous obtiendrons des taux pour l'Accident et la Maladie pour toutes les durées de franchise.

Bien que cette approche reste élémentaire, elle nous permet d'obtenir des taux cohérents et significatifs. Voici un aperçu de la table de maintien en incapacité pour les âges 18 à 30 ans :

Âge	Maintien à 0 jours	Maintien après 3	Maintien après 7	Maintien après 15	Maintien après 30
18	10000	9451	8718	7253	4414
19	10000	9451	8718	7253	4414
20	10000	9451	8718	7253	4414
21	10000	9473	8769	7363	4638
22	10000	9493	8818	7466	4848
23	10000	9511	8859	7554	5026
24	10000	9524	8889	7619	5159
25	10000	9535	8915	7675	5273
26	10000	9543	8934	7716	5357
27	10000	9550	8950	7751	5427
28	10000	9555	8962	7775	5477
29	10000	9560	8973	7800	5526
30	10000	9564	8983	7820	5567

Figure 11.3: Extrait de la table de maintien en incapacité

Au jour zéro, 10 000 individus entrent en incapacité. Pour les individus âgés de 18 ans, nous lisons que 9 451 d'entre eux se maintiennent après trois jours. Après trente jours, il ne restera néanmoins plus que 4 414 individus.

Si l'on essaie de faire la même interprétation pour un individu d'âge 30 ans, nous constatons une légère différence. Après trente jours, il semble qu'il reste encore 5 567 individus en incapacité. Cela est significativement plus que pour les individus de 18 ans.

Cette table de maintien en incapacité tient en effet compte des différents profils des individus en fonction de leur âge. Plus un individu est âgé, plus il est susceptible de rester longtemps en incapacité, au contraire d'un individu plus jeune. Comme la loi modélisée pour la maladie est modélisée sur une durée de franchise de quinze jours, nous aurons besoin des coefficients de passage entre quinze jours et les autres durées.

En calculant simplement les rapports entre nombre d'individus restants à l'instant $t+1$ et nombres d'individus présents à l'instant t, nous obtenons pour les âges de 18 à 30 ans les coefficients suivants :

Âge	Coefficient 0j - 15j	Coefficient 3j - 15j	Coefficient 7j - 15j	Coefficient 30j - 15j
18	1,38	1,30	1,20	0,61
19	1,38	1,30	1,20	0,61
20	1,38	1,30	1,20	0,61
21	1,36	1,29	1,19	0,63
22	1,34	1,27	1,18	0,65
23	1,32	1,26	1,17	0,67
24	1,31	1,25	1,17	0,68
25	1,30	1,24	1,16	0,69
26	1,30	1,24	1,16	0,69
27	1,29	1,23	1,15	0,70
28	1,29	1,23	1,15	0,70
29	1,28	1,23	1,15	0,71
30	1,28	1,22	1,15	0,71

Figure 11.4: Coefficients de passage pour la maladie à 15 jours

La loi modélisée pour l'accident étant modélisée sur une durée de franchise de sept jours, nous aurons besoin des coefficients de passage entre sept jours et les autres durées :

Âge	Coefficient 0j - 7j	Coefficient 3j - 7j	Coefficient 15j - 7j	Coefficient 30j - 7j
18	1,15	1,08	0,83	0,51
19	1,15	1,08	0,83	0,51
20	1,15	1,08	0,83	0,51
21	1,14	1,08	0,84	0,53
22	1,13	1,08	0,85	0,55
23	1,13	1,07	0,85	0,57
24	1,12	1,07	0,86	0,58
25	1,12	1,07	0,86	0,59
26	1,12	1,07	0,86	0,60
27	1,12	1,07	0,87	0,61
28	1,12	1,07	0,87	0,61
29	1,11	1,07	0,87	0,62
30	1,11	1,06	0,87	0,62

Figure 11.5: Coefficients de passage pour l'accident à 7 jours

Dans les deux cas, nous constatons que :

— Lorsque la durée de franchise d'arrivée est supérieure à la durée de franchise de départ, le coefficient calculé est inférieur à 1.

— Lorsque la durée de franchise de départ est supérieure à la durée de franchise d'arrivée, le coefficient calculé est supérieur à 1.

Ce résultat est cohérent : plus la durée de franchise est courte, plus le taux d'incidence en incapacité est grand et inversement. Finalement, après multiplication des taux entre eux, nous obtenons les taux d'incidence finaux suivants :

Âge	acc_0	acc_3	acc_7	acc_15	acc_30
18	2,15%	2,04%	1,88%	1,56%	0,95%
19	2,15%	2,03%	1,87%	1,56%	0,95%
20	2,14%	2,02%	1,87%	1,55%	0,94%
21	2,12%	2,01%	1,86%	1,56%	0,99%
22	2,11%	2,01%	1,86%	1,58%	1,02%
23	2,11%	2,00%	1,87%	1,59%	1,06%
24	2,10%	2,00%	1,87%	1,60%	1,09%
25	2,11%	2,01%	1,88%	1,62%	1,11%
26	2,11%	2,02%	1,89%	1,63%	1,13%
27	2,12%	2,03%	1,90%	1,65%	1,15%
28	2,13%	2,04%	1,91%	1,66%	1,17%
29	2,15%	2,05%	1,93%	1,67%	1,19%
30	2,47%	2,36%	2,22%	1,93%	1,37%
31	2,48%	2,37%	2,23%	1,95%	1,39%
32	2,49%	2,39%	2,24%	1,96%	1,41%
33	2,50%	2,40%	2,25%	1,97%	1,42%
34	2,42%	2,32%	2,18%	1,91%	1,37%
35	2,44%	2,33%	2,19%	1,91%	1,37%
36	2,45%	2,34%	2,20%	1,91%	1,37%
37	2,45%	2,34%	2,20%	1,91%	1,36%
38	2,37%	2,26%	2,12%	1,85%	1,31%
39	2,35%	2,25%	2,11%	1,83%	1,30%
40	2,33%	2,23%	2,09%	1,82%	1,29%
41	2,17%	2,08%	1,95%	1,70%	1,21%
42	2,15%	2,06%	1,93%	1,69%	1,21%
43	2,14%	2,05%	1,93%	1,68%	1,21%
44	1,87%	1,79%	1,68%	1,47%	1,06%
45	1,91%	1,83%	1,72%	1,50%	1,08%
46	1,59%	1,52%	1,43%	1,26%	0,91%
47	1,67%	1,60%	1,51%	1,32%	0,97%
48	1,78%	1,70%	1,61%	1,42%	1,04%
49	1,70%	1,63%	1,54%	1,36%	1,02%
50	1,83%	1,76%	1,66%	1,48%	1,11%
51	1,96%	1,89%	1,79%	1,60%	1,22%
52	2,40%	2,31%	2,20%	1,97%	1,52%
53	2,56%	2,47%	2,35%	2,12%	1,66%
54	2,71%	2,62%	2,50%	2,26%	1,79%
55	2,81%	2,72%	2,60%	2,36%	1,88%
56	2,94%	2,85%	2,72%	2,47%	1,99%
57	3,05%	2,95%	2,83%	2,57%	2,09%
58	3,14%	3,04%	2,91%	2,66%	2,16%
59	3,22%	3,12%	2,98%	2,72%	2,20%
60	3,27%	3,17%	3,03%	2,75%	2,22%
61	3,30%	3,20%	3,06%	2,78%	2,23%
62	3,32%	3,21%	3,07%	2,78%	2,23%
63	3,30%	3,20%	3,05%	2,77%	2,21%
64	3,27%	3,16%	3,02%	2,74%	2,18%
65	3,22%	3,11%	2,97%	2,69%	2,14%

Figure 11.6: Taux d'incidence finaux pour l'accident

Âge	mal_0	mal_3	mal_7	mal_15	mal_30
18	2,84%	2,69%	2,48%	2,06%	1,26%
19	2,83%	2,68%	2,47%	2,05%	1,25%
20	2,83%	2,67%	2,46%	2,05%	1,25%
21	2,78%	2,63%	2,44%	2,05%	1,29%
22	2,74%	2,60%	2,42%	2,05%	1,33%
23	2,71%	2,58%	2,40%	2,05%	1,36%
24	2,70%	2,57%	2,40%	2,06%	1,39%
25	2,69%	2,56%	2,40%	2,06%	1,42%
26	2,69%	2,56%	2,40%	2,07%	1,44%
27	2,69%	2,57%	2,41%	2,09%	1,46%
28	2,70%	2,58%	2,42%	2,10%	1,48%
29	2,71%	2,59%	2,43%	2,12%	1,50%
30	2,37%	2,26%	2,13%	1,85%	1,32%
31	2,37%	2,27%	2,13%	1,86%	1,33%
32	2,38%	2,28%	2,14%	1,87%	1,34%
33	2,39%	2,29%	2,15%	1,88%	1,36%
34	2,51%	2,40%	2,26%	1,97%	1,42%
35	2,52%	2,41%	2,27%	1,98%	1,42%
36	2,54%	2,43%	2,28%	1,99%	1,42%
37	2,54%	2,43%	2,28%	1,99%	1,41%
38	2,63%	2,51%	2,36%	2,05%	1,45%
39	2,61%	2,50%	2,35%	2,04%	1,45%
40	2,59%	2,47%	2,32%	2,02%	1,43%
41	2,71%	2,59%	2,43%	2,12%	1,51%
42	2,68%	2,56%	2,41%	2,10%	1,50%
43	2,67%	2,55%	2,40%	2,10%	1,50%
44	3,02%	2,89%	2,72%	2,38%	1,71%
45	3,08%	2,95%	2,78%	2,43%	1,75%
46	3,63%	3,47%	3,27%	2,86%	2,07%
47	3,80%	3,64%	3,43%	3,01%	2,20%
48	4,02%	3,86%	3,64%	3,21%	2,36%
49	4,52%	4,34%	4,10%	3,63%	2,70%
50	4,84%	4,65%	4,40%	3,91%	2,95%
51	5,18%	4,98%	4,73%	4,21%	3,22%
52	5,19%	5,00%	4,75%	4,26%	3,29%
53	5,51%	5,31%	5,06%	4,55%	3,57%
54	5,81%	5,61%	5,35%	4,83%	3,83%
55	6,13%	5,93%	5,66%	5,13%	4,10%
56	6,38%	6,18%	5,91%	5,37%	4,33%
57	6,60%	6,40%	6,12%	5,58%	4,52%
58	6,79%	6,59%	6,31%	5,75%	4,67%
59	6,97%	6,75%	6,46%	5,88%	4,76%
60	7,11%	6,88%	6,58%	5,98%	4,82%
61	7,18%	6,95%	6,65%	6,03%	4,85%
62	7,22%	6,99%	6,67%	6,05%	4,84%
63	7,20%	6,96%	6,65%	6,03%	4,82%
64	7,13%	6,90%	6,59%	5,96%	4,76%
65	7,02%	6,78%	6,48%	5,86%	4,66%

Figure 11.7: Taux d'incidence finaux pour la maladie

Maintenant que nous avons obtenu les taux d'incidence finaux, nous pouvons passer au calcul des tarifs et à l'interprétation des résultats.

12 Calcul des tarifs finaux

Dans cette dernière partie, l'objectif est de présenter les résultats obtenus après calcul des tarifs finaux par âge, offre et groupe de risque. Pour des soucis de confidentialité, les résultats chiffrés que nous présenterons ont été désensibilisés, sans perte de généralisation.

12.1 Calcul d'une nouvelle grille tarifaire

Soit un individu ayant souscrit un contrat avec les caractéristiques suivantes :

- **Type de franchise :** Absolue.
- **Risques couverts :** Accident et Maladie en ITT.
- **Durée des franchises :** 7 jours pour l'Accident et 15 jours pour la Maladie.
- **Option hospitalisation :** Non.
- **Montant assuré :** 1 €.

Le tarif calculé à partir de l'ancienne loi est le suivant :

Âge	Tarif		
18	80 €	42	119 €
19	80 €	43	124 €
20	80 €	44	129 €
21	83 €	45	137 €
22	86 €	46	150 €
23	88 €	47	167 €
24	90 €	48	186 €
25	92 €	49	210 €
26	95 €	50	236 €
27	97 €	51	260 €
28	100 €	52	283 €
29	102 €	53	305 €
30	98 €	54	328 €
31	101 €	55	356 €
32	105 €	56	384 €
33	107 €	57	428 €
34	109 €	58	474 €
35	109 €	59	505 €
36	110 €	60	527 €
37	110 €	61	533 €
38	111 €	62	525 €
39	112 €	63	516 €
40	114 €	64	499 €
41	117 €	65	446 €
		66	343 €

Figure 12.1: Tarif calculé à partir de l'ancienne loi

Nous calculons aussi le tarif à partir de la nouvelle loi modélisée :

Âge	Tarif
18	109 €
19	109 €
20	109 €
21	113 €
22	116 €
23	119 €
24	123 €
25	126 €
26	130 €
27	134 €
28	139 €
29	143 €
30	145 €
31	149 €
32	153 €
33	157 €
34	160 €
35	163 €
36	165 €
37	166 €
38	168 €
39	169 €
40	170 €
41	172 €
42	173 €
43	176 €
44	181 €
45	188 €
46	201 €
47	219 €
48	242 €
49	271 €
50	306 €
51	345 €
52	385 €
53	429 €
54	473 €
55	519 €
56	562 €
57	597 €
58	619 €
59	621 €
60	613 €
61	609 €
62	600 €
63	585 €
64	560 €
65	490 €
66	377 €

Figure 12.2: Tarif calculé à partir de la nouvelle loi

Par simple lecture des valeurs, nous constatons que le tarif calculé à partir de la nouvelle loi est légèrement plus prudent que celui calculé à partir de l'ancienne loi. Nous devons aussi penser à considérer les coefficients de majoration et de minoration :

Coefficients de majoration	
A	0,82
B	1
C	1,21

Figure 12.3: Coefficients de majoration / minoration pour le nouveau tarif

12.2 Application au portefeuille d'assurés

Afin d'aller jusqu'au bout de la comparaison, nous décidons d'utiliser l'historique de données à disposition pour calculer de nouveaux tarifs. Nous recalculons donc pour les assurés de la base historisée de 2018 les valeurs des primes à partir de l'ancienne loi et de la nouvelle loi et effectuons une comparaison.

Voici un extrait des résultats obtenus :

Exposition	Nouveau groupe de tarif	Âge	Risques couverts	Type de franchise	Durée de l'accident	Durée de la maladie	Hospitalisation	Montant assuré ITT	Montant assuré IPP / IPT	Prime loi 2018	Prime nouvelle loi
1,00	A	64	ITT	Relative	30	30	Non	2 545,83 €	- €	1 312 €	1 207,10 €
1,00	C	60	ITT	Relative	30	30	Non	1 501,73 €	- €	813 €	1 148,97 €
1,00	B	64	ITT	Relative	30	30	Non	2 545,83 €	- €	1 312 €	1 472,08 €
1,00	B	60	ITT	Relative	30	30	Non	3 818,83 €	- €	2 069 €	2 414,69 €
1,00	B	62	ITT	Relative	30	30	Non	2 545,83 €	- €	1 376 €	1 574,38 €
1,00	B	63	ITT	Absolue	7		Non	1 143,37 €	- €	209 €	216,70 €
1,00	A	60	ITT	Relative	30	30	Non	1 290,79 €	- €	699 €	669,27 €
1,00	B	45	IPT+ITT	Relative	30	30	Non	800,34 €	9 604,56 €	158 €	215,41 €
1,00	C	60	ITT	Relative	30	30	Non	833,84 €	- €	452 €	637,97 €
1,00	B	56	ITT	Relative	30	30	Non	1 109,61 €	- €	440 €	642,17 €
1,00	B	61	ITT	Relative	30	30	Non	1 527,49 €	- €	837 €	958,58 €
1,00	C	60	ITT	Absolue	30	30	Non	3 003,58 €	- €	1 442 €	2 036,42 €
1,00	B	59	ITT	Relative	30	30	Non	2 382,55 €	- €	1 235 €	1 525,30 €
1,00	C	60	ITT	Absolue	7		Non	1 059,28 €	- €	195 €	254,80 €
1,00	C	58	ITT	Relative	30	30	Non	1 829,39 €	- €	890 €	1 411,02 €

Figure 12.4: Recalcul des primes sur la base historisée de 2018

De cette comparaison ressort le fait que notre nouvelle loi d'incidence modélisée semble prudente. En effet, les tarifs calculés sont souvent légèrement plus élevés que ceux calculés à partir de l'ancienne loi. Toutefois, nous observons que pour les personnes identifiées comme peu risquées, les nouveaux tarifs sont parfois plus faibles.

Pour terminer la comparaison, nous avons aussi essayer de comparer les ratios sinistres sur prime. Il est cependant difficile de tirer des conclusions de cette comparaison. Les provisions en AT sont en effet souvent très élevées à cause des invalidités potentielles. L'approche dans le calcul du tarif est donc en général prudente ce qui conduit à des ratios S/P faibles. Ainsi, dans le cadre de notre étude, les deux ratios S/P calculés avec l'ancienne loi et la nouvelle loi sont assez proches.

Pour expliquer cette marge de prudence rajouté par la nouvelle loi d'incidence, nous soulignons le fait que :

— D'une part, les taux ont été modélisés *avant refus*. C'est à dire qu'en réalité, un pourcentage de l'incidence modélisée est refusée par l'assureur et donc ne fait pas partie de l'incidence à proprement parler.
— D'autre part, les taux ont été re-ventilés ensuite en utilisant la table de maintien du BCAC 2013. Cette table propose une approche prudente et donc participe à augmenter implicitement le taux d'entrée modélisé.

Quatrième partie

Conclusion

Dans le cadre de ce mémoire, nous avons souhaité mettre en avant des axes innovant de tarification pour le risque d'arrêt de travail et challenger la loi d'expérience précédemment existante. Cette étude trouve son importance dans le fait qu'il est nécessaire d'assurer régulièrement un suivi des lois modélisées en interne pour vérifier que celles-ci parviennent toujours bien à prendre en compte les phénomènes régissant le portefeuille d'assurés et que le tarif proposé est toujours cohérent. D'autre part, la recherche de nouveaux axes discriminants constitue une problématique nouvelle et permet d'envisager la possibilité de développer une offre plus ciblée.

Nous avons décidé d'aborder le problème selon une démarche méthodique. Après avoir présenté l'activité de Natixis Assurances et le Groupe BPCE, ainsi que notre base de travail, nous avons entamé la partie de recherche de nouvelles variables discriminantes.

Dans cette première partie, nous avons effectué une analyse descriptive de notre base de données, avec l'objectif d'identifier des premières connexions entre les variables et l'incidence en incapacité. Dans une optique de perfectionnement, nous avons ensuite implémenté des algorithmes d'apprentissage non-supervisés. Les objectifs étaient multiples : tenter de réduire les modalités de nos variables et continuer l'identification des nouveaux axes discriminants. Enfin, par opposition, nous avons appliqué une méthode d'apprentissage supervisé. Les résultats de cette partie ont mis en évidence l'importance de la Catégorie Socio-Professionnelle et de l'Établissement de souscription du contrat relativement à l'incidence en incapacité. Pour la suite du mémoire, nous avons décidé de conserver ces variables pour créer des groupes d'assurés aux profils de risque distincts.

Dans la deuxième partie, nous avons introduit les modèles de survie et tenté de les appliquer à notre base de données, tout en prenant en compte l'impact discriminant des axes principaux identifiés auparavant. De cette manière, nous avons dans un premier temps effectué un calcul des taux bruts en utilisant l'estimateur de Kaplan-Meier, que nous avons ensuite lissés par application de la méthode de Whittaker-Henderson. À partir des résultats obtenus pour le groupe d'individus de risque neutre, nous avons modélisé par positionnement et par application du modèle de Cox les lois d'incidence des deux autres groupes d'individus : le groupe A (groupe peu risqué) et le groupe C (groupe risqué). Pour mettre en évidence ces groupes, nous avons étudié le croisement des variables *Catégorie Socio-Professionnelles* et *Établissement*. Finalement, nous obtenons une loi d'incidence par âge pour chacun des groupes de risques.

Dans la dernière partie, nous avons introduit les bases de la tarification pour le risque d'incapacité et avons présenté la structure de l'offre *Prévoyance Pro*. Dans un deuxième temps, nous avons présenté les étapes d'adaptation des lois modélisées aux outils de tarification mis à disposition dans l'équipe Actuariat Produits Prévoyance. En effet, il était nécessaire d'effectuer une distinction selon la cause de survenance de l'incapacité : l'accident et la maladie. Il était également nécessaire de produire des lois d'incidence dépendantes de la durée de franchise. Dans un dernier temps, nous avons effectué un calcul des primes et interprété les résultats finaux. Nous avons observé que les nouveaux tarifs calculés étaient plus élevés en moyenne pour les individus du portefeuille. Cela s'est justifié par nos différents positionnements prudents au cours de l'étude.

En conclusion, cette étude nous a permis de mettre en lumière plusieurs évènements intéressantes. Par exemple :

1. Nous avons pu montrer qu'il existait des axes discriminants pour l'incidence en incapacité différents de ceux employés usuellement.
2. Nous avons pu montrer la complémentarité de l'utilisation d'algorithmes de classification et de prédiction dans le cadre de l'estimation de nouveaux taux d'incidence pour l'incapacité.
3. Nous avons pu challenger la loi précédemment utilisée au sein du service et tirer des conclusions sur les résultats que nous observons aujourd'hui.

Pour les années à venir, il sera intéressant d'étudier l'évolution des taux d'incidence en incapacité pour confirmer la pertinence de cette nouvelle loi modélisée et pour nous permettre de prendre une décision.

Bien que nous sommes parvenus dans ce mémoire à mettre en lumière plusieurs évènements intéressants, il reste certains points sur lesquels nous aurions pu parfaire notre étude.

Lors de la création des groupes de risque, nous nous sommes limités au deux variables *Catégorie Socio-Professionnelles* et *Établissement* par soucis d'exposition. Il aurait été intéressant, si cela avait été possible, d'essayer de prendre en compte un nombre de nouvelles variables plus important.

Au moment de la ventilation des taux d'incidence selon les durées de franchise, nous avons utilisé la table de maintien BCAC 2013. Bien que l'approche reste cohérente, cette table présente une marge de prudence non-négligeable que nous avons donc embarqué, en plus de la sélection des taux *avant refus*, dans notre modélisation des taux.

Pour approfondir l'étude actuelle, le croisement des données avec d'autres services d'assurance est envisageable. Nous pourrions aussi appliquer des méthodes de Machine Learning plus complexes. Ce mémoire s'inscrit dans une démarche de recherche de la part de l'équipe Actuariat et de la prévoyance individuelle. Les résultats mis en avants ici ont permis au développement des connaissances métiers et ouvrent la porte à l'utilisation de nouvelles variables, non utilisées jusqu'à-lors, dans un contexte de tarification.

Bibliographie

[1] Agudelo, F. (2016). *Analyse géospatiale des modèles de durée appliqués aux risques de souscription en Prévoyance.* Mémoire d'actuariat.

[2] Brownlee, J. (2020). *Supervised and Unsupervised Machine Learning Algorithms.*

[3] Cox, D. (1975). *Partial Likelihood.*

[4] Garbade, M. (2018). *Understanding K-Means clustering in Machine Learning.*

[5] Gill, J. (2000). *Generalized Linear Models : A Unified Approach.*

[6] INSEE (2003). *Nomenclature des professions et des catégories socio-professionnelles.*

[7] Kaplan, E. (1958). *Nonparametric estimation from incomplete observations.*

[8] Kryshtopenko, A. (2018). *Étude de l'attrition en Prévoyance Individuelle.* Mémoire d'actuariat.

[9] Mishra, S. (2017). *Unsupervised learning and Data clustering.*

[10] Planchet, F. (2020a). *Modèles de durée : méthodes de lissage et d'ajustement.*

[11] Planchet, F. (2020b). *Modèles de durée : statistique des modèles non-paramétriques.*

[12] Reddy, C. (2018). *Understanding the concept of Hierarchical clustering Technique.*

[13] Rouaud, M. (2013). *Probabilités, statistiques et analyses multicritères.*

[14] Saporta, G. (2006). *Probabilités, analyse des données et statistiques.*

[15] STHDA (2017). *Analyse Factorielle des Données Mixtes.*

[16] Székely, G. (2005). *Hierarchical clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method.*

Table des figures

2.1 Organigramme du Groupe BPCE et de Natixis Assurances 2
2.2 Distribution des produits au sein du portefeuille des Banques Populaires 4
3.1 Les trois premières colonnes de la base solvency 8
3.2 Base à la maille année d'observation / client 9
3.3 Base après retraitement 11
3.4 Exposition par région 13
3.5 Exposition par banque 14
3.6 Taux d'expatriation par banque 15
3.7 Les catégories socio-professionnelles avant 1982 (source : INSEE) 16
3.8 Les catégories socio-professionnelles après 1982 (source : INSEE) 17
3.9 La répartition des clients par CSP dans notre base de travail 18
4.1 Taux d'incidence AT par région 20
4.2 Taux d'incidence AT par entité bancaire 21
4.3 Taux d'incidence en incapacité par âge 22
4.4 Taux d'incidence AT par ancienneté 23
4.5 Taux d'incidence en incapacité par Catégorie Socio-Professionnelle 24
4.6 Taux d'incidence AT par tranche de montant garanti 25
4.7 Taux d'incidence AT par montant de prime 26
5.1 Exemple de classification produite par l'algorithme CAH 29
5.2 La base de travail pour la réduction des modalités de région 33
5.3 Corrélation entre les variables de sinistralité par région 34
5.4 Tableau des valeurs propres de l'ACP 34
5.5 Contribution de chaque dimension à expliquer la variabilité totale 34
5.6 Position des individus sur l'espace des dimensions de l'ACP 35
5.7 Représentation de la classification sous forme de clusters 36
5.8 Représentation de la classification sous forme d'une arborescence 37
5.9 Position des modalités de la variable banque sur l'espace de dimension réduit défini par l'ACP 38
5.10 Représentation de la classification sous forme de clusters 38
5.11 Représentation en arborescence de la classification de la variable Banque 39
5.12 Position des individus CSP sur l'espace de l'ACP 40
5.13 Représentation en arborescence des classes de modalités de la CSP 41
5.14 Base de travail pour l'AFDM 43
5.15 Tableau des valeurs propres 44
5.16 Analyse des corrélations des variables quantitatives 45
5.17 Analyse des corrélations des variables qualitatives (Composantes 2 et 3) 46
5.18 Analyse des corrélations des variables qualitatives (Composantes 4 et 5) 46
5.19 Variables qualitatives et quantitatives 47
5.20 Coloration des individus par CSP 48
5.21 Coloration des individus par région et par banque 48
5.22 Distribution de la sinistralité au sein des clusters 50
5.23 Distribution de l'âge pour le cluster de sinistralité faible 52
5.24 Distribution de l'âge pour le cluster de sinistralité faible 52
5.25 Distribution des régions pour le cluster de sinistralité faible 53
5.26 Distribution des banque pour le cluster de sinistralité faible 54
5.27 Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité faible 54
5.28 Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité faible 55

5.29 Distribution des CSP pour le cluster de sinistralité faible 55
5.30 Distribution de l'âge pour le cluster de sinistralité forte 56
5.31 Distribution de l'âge pour le cluster de sinistralité forte 56
5.32 Distribution des régions pour le cluster de sinistralité forte 57
5.33 Distribution des banques pour le cluster de sinistralité forte 58
5.34 Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité forte 58
5.35 Distribution des montants garantis et de la sélection médicale pour le cluster de sinistralité forte 59
5.36 Distribution des CSP pour le cluster de sinistralité forte 59
6.1 Premiers résultats après application du GLM 66
6.2 Résultats retraités du GLM 67
6.3 Résultats finaux du GLM 69
7.1 Taux d'incidence après croisement des modalités des variables explicatives 73
7.2 Taux d'incidence après croisement des modalités des variables explicatives (2) . . 73
8.1 Transformation de la base en vue de l'application des modèles de survie 78
8.2 Taux d'incidence bruts par tranches d'âge 79
8.3 Taux d'incidence lissés par âge 81
8.4 Observés / Attendus : validation de la loi pour le groupe B 82
8.5 Taux d'incidence (base complète) lissés par âge pour le groupe B 83
9.1 Vérification de l'hypothèse de log-linéarité pour le groupe A 85
9.2 Vérification de l'hypothèse de log-linéarité pour le groupe C 86
9.3 Application du modèle de Cox 86
9.4 Validation de la loi positionnée pour le groupe A 87
9.5 Validation de la loi positionnée pour le groupe C 88
9.6 Taux d'incidence finaux pour tous les groupes de risque 88
11.1 Répartition Accident / Maladie par âge 94
11.2 Taux d'incidence pour l'accident à 7 jours et la maladie à 15 jours 95
11.3 Extrait de la table de maintien en incapacité 96
11.4 Coefficients de passage pour la maladie à 15 jours 97
11.5 Coefficients de passage pour l'accident à 7 jours 97
11.6 Taux d'incidence finaux pour l'accident 98
11.7 Taux d'incidence finaux pour la maladie 98
12.1 Tarif calculé à partir de l'ancienne loi 99
12.2 Tarif calculé à partir de la nouvelle loi 100
12.3 Coefficients de majoration / minoration pour le nouveau tarif 100
12.4 Recalcul des primes sur la base historisée de 2018 101

Liste des tableaux

3.1 Tableau comparatif avec les flux comptables (en euros) 11
3.2 Tableau comparatif avec les flux comptables (en euros) 12
4.1 Taux d'incidence en fonction de la sélection médicale 24
4.2 Taux d'incidence par franchise 25
4.3 Caractéristiques d'un individu peu risqué 27
4.4 Caractéristiques d'un individu risqué 27
5.1 Exposition par groupe 51
5.2 Caractéristiques des groupes de risque identifiés 60